Understanding Incentives in Subjective Evaluations: Evidence from Educators

Andrew J Morgan *

November 15, 2020

Abstract

Employee evaluation is a central function of any firm, yet implementing a system that aligns incentives between the firm, employees, and supervisors while providing meaningful feedback remains a difficult task. Supervisor ratings that are more lenient and more compressed than the underlying distribution of employee productivity remain a subject of concern for organizations in the public and private sectors. I investigate the use of subjective evaluations in a novel setting in a large, urban public school district in which supervisors are penalized for reporting ratings that misalign with an objective measure of individual employee productivity. First, I document that subjective evaluations of teachers consistently increased throughout my sample period, with mixed evidence on the relationship of this increase to teacher productivity. I then turn to investigating the mechanisms by which ratings assignment could be influenced by the incentives principals and teachers face in this system. I document no evidence that supervisors alter their rating behavior in response to the penalty, using a difference in differences approach to determine that teachers who were and were not included in the penalty calculation received no different ratings after the introduction of the penalty. Using regression discontinuity and difference in discontinuity estimates I also find that teachers who are marginally close to a salary increase and thus face a strong financial incentive for higher ratings are assigned higher ratings that do not correspond to productivity improvements. Finally, observation scores appear to rise with every repeated interaction with a supervisor, accounting for other factors that may influence productivity, but whether this reflects increased productivity or more lenient ratings is less clear.

JEL Classification: I20, I28, H75, J33, M520

*PhD Candidate in Economics, University of Illinois at Chicago, Email: amorga27@uic.edu. For most recent version of the paper, please see https://sites.google.com/view/andrewjmorgan

1 Introduction

A central task of supervisors is to assess the performance of and provide feedback to employees. However, use of subjective measures of performance, particularly when used in compensation, often comes with substantial biases. Supervisors are apt to evaluate employees more leniently than their true underlying performance, and evaluation distributions are frequently more compressed, generally coming from only the upper end of the ratings set (Weisberg et al. (2009), Frederiksen, Lange and Kriechel (2017)). In spite of this, subjective evaluations are used extensively across many industries. While there is ample evidence of the use of subjective evaluations of performance, less is known about the ways in which firms can induce more effective subjective ratings toward more productive ends, particularly in firms where workers engage in complex, multi-task occupations.

I study a novel policy in a large, urban public school system in which supervisors are financially incentivized to report more accurate ratings of the teachers in their schools, as a part of a large-scale overhaul of the way educators are evaluated and paid. The district, henceforth referred to simply as the "District", began the reform in the 2012-2013 school year, with a focus on improving achievement through improving educator effectiveness. I investigate the ratings assignment behavior from the outset of this system as well as the response to a penalty in which supervisors receive a potential financial cost for assigning ratings that less closely align with an objective measure of employee quality. Using a difference in differences approach I show that supervisors do not significantly change their reporting behavior in response.

In contrast to most other work done on subjective evaluations in teaching, my setting is relatively unique in being in a particularly high stakes context where a majority of teacher yearly compensation is determined by subjective evaluations. I investigate how ratings evolve when both principals and teachers face financial consequences for their ratings behavior. I show how the compensation structure in the district impacts the reported ratings and the efficacy and reliability of these ratings in evaluating employees, using teacher value-added to student test scores as a more objective measure to compare against. While subjective ratings increase sharply over my sample period, ratings do not

appear to become less correlated with student outcomes over time.

However, it is less evident to what extent these ratings correspond to changes in employee productivity. I identify key potential mechanisms by which ratings may diverge from observed performance. I present regression discontinuity and difference in discontinuity evidence showing that principals appear to assign ratings that diverge from student outcomes for teachers that face a higher financial incentive from increased ratings. Ratings also are significantly higher for every repeated interaction between supervisor and employee after controlling for other factors that could influence ratings and development. It is less clear to what extent this reflects more lenient ratings or productivity improvements through repeated, continual feedback.

The evidence on the appropriate utilization of subjective employee evaluations remains mixed. Use of subjective evaluations has been shown to be associated with higher levels of employee producitivity (Gibbs et al. (2004)) and school accountability systems frequently employ subjective evaluations like classroom observations as a way to evaluate employees (NCTQ (2016)). At the same time, evaluations that are more lenient and compressed than the underyling distribution of employee quality has long been cited as a consistent problem in subjective evaluation systems (Prendergast and Topel (1996), Jawahar and Williams (1997), Golman and Bhatia (2012)). Inaccurate ratings can lead employees to not know their true productivity and blunt the incentive effects from performance pay systems, and indeed higher levels of compression have been shown to lead to lower productivity in the firm (Kampkötter and Sliwka (2018)). These problems may be particularly pronounced in the public sector, where there may be weaker incentives on supervisors to maximize productivity.

Additionally, subjective evaluations have been shown both to be subject to bias from employer favoritism and to be subject to significant racial and gender biases, further calling into question the reliability of these types of performance ratings, particularly when ratings determine a significant portion of pay (Prendergast and Topel (1996), Jawahar and Williams (1997), Elvira and Town (2001), Moers (2005), Castilla (2012), Drake, Auletto and Cowen (2019)). When specifically used in evaluating teachers, performance

ratings have also been shown to be subject to significant bias from classroom composition, with teachers assigned higher ability students showing much higher evaluation ratings (Whitehurst, Chingos and Lindquist (2014), Steinberg and Garrett (2016)).

Yet the vast majority of firms and state education agencies use some type of subjective performance in evaluating employees (Murphy and Cleveland (1991), NCTQ (2016)). In investigating the determinants and efficacy of subjective evaluations, most prior work focuses on a narrowly-defined measure of productivity in a singular task or low-skilled environment (Bandiera, Barankay and Rasul (2009), Breuer, Nieken and Sliwka (2013)). Studies in most occupations do not allow for measures of individual productivity where output is determined multi-dimensionally. In contrast, investigating questions of efficacy and determinants of subjective evaluations in an education setting affords me a unique opportunity to construct well-defined measures of individual employee effectiveness – value-added. A teacher's value-added to student test scores is an objective measure of productivity that provides for an opportunity to determine divergence of subjective evaluations from individual productivity in a high-skilled, multitask environment.

The proper way to determine effective teaching remains a consistent empirial question. Numerous studies have shown that there is substantial variation in teacher quality (Rockoff (2004), Rivkin, Hanushek and Kain (2005), Kane and Staiger (2008)). Additionally, prior evidence has found that principals are generally *able* to distinguish between low and high effectiveness teachers, but less is known about the decisions that go into how principals choose their ratings for teachers (Jacob and Lefgren (2008), Bacher-Hicks et al. (2019), Harris and Sass (2014)). While value-added has been shown to be an effective measure of teacher effectiveness, value-added measures are only available for a small subset of teachers (Chetty, Friedman and Rockoff (2014)). Thus the vast majority of teachers must be evaluated using alternative approaches. At the same time, value-added may only measure one aspect of student achievement – a student's test score growth. While we would expect value-added and subjective evaluations to be related, the two measures inherently assess different aspects of productivity. My work then fits into the broader literature on teacher evaluations and school accountability and investigates the

reliability and potential improvement of the reliability of teacher evaluations.

At the same time, however, most prior work uses subjective evaluations that may not be applicable when evaluations are used in a high stakes context. Because ratings assigned when stakes are high could be more likely to be lenient (Jawahar and Williams (1997)), there could be substantial differences in the ratings assignment when applied to a setting where principal ratings determine compensation. More recent studies focus their attention on high-stakes evaluation systems, but the work that has been done generally shows that ratings remain modestly correlated with value-added measures (Sartain et al. (2011), Kraft, Papay and Chi (2020)). Grissom and Loeb (2017) also study this question and show that evaluations made in low stakes environment differ slightly from evaluations that impact teacher renewal, but in both instances evaluations are quite high and compressed. My work follows more closely in this line of work, where ratings assignment affects both teacher and principal compensation.

The paper proceeds as follows. I detail the institutional background in Section 2. Section 3 then outlines a conceptual model for supervisor rating assignment and identifies the potential trade-offs supervisors face when assigning ratings. I then proceed to describe the data I use for my empirical approaches in Section 4. I document ratings evolution in Section 5, present evidence on the potential mechanisms by which ratings can diverge from observed performance in Section 6, and conclude in Section 7.

2 Institutional Background

Beginning in the 2012-2013 school year, the District undertook an extensive overhaul of the compensation and evaluation system for its educators. The District first implemented the principal reform and followed it up two years later in 2014-2015 with a parallel program for teachers. Both programs are a broad-based reform of the previous compensation systems with an intense focus on educator development. For both administrators and teachers, the prior compensation system determined primarily by education and experience was replaced with a system that is entirely determined by a formula that incorporated

supervisor observations, surveys and student test scores.

The formula to determine compensation assigns points across these three main categories, with different weights for each of the three categories based on a teacher's subject, grade, and population taught. Total evaluation scores are calculated by a weighted combination of these factors. The district then assigns ranges, with fixed proportions of teachers in each range by year, for the two-year average of the total scores where each range determines salary level. Compensation levels begin at $47,000 for teachers in the Unsatisfactory level and increase to upwards of $90,000 for teachers in the Master level. Compensation is therefore not linear in score, with an increase in level corresponding to between a $3,000 and $7,000 increase in salary. Additionally, once a teacher has reached a specific compensation level, their salary is fixed for up to three years. Only if their total evaluation score corresponds to a lower salary level for three consecutive years can their salary be lowered.[1] Additionally, teachers receive the higher of their reformed salary or their salary in the 2014-2015 school year – their salary in the year prior to policy change – if available. I exploit this nonlinearity and stickiness in wage acquisition in a regression discontinuity design to identify one possible avenue of leniency.

The focus of my paper is on the more subjective nature of classroom observations, which constitute the majority of the total evaluation score for all teachers. Teachers with their own student test scores and surveys receive a 50% weighting on classroom observations while teachers without their own student test scores (i.e. grades and subjects not subject to state-standardized test scores) and without student survey scores receive the highest weight of 80%. Notably, observation scores make up a slightly larger portion of total evaluation ratings in the District than in systems with other similar policies (Putman et al. (2018)). This means that a substantial subset of teachers receive points from both a subjective classroom observations rating as well as from student achievement.[2] For this group of teachers, I estimate value-added to test scores as a way to generate a

[1] In practice, no teacher in my sample has moved to a lower compensation level in any year of my study.

[2] Student achievement points towards a teacher's pay are determined by the maximum of three measures: student pass rate, test score growth compared to a student's peer group, and a district-defined measure of value-added.

more objective measure of quality against which to compare more subjective classroom observations.

Principals are evaluated and compensated across a similar scoring system, where their total evaluation score is determined by their own outside observer's subjective points assignment, as well as parent surveys and test scores of the school they oversee. Additionally, they are required to evaluate teachers by assigning points based on how well the principal believes the teacher adheres to a set rubric, both on a year-long review and from in-classroom observations. These observations are intended to serve two main roles: first, to highlight improvement areas the principal believes the teacher should focus on and second to determine the classroom observations portion of teacher compensation score.

Figure 1 gives an example of the types of standards for which principals are asked to evaluate teachers. In total, principals evaluate teachers on 18 target categories across 4 domains, two of which are used exclusively as year-long reviews rather than strictly in-classroom metrics. For classroom observations, principals observe teachers between 5 and 9 times throughout the school year, decreasing with teacher evaluation level. Principals assign each category a score between 0 and 3 dependent on how well the principal perceives the teacher's performance on that specific metric to be. Scores assigned to each of the categories are then averaged over all observation periods in the year, and average category scores are combined using a weighted average, where more weight is given to categories used in classroom observations. This process generates a metric for observation scores that ranges between 0 and 100 from which the district assigns points that partly determine evaluation level and compensation. Throughout the paper, I refer to this computed total observation score as the teacher's "observation" score.

The District has also made it a focus to attempt to increase the accuracy of observations by requiring principals to receive annual training and certification in observation scoring. Additionally, recognizing that subjective measures of worker productivity are potentially prone to inaccuracy and lenient ratings, the District created a component of the principal's own evaluation score that penalizes principals for having a higher average mismatch between scores they assign teachers in classroom observations and the scores

teachers receive from the test score component of their own evaluation score.

Principals also receive points on a number of separate components as well. One that could directly influence the ratings assignment of teachers is a separate component determined by teacher development, where principals are rewarded for the average growth in their teachers' total evaluation scores across years. Principals with the highest average teacher growth, relative to other principals in the district in that year, receive higher points on this category. The mismatch penalty and the growth component constitute roughly 5% each towards a principal's total pay.

With the introduction of the penalty metric, the District explicitly defines one way in which principals must adhere to a pre-defined accuracy of ratings. However, teachers without their own student test scores were not (could not be) included in this penalty calculation, allowing me to compare observation scores for two groups of teachers–those who were included and those who were excluded. Additionally, this penalty went into effect the second year of the program (2015-2016), giving me the opportunity to compare score differences for teachers with and without their own student test scores for one time period prior to implementation. This allows me to test the efficacy of this penalty by examining any divergence in the trajectory of subjective evaluations between these two groups in a difference in differences design to test to some extent the efficacy of this penalty, with the caveat that I have no substantial pre-period to test for significantly different pre-trends.

3 Conceptual Framework

To illustrate how principals make decisions for subjective ratings and the trade-offs they face, I construct a simplified model of how principals optimize their rating behavior in this context. The following model should not be taken as an exhaustive model of principal and teacher behavior and interactions, but rather it serves to identify the main, broad factors principal face in assigning ratings. Principal rating behavior is determined by multiple factors, and principals may face potentially competing incentives in assigning ratings.

Principals may prefer to accurately assess performance, but at the same time may care about the well-being of their employees. Interpersonal factors based on the long-term nature of a supervisor-employee relationship and costs associated with replacing teachers may lead to principals assigning ratings that are more lenient. For instance if a principal would in isolation assign an accurate, but low, rating to a teacher, they may choose instead to assign a rating that is more lenient and higher if they face interpersonal pressure from a teacher, especially if they believe that they would incur high costs associated with replacing that teacher.

Supervisors s observe the performance p_i of agent i and assign a subjective rating of their performance r_i based on observed performance. Supervisors choose a rating that optimizes their own utility U_s given by the following equation.

$$U_s = -\lambda(r_i - p_i)^2 + \gamma_{is} U_A \tag{1}$$

Following the prior literature, I model supervisors as having a distaste for inaccurate ratings, with λ denoting a supervisor's proclivity for ratings that less closely align with observed performance. Principals may derive negative utility from inaccurately rating employees from two factors. First, inaccuracy likely affects productivity of the firm. If employees are not rated accurately, this impedes their ability to effectively develop skills and weakens the alignement between performance incentives and productivity. If pay is determined by firm productivity, the supervisor may suffer from lower productivity stemming from less well-developed workers. Second, to the extent that accuracy is verifiable by the firm, principals may face reprimand if their reporting deviates too far from true performance.[3] In this framework, I do not differentiate between these factors and combine them into an overall taste for accuracy. I restrict λ to be positive, representing that all principals have some distaste for inaccuracy.

Supervisors also derive utility from their employee's own utility U_A. I model the degree to which this interpersonal relationship matters with a component γ_{is} that can differ for

[3]See Kampkötter and Sliwka (2018), Golman and Bhatia (2012), and Jawahar and Williams (1997) for more in depth discussion of modeling supervisor accuracy behavior.

each supervisor-agent match.[4] γ can be positive or negative where negative values suggest a dislike of individuals, where their disutility benefits the supervisor.

Agents choose effort to maximize

$$U_A = w(r_i) = r_i \tag{2}$$

Employees care about their wage, which is directly determined by their rating. In this simplified model, I model the wage function with respect to rating as a linear and differentiable function, but in my context, neither is the case. Wages do increase in rating, but do so non-linearly and discontinuously, which I exploit later in a regression discontinuity approach.

The optimal rating for each supervisor is then

$$r_i^* = \frac{\gamma_{is}}{2\lambda} + p_i \tag{3}$$

Ratings then differ from observed performance by a factor $\frac{\gamma_{is}}{2\lambda}$. Supervisors who positively value their employee's well-being, $\gamma > 0$, will then assign ratings greater than performance, and vice versa.

I explore the interpersonal component of the model in two ways. First, I identify a subset of teachers that I believe should have the highest likelihood to receive potential special treatment – those that just missed hitting a higher compensation level the year prior. If principals take expected worker pay into account, we might observe a higher observation score for the teachers who just missed out on hitting a higher compensation level relative to the teachers who did not. I also observe whether or not any change in rating corresponds to a change in student outcomes to determine if a change in rating corresponds to a change in potential observed performance. Next, given the panel nature of my data, I observe which principals have rated which teachers in prior observation periods. It may be the case that by observing a teacher previously and establishing a prior rapport a principal then has a higher incentive to deviate from observed performance.

[4]See Prendergast and Topel (1996), Prendergast (2002), Sliwka (2007), and Giebe and Gürtler (2012) for a discussion on employee favoritism and interpersonal relationships in the firm.

Both of these factors should serve to increase the extent to which supervisors weight their employee's utility, γ_{is}. $\frac{\partial r_i^*}{\partial \gamma_{is}} > 0$ given positive λ, meaning that principals who assign more weight for an employee are more lenient with their ratings.

A rather unique aspect of the reform I study is the district-wide implementation of a penalty component for inaccurate ratings from principals. Naturally, this penalty should serve to incentivize principals to more accurately rate teachers, increasing λ. The change in the optimal rating $\frac{\partial r_i^*}{\partial \lambda_s}$ is then $\frac{-\gamma_{is}}{2\lambda^2}$ which is positive if γ_{is} is negative, and vice versa. This implies that ratings for teachers for which a principal has a less favorable view will increase, and decrease for teachers a principals favors, leading to more accurate ratings. Given prior evidence of this and the logic that teachers and principals that like each other likely sort together, I would expect to see that any change in accuracy would likely come from a decrease in ratings. In my analysis, I exploit the fact that the penalty calculation was introduced later in the program and that some teachers were not included in the penalty calculation. If the relative proportion of teachers for which principals have a favorable view is not changing between the two groups, we should expect to see the ratings of teachers that were included fall relative to teachers that were excluded post-introduction.

4 Data

The panel nature of the unique administrative data I use allows me to generate precise measures of individual subjective and objective performance. Because I also know the date, score, supervisor, and employee for each observation period, I am also able to construct differences in observation scores by the number of times a supervisor observes an employee. I use files from the District's administrative systems to construct teacher and principal panel data sets for the school years 2012-2013 to 2018-2019. These data include detailed measures of each of the components and sub-components of the evaluation systems as well as district-calculated total scores from each of these measures. The data also include the raw data needed to calculate these measures, including data from each

observation period for each teacher and an identification to link each observation period to an observing principal.

I merge these data with data from the state administrative system containing employment and demographic information for all staff in the district. State data detail where and in what capacity each District employee serves in each year, with demographic information on each employee's years of professional experience, sex, and race and ethnicity. Using these data, I am also able to link each teacher to the students that they teach in each year.

These data also contain a rich panel of demographics and test scores for each student in the district in each year. They also list classroom enrollment for each of the roughly 150,000 students in the district in each year of my sample period. I merge these to student demographic information containing sex, race and ethnicity, free and reduced price lunch status, limited English proficiency, as well as special education status. Using these data, I am able to link students to teachers in order to construct value-added to student test scores for each teacher to generate a traditionally more objective measure of teacher evaluations.

Student test scores come from the state standardized tests, which were administered to every student in grades 3-8 reading and math. I standardize all student test scores to the state averages of each grade and year, so that each score is normalized around the state mean of 0 and standard deviation 1 within grade and year.

Virtually all (93%) of the approximately 10,000 teachers in the district in each year are observed in classroom and assigned a total evaluation score in each year. Teachers that are not required to be evaluated include certain types of special education teachers and teachers classified as guest or substitute teachers. Each teacher that is evaluated is required to be observed between 5 and 9 times throughout the year, with teachers with lower evaluation levels required to receive more observations.

Table 1 presents information on teacher evaluation components as well as demographic information for two groups of teachers. Because I can only identify the total number of observations for the years 2014-2015 to 2017-2018, this table excludes statistics for teachers

in 2018-2019. Columns 1 and 2 break down these summary statistics by the two analysis groups I use in this paper – all teachers with evaluation scores and a subgroup that contains both evaluations and for which I can construct test score value-added measures.[5] The second group is comprised of grades 3-8 reading and math subject teachers. Across all years, the average total observation score is roughly 73 out of 100, coming from an average of 8 total observations per year for both groups of teachers. The value-added sample is slightly less experienced with an average of around 9 years of experience compared to 10 years for the full sample. Approximately 70-80% of teachers are female, and a little under a third hold an advanced degree. A little over one-third of teachers in both samples identify as Black or African American, a little under a third identify as white, and a little under a third identify or Hispanic or Latine.

5 Understanding Ratings Evolution and Assignment

To understand the way in which evaluations are determined given the changes in the district, I first document the evolution of the observation scores we observe over time. First, I document a pattern that shows observation scores increasing consistently throughout my sample period, indicative of either more lenient ratings or more productive employees, or both. To determine the degree to which observation scores correspond to changes in achievement, I then document the relationship between observation scores and value-added throughout my sample period, finding a modest relationship that persists across the years of my study.

Tying pay to evaluations can create strong incentives for leniency and at the same time could also improve student outcomes if teachers are incentivized to develop more skills and exert more effort. Indeed, taking the district as a whole, my coauthors and I in a separate paper use a synthetic control approach to show that student achievement in the District shows signs of improving relative to comparable districts in the state. An increase in ratings may then reflect some improvement to student achievement but could

[5]Roughly 97% of teachers for whom I can construct value-added have evaluation scores in each year–for roughly 80 teachers out of approximately 2500 in each year.

also be reflective of lenient and inaccurate ratings. It is important to then explore how changes in observation scores correspond to changes in student outcomes at the teacher level to determine how indicative ratings can be for teacher effectiveness.

To document the degree to which observation scores may diverge from student outcomes, I estimate a value-added to each teacher's students' math or reading test scores. I construct yearly value-added estimates for each teacher in the following way:

$$A_{igjt} = \Omega(f(A_{igt-1})) + X\beta + \delta_{gt} + \gamma_{jt} + \varepsilon_{igjt} \tag{4}$$

A_{igjt} measures each student's i math or reading test score in grade g at time t with teacher j. I regress this on a cubic in past student achievement and control for student demographics, X, including student sex, race and ethnicity, free and reduced price lunch status, special education status and limited English proficiency. X also includes school-by-year means of each of these student demographic variables. I additionally control for grade-by-year fixed effects, δ_{gt}. γ_{jt} captures the teacher-by-year fixed effect and represents the teacher value-added to student achievement in each year.

I first document the time trends in observation scores I observe. Figure 2 shows that the broad distributions of observation scores continually shift towards the right since the start of the teacher reform in the 2014-2015 school year. By the last year for which I have data, the most common value of total observation scores hovers around 95 points (out of 100). Table 2 shows the mean observation ratings is increasing across all years of my data. While the overall standard deviation of observation scores does not substantially change over my sample period, ratings are still rather compressed, evident in less willingness on the part of principals to rate employees using the lower ends of the ratings scale. Across years, I observe little usage of the two lowest ratings. Table 2 shows that roughly 56% of teachers in the first year receive evaluations corresponding to an average score within the two highest levels (66 points), meaning that on average very few teachers receive a "progressing" or "unsatisfactory" average rating across all metrics. This number increases to over 80% of teachers in the last year of my study period. Additionally, an increasing number of teachers show virtually no avenue for improvement, receiving an evaluation

score of 95 points or higher. In the latter years, roughly 20% of teachers receive scores at or above 95 points.

We can also see that this change has happened slowly and progressively throughout the sample period. The mean growth in a teacher's observation score year-on-year is similarly high. Table 3 shows the changing distribution of the growth of observation scores over time. The mean teacher receives a modest increase in score each year, ranging from around 6 points to a little over 3 points in the last year of my sample. The change in observation scores remains right skewed throughout the years, with roughly 25% of teachers receiving double digit increases in observation score in any particular year. However, there does exist a sizable fraction of teachers with any decrease in score from the year prior. Roughly 30% of teachers in any year receive any lower observation score than in the year prior. Decreases remain small relative to increases, however, leading to an overall continuous increase in scores.

I also document a general increase in ratings that is coming from across all experience levels of teachers. Figure 3 shows the average observation scores for all teachers in the district by experience level for a given year. In each year, ratings are higher for teachers with higher levels of experience up to roughly 5 years of experience and then level off, similar to what is observed in experience profiles to value-added where, in the early years, there exists an increase in the evaluation metric and then a leveling off after 5-10 years. Note, however, that these plots do not plot the returns to experience, but rather the broad means of observation scores for teachers with a given level of experience in each year.[6] Increases appear to be occurring in all experience levels, with teachers with any given number of years of experience having higher average observation scores than similarly experienced teachers in any prior year.

There is also some evidence that teachers at higher levels of experience have higher ratings increases than teachers at lower levels of experience. Figure 4 plots the difference between the 2018-2019 school year ratings and the 2014-2015 ratings for teachers with a given level of experience in each year. While score increase is positive for all experience

[6]I cap experience at 25 years, but the pattern holds when including teachers with up to 40 years of experience. This constitutes around 95% of all teachers.

levels, the difference is increasing nearly throughout the experience distribution. Estimates for teachers past roughly 15 years of experience fluctuate, but are generally higher than teachers with 10 years of experience, and the difference for teachers with 10 years of experience is generally higher than teachers with 5 or fewer years. While observation scores are not expected to be perfectly aligned with productivity measured by student test scores, this is somewhat contrary to what we would expect of teacher professional development from the literature on value-added.

Potential explanations of this relationship with experience could include an increased pressure principals face from more experienced teachers holding higher influence in the school or from higher professional development. Kalogrides, Loeb and Beteille (2013) find that higher teacher experience is associated with higher prior student achivement within a school, suggesting that more experienced teachers may be better able to influence a principal when it comes to classroom assignment. At the same time, Kraft, Papay and Chi (2020) document a similar increase in observation scores in a returns to experience model and attribute the increase to professional development not measured by test scores, but can only examine this question for teachers with 10 or fewer years of experience due to data limitations.

However, to the extent that observations and student achievement are related, if increases in teacher evaluations by experience level correspond to increases in student achievement, we should expect to see the difference in value-added between the two years for more experienced teachers be higher than the difference in value-added for less experienced teachers as well. This should especially be the case given that the District has placed a strong focus on improvement to student achievement. I present comparable experience figures to Figure 4 for yearly math and reading value-added to determine if the difference in value-added is changing at a commensurate level to observation scores. However, Figures 5 and 6 show differences in value-added by experience with no similar pattern. The change in value-added for teachers with higher levels of experience is no different than the change in value-added for teachers with lower levels of experience. Together, these suggest that observation scores are increasing for more highly experienced

teachers with no noticeable increase in value-added for these same teachers. Thus, it appears to be the case that more experience teachers may be receiving ratings that do not accurately reflect improved student achievement, relative to less experienced teachers.

The use of value-added also allows me to have a reliable measure of individual objective performance, and I now turn to directly measuring how related subjective measures are to objective measures of individual performance in this high stakes context. Table 4 shows that overall correlations between observations and value-added are positive and modest in magnitude. Correlations between both reading and math value-added fluctuate throughout the years, hovering around 0.2 to 0.3, with math being slightly more positively correlated than reading– roughly in line with what prior work has found (Jacob and Lefgren (2008), Kraft, Papay and Chi (2020)). Notably, for both subsets of teachers, correlations do not appear to exhibit an increasing or decreasing pattern, suggesting that there is no rank order shifting across the years – that principals remain fairly consistent in rating the most effective teachers more highly.

A concern with subjective evaluations in a system where there may be a strong incentive towards lenient ratings is that scores may be inflated so as to be less useful in determining a teacher's future productivity. Given the trend towards higher ratings in the district, to identify if this is happening, I follow a similar procedure to that outlined in Bacher-Hicks et al. (2019) and Rockoff and Speroni (2010) in identifying the predictive validity of subjective observations on future student outcomes. If increases in ratings are becoming less predictive of student achievement, we would expect to see the relationship between predicted observation scores and student achievement fall.

To determine if this is the case, I regress outcomes of students matched to their teacher in a particular year on predicted yearly-standardized observation scores. To generate predicted observations scores and to account for measurement error and fluctuations in measurements, I follow the procedure outlined in the following three regressions. I first regress the two-year prior year observation score $Score_{jt-1}$ on the score from the one-year prior, $Score_{jt}$ (Equation 5). I capture this coefficient, $\hat{\beta}$, and multiply it to the observation score from t to generate a predicted observation score for teacher j in the following year,

$\widehat{Score_{jt+1}}$ (Equation 6). Finally, I match students to the teachers they have in each year, and regress student outcomes in year $t+1$ on predicted observation score, controlling for student demographics, prior achievement scores and prior absences and discipline in each regression as well.

$$Score_{jt-1} = \beta Score_{jt} + \varepsilon_{jt-1} \tag{5}$$

$$\widehat{Score_{jt+1}} = \hat{\beta} * Score_{jt} \tag{6}$$

$$Outcome_{ijt+1} = \delta \widehat{Score_{jt+1}} + X\Omega + \varepsilon_{ijt+1} \tag{7}$$

Table 5 shows the results from this process. If classroom observations less accurately predict student outcomes over time, we should expect to see the coefficient on predicted observations decreasing over time. However, this does not appear to be the case. The coefficients on math and reading test scores in the top and bottom panels, respectively, show that the coefficient for predicted score in 2016-2017 (Column 1) is no smaller for math and reading test scores than the coefficient in 2018-2019 (Column 3). Having a teacher that is one standard deviation higher in predicted observation score increases math test scores by around 0.11 standard deviations and around 0.073 standard deviations for reading teachers. Observation scores overall appear to be moderately correlated with student achievement, with stable predicted estimates over time.

Taken together, use of subjective evaluations in a high stakes compensation context does not appear to be driving a divergence between classroom observations and student outcomes. Observation scores are moderately related to value-added and they appear to not be losing predictive power over time, however magnitudes are relatively modest in each case. Given the steady growth in scores over time, it is important to determine the ways in which principals may be rating teachers that may or may not reflect true productivity growth. In the next section, I turn to directly investigating certain avenues where ratings may be more likely to be applied inaccurately.

6 Avenues for Divergence between and Ratings and Performance

Given the incentive structures faced by both principals and teachers, we might expect more or less lenient ratings for different groups of teachers. Principals may face a financial penalty for inaccurately rating teachers included in the penalty, but face no such trade-off for the teachers excluded from the calculation. Teachers who just missed out on hitting a compensation level in the prior year may have strong incentives surrounding the receipt of ratings in the following year, relative to those who achieved a higher compensation level. Additionally, principals and teachers may face stronger interpersonal effects from having been assigned to ratings periods in prior instances. Investigating the magnitude of these factors is necessary in determining the significance of the incentives principals and teachers face in ratings behavior and assignment.

I estimate three main empirical approaches to identify the extent and determinants of observation score bias in teaching. First, to estimate if principals respond to the penalty by being less lenient, I construct difference in differences estimates, comparing observations ratings of teachers who were and were not included in a principal's penalty calculation before and after the introduction of the penalty. If the penalty was effective in reducing leniency, we should see this manifest as lower observation scores relative to teachers who were not included, after the introduction of the penalty. Second, to investigate the effect of strong financial incentives on ratings, I present regression discontinuity estimates that compare subjective and objective measures of teacher evaluations for teachers who fall just to the left or right of a compensation cutoff the year prior. If principals are assigning more lenient ratings due to the discontinuous effects of teachers' utility and wages, we would expect to see an increase in ratings for those teachers just to the left of the cutoff relative to those just to the right in the year prior, with no comparable increase in student achievement. Lastly, to determine the importance of repeated observations on ratings assignment, I investigate the impact of having the same observer in prior observations periods, controlling for principal and teacher observation experiences.

6.1 Penalty Effects on Supervisor Ratings

To investigate the effect of the penalty I exploit the fact that the penalty calculation came into effect in the second year of the program and that only a subset of teachers under any particular principal were included in the penalty calculation. I calculate difference in differences estimates for each year, comparing teachers who were and were not included in the calculation before and after the introduction of the penalty. This roughly corresponds to calculating the difference in each year for teachers with and without their own student's test scores, before and after the 2015-2016 school year. Because these two types of teachers face different pay schemes, it is important to note that their baseline levels of observation ratings may differ. This should be captured by the level differences in the period before the penalty was introduced.

Specifically, I estimate

$$score_{jt} = \beta_1 included_{jt} + \theta_t + \sum_{t=2015}^{2019} \theta_t * included_{jt} + \varepsilon_{jt} \tag{8}$$

where *included* equals 1 for teachers j who were included in a principal's penalty calculation, and 0 otherwise. θ represents year fixed effects and the interactions between each θ and the *included* dummy variable represent the effect of the penalty in each year. Some estimates also include teacher level demographic controls of years of professional experience and having an advanced degree.

Examining average observation scores between these two groups of teachers suggest that principals do not appear to be differentially altering their rating behavior for the teachers that were and were not included in the penalty calculation. Figure 7 plots the average observation scores for teachers with (blue line) and without their own student test scores (red line).[7] The differences between the two lines after the introduction of the penalty in the 2015-2016 school year do not appear to be meaningfully different from the year in which the penalty was not in place. The difference between the observation scores

[7]Note that this sample differs slightly from the value-added sample used in the other portions of this paper. The district calculates the penalty based on points gained from all teachers with their own student achievement–not just those for whom value-added can be calculated – and thus constitutes a more broad group of teachers.

for teachers that were and were not included in the penalty calculation is highest in the 2015-2016 school year, where teachers that were not included in the penalty calculation have an approximately 1 point higher average score, but all differences are small in every other year.

Although it could be the case that principals choose to rate the different teacher groups differently for other reasons, the effect of this should be captured by the pre-period difference between the two groups. Difference in differences estimates in Table 6 show the average observation score difference for each year relative to the year before the penalty was introduced. The estimates imply that observation scores for each year are not significantly changed between the two groups after the penalty. The only significant effect comes in 2017, and the sign of which implies, if anything, that teachers with their own student test scores increase in average observation score relative to the pre-period, counter to what we might expect after the introduction of the the penalty. Changing demographics of teachers does not explain this result, as after accounting for teacher demographic controls in Column 2 the estimates remain small and largely insignificant. Thus, this evidence suggests that principals are not taking the penalty into account when rating teachers.

Because I do not have data prior to the 2014-2015 school year to be able to estimate the likelihood of divergent trends unrelated to treatment, difference in differences estimates should be treated as more descriptive. It may also be the case that there could be potential spillover effects of the penalty on a principal's ratings for excluded teachers. If this was the case, this would create downward bias in our difference in differences effect, and could explain why I find no significant differences.

6.2 Financial Incentives on Teacher Ratings and Performance

I present evidence that suggests that principals may be more inclined to assign more lenient ratings if they believe this could help increase a teacher's pay a meaningful amount. Since teachers are only subject to a pay increase if their score puts them in a higher compensation level, it may be the case that teachers and principals have the strongest incentive

to be more lenient for teachers who just missed out on attaining a higher compensation level in the year prior. Teachers in this group should have the highest incentive to lobby for more lenient scores, put in higher effort to pass the threshold through higher student achievement and teacher development, or both. Additionally, because teachers receive a salary protection for up to three years after achieving a compensation level, teachers just to the right of a compensation cutoff should face discontinuously less incentive than teachers just to the left.

To test this, I estimate regression discontinuity estimates for teachers just around a compensation threshold for their observation scores and math and reading value-added in the next year. While there is some incentive no matter where a teacher is in the distribution of ratings for continuous improvement, the incentive for lenient ratings should be much less strong for teachers who are "far" enough away from a compensation cutoff. I present estimates showing that while value added does not appear to be discontinuously changing for teachers around a compensation cutoff, observation scores differ, albeit insignificantly.

However, other factors may also be changing around a threshold. Teachers that do not achieve a higher level in the year prior may lose motivation if they believe their low rating does not reflect their true ability. Principals may also derive information effects based on evaluation level, such that it could be the case that principals are incorrectly rating teachers who hit just above a threshold substantially higher than those just below solely because they have a higher rating level. Cullen et al. (2016) show some evidence that educators may not be correctly exploiting the full information effects of similar ratings. If this is the case, principals may be rating teachers more highly simply based on a recognition of higher evaluation level. Because of this, we might expect to see a higher rating for teachers just above the cutoff relative to those below, in the absence of other factors. To account for these possibilities, I present separate estimates for teachers who did and did not have grandfathered pay from before the teacher reform was introduced and additionally estimate difference in discontinuity estimates, following Grembi, Nannicini and Troiano (2016), to specifically eliminate other effects of the cutoff that may influence

ratings. Teachers without a financial incentive would likely still be subject to demotivating factors, and principals' ratings could still reflect asymmetric information for this group as well.

To investigate whether or not principals are responding to this incentive faced by their teachers, I estimate regression discontinuity estimates for teachers who just missed out on attaining a compensation level in the prior year. Specifically, I estimate regression discontinuity equations in the following way

$$C_{jt+1} = \beta_0 above_{jt} + \beta_1 point_{jt} + \beta_2 above * point_{jt} + \varepsilon_{jt} \tag{9}$$

where C represents either the observation score or math or reading value added of teacher j in year $t+1$. I regress this on whether or not a teacher's score in the prior year placed them in a higher compensation level, *above*, the distance from the threshold value in the year prior, *point*, and the interaction between these variables. β_0 represents the linear approximation of the discontinuity for just achieving a higher compensation level on a teacher's evaluation components in the next year. I present estimates for multiple bandwidths of teachers close to a cutoff, with main estimates shown for teachers within 4 points of a threshold.

To eliminate any other concurrent effects around the cutoff, I compare discontinuities for two groups of teachers, those with and without grandfathered salary from prior to the implementation of the reform. I additionally present formal difference in discontinuity estimates, estimated from the following regression.

$$\begin{aligned} C_{jt+1} = \beta_0 above_{jt} + \beta_1 point_{jt} + \beta_2 above * point_{jt} \\ + \beta_3 fincentive_{jt} + \beta_4 fincentive * point_{jt} \\ + \beta_5 fincentive * above_{jt} + \beta_6 fincentive * above * point_{jt} + \varepsilon_{jt} \end{aligned} \tag{10}$$

The variable $fincentive_{jt}$ is an indicator for whether or not a teacher has a financial incentive for hitting a compensation level in the next year, i.e that they did not have

grandfathered pay that was higher than the next compensation level they face. β_5 represents the difference in the discontinuity between these two groups, isolating the financial incentive faced by teachers and eliminating potential other factors changing around this cutoff.

Figure 8 presents a regression discontinuity plot of classroom observation scores for teachers who were not subject to grandfathered pay around the threshold and thus had a financial incentive (Figure 8a) and teachers who did not have a financial incentive (Figure 8b). We see some evidence in Figure 8a that teachers within 4 points of a cutoff on the left of a threshold may have slightly higher observation scores in the next year after missing the cutoff than teachers just to the right of the threshold. Table 7, Column 1 and 2 of Panel A shows that the discontinuity with and without controlling for teacher experience, degree status and sex results in an insignificant roughly 1.5 point decrease after the threshold. In contrast, teachers on the right side of the threshold and without a financial incentive in fact have a much higher observation score than teachers who missed the cutoff, potentially reflecting asymmetric information around the cutoff. Table 7, Column 1 and 2 of Panel B shows these teachers receive a substantial but insignificant increase of roughly 3.5 points.

The comparable estimates for value-added do not show similar changes. When turning to Figures 9 and 10 we see small, insignificant and inconsistent results for math and reading value-added. These results are not subject to information asymmetry across the threshold, so we should not expect to see differences between teachers with and without financial incentive, other than through effects of (de)motivation. Teachers just to the left of the cutoff, both with and without financial incentive appear to have slightly higher but insignificant math value-added, with Column 4 of Table 7 showing a 0.008 and -0.174 change in math value-added for teachers with and without a financial incentive, respectively. The case is reversed for reading. Figure 10 and Table 7 Column 6 show an estimated discontinuity of around 0.04 and 0.07 for reading value-added for the two groups. Thus while there may be some slight evidence of increased observation scores, no such pattern holds for math or reading value-added.

Table 8 shows that this pattern holds for varying bandwidths, but results remain noisy. Columns 1-3 show results for observation scores, math and reading value-added for teachers who were within 5 points of the threshold in the year prior, and Columns 4-6 show for these for a bandwidth of 3. Both sets of results suggest a slight, insignificant fall in subjective evaluation scores for teachers just after the threshold, with no clear change in value-added for teachers with a financial incentive, with a slight, insignificant increase in observation scores and no change in value-added for teachers without a financial incentive.

Putting these two sets of results together in a formal difference in discontinuity design results in a significant and substantial drop in observation scores for teachers just after the threshold with a financial incentive relative to the discontinuity for those without an incentive. Column 1 of Table 9 shows that there is a 5 point decrease in observation scores for teachers who have hit a compensation level, with no corresponding significant change in value-added, suggesting that teachers may in fact receive more lenient ratings as a result of being "close" to receiving more pay.

A concern with any regression discontinuity design is that individuals can perfectly manipulate the side of the threshold onto which they fall, leading to biased estimates from incomparable individuals on either side of the cutoff. In my case, while I make the case that individual teachers and principals may try to manipulate scores around a cutoff in this way, they do not appear able to perfectly influence their status. First, given that thresholds for each compensation level varies from year to year and is assigned by the district, teachers (and principals) cannot know ahead of time *precisely* how much they must increase their scores from the prior year. The empirical evidence also bears this out. Figure 11 shows no discontinuity in the number of teachers on either side of the threshold for both groups of teachers and Table 10 shows virtually no change in observable characteristics for these teachers as well. Neither experience nor advanced degree status changes meaningfully at the cutoff for either group of teachers, but advanced degree status does have a marginally significant discontinuity around the cutoff for teachers with a financial incentive.

Taken together, these regression discontinuity and difference in discontinuity estimates

provide some evidence that principals may be more inclined to be more lenient with the evaluations they assign to teachers who have strong compensation incentives attached to their evaluations.

6.3 Effect of Repeat Interactions on Ratings

I next turn to investigating the extent to which repeated interactions may play a role in rating assignment. It may be the case that principals rate teachers more accurately based on if they rated them in previous instances, all things equal. Prior evidence has shown that subjective ratings can be influenced by repeated interactions between supervisors and employees (Rothstein (1990), Borman, White and Dorsey (1995)). Ratings are then likely to become more accurate for each additional observation that a supervisor observes the same employee. At the same time, two additional factors may influence ratings upward. Repeated interactions may create more opportunity for increased professional development if principals are better able to identify areas of improvement and teachers are able to respond to identified improvements through repeated feedback. But principals may also feel more pressure to increase ratings if they have developed a stronger personal relationship with their employee and may reward teachers more highly for responding to individualized feedback. While accuracy implies a positive or negative change in score for each additional rating, stronger personal relationships are more likely to increase observation scores.

To determine the role of repeated observations on score, I estimate a model that predicts yearly-standardized score on each metric of the observation rubric according to observation histories of the principal and teacher. Because teachers may develop skills by having been observed more often and principals may be better able to more accurately rate teachers by observing others more often, I additionally control for the total observations done by the observer and the total number of observations of that employee by any observer.

$$score_{jrt} = \eta_r + \rho_{jt} + \beta_0 exp_{rt} +$$
$$\beta_1 observed_{rt} + \beta_2 observee_{jt} + \beta_3 observed\_observee_{jrt} + \varepsilon_{jrt} \quad (11)$$

Equation 11 regresses a teacher j's metric score on the total observations done by an observer, *observed*, observer years of experience in an administrative role *exp*, the total number of observations on that teacher *observee*, and the total number of observations by that observer for that teacher *observed_observee*, while also accounting for observer η and teacher-by-year fixed effects ρ. β_3 then captures the additional effect on a teacher's score of each additional observation by the same observer.

The results of this model are shown in Table 11. First, it is the case that observers with more observations and teachers with more observations have higher average scores. Each additional observation done by an observer increases the average score on a metric by 0.0005 standard deviations. Moving up one standard deviation of roughly 200 observations in observer observation experience then would result in a 0.1 standard deviation increase in metric score. This is a separate effect from experience serving in a principal capacity, where there is no significant effect on observation scores. Teachers also increase in average score with additional observations. On average, each observation results in a 0.03 standard deviation increase in average score, and moving up one standard deviation (around 9 observations) in teacher observations results in an increase of 0.27 standard deviations of average score. The last row of Table 11 shows the effect of an observer observing the same individual teacher one additional time. After controlling for total observations and principal and teacher-by-year fixed factors, an observer observing a teacher one additional time yields an additional 0.01 standard deviation increase in score. This finding fits with the model where principals may be rewarding teachers more strongly for implementing more individualized feedback, either through increased professional development or more lenient ratings.

Overall, the evidence presented in this section suggests there may be meaningful im-

pacts of differing incentive schemes on the determinants of subjective ratings. While principals do not appear to alter their rating behavior in response to a modest financial penalty for themselves, they may take into account the financial incentives faced by their employees. It is important for policy-makers to then account for the different incentives faced by both principals and teachers, and to consider the different avenues by which observations can deviate from objective measures of performance.

7 Conclusion

Given the increased focus on the utilization of teacher evaluations and school accountability, the evidence I have presented is highly relevant for policy-makers. Careful consideration should be taken when designing subjective evaluation systems to account for the incentives faced by both supervisors and employees. Understanding the ways in which supervisors assign ratings is key to implementing a system that aligns incentives between the firm, supervisors, and employees while still allowing for ratings that provide meaningful employee feedback.

I add to the growing literature on the use and determinants of subjective evaluation in the education sector, speaking to the concerns in subjective evaluation systems of ratings that less accurately reflect true productivity. I find that implementing a system of subjective evaluations tied to performance compensation in a large, urban public school system produced evaluations that grew sharply over time, and while student achievement also grew considerably during this same time period, the evidence is more mixed on if this increase fully corresponds to increased employee productivity. While subjective evaluations remain predictive of student achievement over time, I present evidence that suggests that supervisors may face strong incentives towards ratings that less accurately reflect employee productivity in certain instances. Ratings are significantly higher when teachers face strong financial incentives from their ratings, with no corresponding increase in other measures of performance. Teachers who just miss out on hitting a higher compensation level in the year prior have significantly higher observation scores in the next

year compared to teachers who did attain a higher level, with no significant change in student achievement.

Altering the ratings assignment behavior for supervisors also may be difficult. Using a difference in differences design, I show that supervisors appear to have no response to a modest financial penalty for assigning ratings that deviate from district-defined measures of performance. Additionally, repeated observations between supervisors and teachers results in higher subjective ratings, but the extent that this may represent more lenient ratings or increased professional development is less clear.

References

Bacher-Hicks, Andrew, Mark J Chin, Thomas J Kane, and Douglas O Staiger. 2019. "An Experimental Evaluation of Three Teacher Quality Measures: Value-added, Classroom Observations, and Student Surveys." *Economics of Education Review*, 73: 1–15.

Bandiera, Oriana, Iwan Barankay, and Imran Rasul. 2009. "Social Connections and Incentives in the Workplace: Evidence From Personnel Data." *Econometrica*, 77(4): 1047–1094.

Borman, Walter C, Leonard A White, and David W Dorsey. 1995. "Effects of Ratee Task Performance and Interpersonal Factors on Supervisor and Peer Performance Ratings." *Journal of Applied Psychology*, 80(1): 168–177.

Breuer, Kathrin, Petra Nieken, and Dirk Sliwka. 2013. "Social Ties and Subjective Performance Evaluations: An Empirical Investigation." *Review of Managerial Science*, 7(2): 141–157.

Castilla, Emilio J. 2012. "Gender, Race, and the New (Merit-Based) Employment Relationship." *Industrial Relations (Berkeley)*, 51: 528–562.

Chetty, Raj, John N Friedman, and Jonah E Rockoff. 2014. "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates." *The American Economic Review*, 104(9): 2593–2632.

Cullen, Julie Berry, Gregory Phelan, Eric A Hanushek, and Steven G Rivkin. 2016. "Performance Information and Personnel Decisions in the Public Sector: The Case of School Principals."

Drake, Steven, Amy Auletto, and Joshua M Cowen. 2019. "Grading Teachers: Race and Gender Differences in Low Evaluation Ratings and Teacher Employment Outcomes." *American Educational Research Journal*, 56(5): 1800–1833.

Elvira, Marta, and Robert Town. 2001. "The Effects of Race and Worker Productivity on Performance Evaluations." *Industrial Relations (Berkeley)*, 40(4): 571–590.

Frederiksen, Anders, Fabian Lange, and Ben Kriechel. 2017. "Subjective Performance Evaluations and Employee Careers." *Journal of Economic Behavior and Organization*, 134: 408–429.

Gibbs, Michael, Kenneth A Merchant, Wim A. Van der Stede, and Mark E Vargus. 2004. "Determinants and Effects of Subjectivity in Incentives." *The Accounting Review*, 79(2): 409–436.

Giebe, Thomas, and Oliver Gürtler. 2012. "Optimal Contracts for Lenient Supervisors." *Journal of Economic Behavior and Organization*, 81(2): 403–420.

Golman, Russell, and Sudeep Bhatia. 2012. "Performance Evaluation Inflation and Compression." *Accounting, Organizations and Society*, 37(8): 534–543.

Grembi, Veronica, Tommaso Nannicini, and Ugo Troiano. 2016. "Do Fiscal Rules Matter?" *American Economic Journal. Applied Economics*, 8(3): 1–30.

Grissom, Jason A., and Susanna Loeb. 2017. "Assessing Principals' Assessments: Subjective Evaluations of Teacher Effectiveness in Low- and High-Stakes Environments." *Education Finance and Policy*, 12(3): 369–395.

Harris, Douglas N, and Tim R Sass. 2014. "Skills, Productivity and the Evaluation of Teacher Performance." *Economics of Education Review*, 40: 183–204.

Jacob, Brian A, and Lars Lefgren. 2008. "Can Principals Identify Effective Teachers? Evidence on Subjective Performance Evaluation in Education." *Journal of Labor Economics*, 26(1): 101–136.

Jawahar, I. M, and Charles R Williams. 1997. "Where All the Children Are Above Average: The Performance Appraisal Purpose Effect." *Personnel Psychology*, 50(4): 905–925.

Kalogrides, Demetra, Susanna Loeb, and Tara Beteille. 2013. "Systematic Sorting: Teacher Characteristics and Class Assignments." *Sociology of Education*, 86(2): 103–123.

Kampkötter, Patrick, and Dirk Sliwka. 2018. "More Dispersion, Higher Bonuses? On Differentiation in Subjective Performance Evaluations." *Journal of Labor Economics*, 36(2): 511–549.

Kane, Thomas J, and Douglas O Staiger. 2008. "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation."

Kraft, Matthew A, John P Papay, and Olivia L Chi. 2020. "Teacher Skill Development: Evidence from Performance Ratings by Principals." *Journal of Policy Analysis and Management*, 39(2): 315–347.

Moers, Frank. 2005. "Discretion and Bias in Performance Evaluation: The Impact of Diversity and Subjectivity." *Accounting, Organizations and Society*, 30(1): 67–80.

Murphy, Kevin R, and Jeanette N Cleveland. 1991. "Performance Appraisal: An Organizational Perspective." Allyn and Bacon.

NCTQ. 2016. "State Evaluation Briefs." National Council on Teacher Quality.

Prendergast, Canice. 2002. "Uncertainty and Incentives." *Journal of Labor Economics*, 20(S2): S115–S137.

Prendergast, Canice, and Robert H. Topel. 1996. "Favoritism in Organizations." *The Journal of Political Economy*, 104(5): 958–978.

Putman, Hannah, Elizabeth Ross, Kate Walsh, Kelli Lakis, and Kency Nittler. 2018. "Making a Difference: Six Places where Teacher Evaluation Systems Are Getting Results." National Council on Teacher Quality.

Rivkin, Steven G, Eric A Hanushek, and John F Kain. 2005. "Teachers, Schools, and Academic Achievement." *Econometrica*, 73(2): 417–458.

Rockoff, Jonah E. 2004. "The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data." *The American Economic Review*, 94(2): 247–252.

Rockoff, Jonah E, and Cecilia Speroni. 2010. "Subjective and Objective Evaluations of Teacher Effectiveness." *The American Economic Review*, 100(2): 261–266.

Rothstein, Hannah R. 1990. "Interrater Reliability of Job Performance Ratings: Growth to Asymptote Level With Increasing Opportunity to Observe." *Journal of Applied Psychology*, 75(3): 322–327.

Sartain, Lauren, Sara Ray Stoelinga, Eric R Brown, Stuart Luppescu, Kavita Matsko, Frances Miller, Claire Durwood, Jennie Jiang, and Danielle Glazer. 2011. "Rethinking Teacher Evaluation in Chicago: Lessons Learned from Classroom Observations, Principal-Teacher Conferences, and District Implementation." Consortium on Chicago School Research.

Sliwka, Dirk. 2007. "Trust as a Signal of a Social Norm and the Hidden Costs of Incentive Schemes." *The American Economic Review*, 97(3): 999–1012.

Steinberg, Matthew P., and Rachel Garrett. 2016. "Classroom Composition and Measured Teacher Performance: What Do Teacher Observation Scores Really Measure?" *Educational Evaluation and Policy Analysis*, 38(2): 293–317.

Weisberg, Daniel, Susan Sexton, Jennifer Mulhern, and David Keeling. 2009. "The Widget Effect." *The Education Digest*, 75(2): 31–.

Whitehurst, Grover J, Matthew M Chingos, and Katharine M Lindquist. 2014. "Evaluating Teachers with Classroom Observations: Lessons Learned in Four Districts." Brookings Institution.

Figures

2.3 DELIVERY: Facilitates clear, cohesive, and purposeful learning experiences

ESSENTIAL TEACHER SKILLS & ACTIONS	Exemplary (3 Points)	Proficient (2 Points)	Progressing (1 Point)	Unsatisfactory (0 Points)
• Supports objectives, prior learning, and all student populations based on their subject, grade, and level with appropriate instructional strategies • Delivers content clearly, accurately, and coherently • Incorporates appropriately varied digital and/or print and/or hands-on instructional resources • Emphasizes the value and connection of content to overall learning and prior knowledge • Combines differentiated and relevant instructional strategies and questioning techniques to maintain appropriate pace and engagement	**Consistently and effectively** presents the content: • Logically, coherently, and in a grammatically correct fashion • Supporting the learning of the posted objective(s) • Building on content previously mastered • Supporting all student populations based on their subject, grade, and level • **Supporting cross-curricular learning** • **Allowing for student input** Consistently **and appropriately** uses multiple, differentiated: • Strategies/materials • Questioning techniques • Academic language • Technologies to engage and emphasize key concepts and their value with **no** irrelevant information Instructions, procedures, and material usage for participating in activities are clear to **all or nearly all** students	**Consistently** presents the content: • Logically, coherently, and in a grammatically correct fashion • Supporting the learning of the posted objective(s) • Building on content previously mastered • Supporting all student populations based on their subject, grade, and level **Consistently** uses **multiple, differentiated**: • Strategies/materials • Questioning techniques • Academic language • Technologies to engage and emphasize key concepts and their value with **little to no** irrelevant information Instructions, procedures, and material usage for participating in activities are clear to **most** students	**Generally** presents content logically and coherent fashion, but: • **Some** parts are unclear, grammatically inaccurate, or developmentally inappropriate • **May not** effectively support the learning of the posted objective(s) • **May not** build on content previously mastered • **May not** support all student populations based on their subject, grade, and level Uses **limited**: • Verbal and nonverbal techniques to convey concepts and their value • Academic language with **some** irrelevant information Instructions, procedures, and material usage for participating in activities are clear to **some** students	Presents content and purpose: • In a **confusing** way, using unclear, grammatically incorrect, and/or incoherent language • With **little to no evidence** of instruction in support of the posted objective(s) • **Does not** build on content previously mastered • **Does not** support all student populations based on their subject, grade, and level **Rarely** uses: • Verbal and nonverbal techniques to convey concepts and their value • Academic language with some irrelevant or **inaccurate** information Instructions, procedures, and material usage for participating in activities are clear to **very few** students
May 2019 Revision				

Figure 1: An example of one of the 18 metrics on which principals are asked to evaluate teachers in the classroom. Principals assign points on each metric, from 0 to 3 according to how well they believe a teacher follows this rubric.

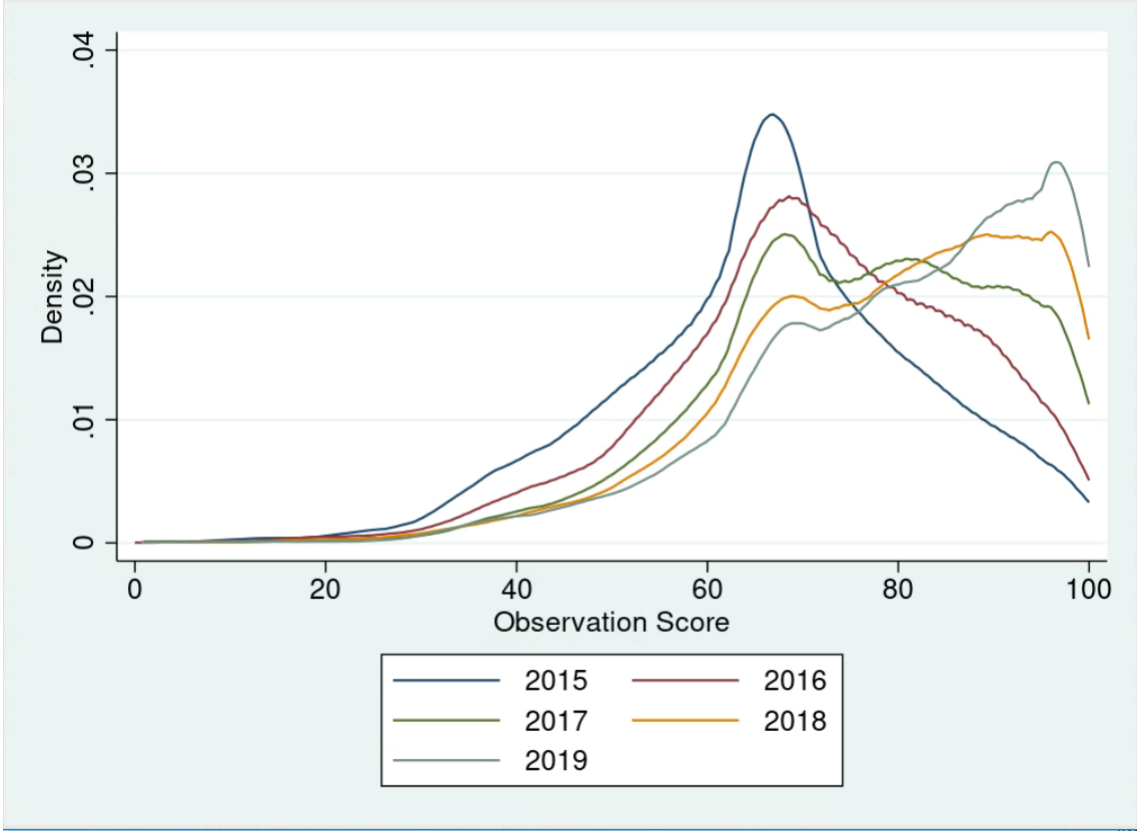


Figure 2: Distributions of classroom observation scores over each school year of my sample period. Distributions consistently shift towards the upper end of the ratings set across my sample period, to where the modal observation score is around 95 points out of 100 in the 2018-2019 school year.

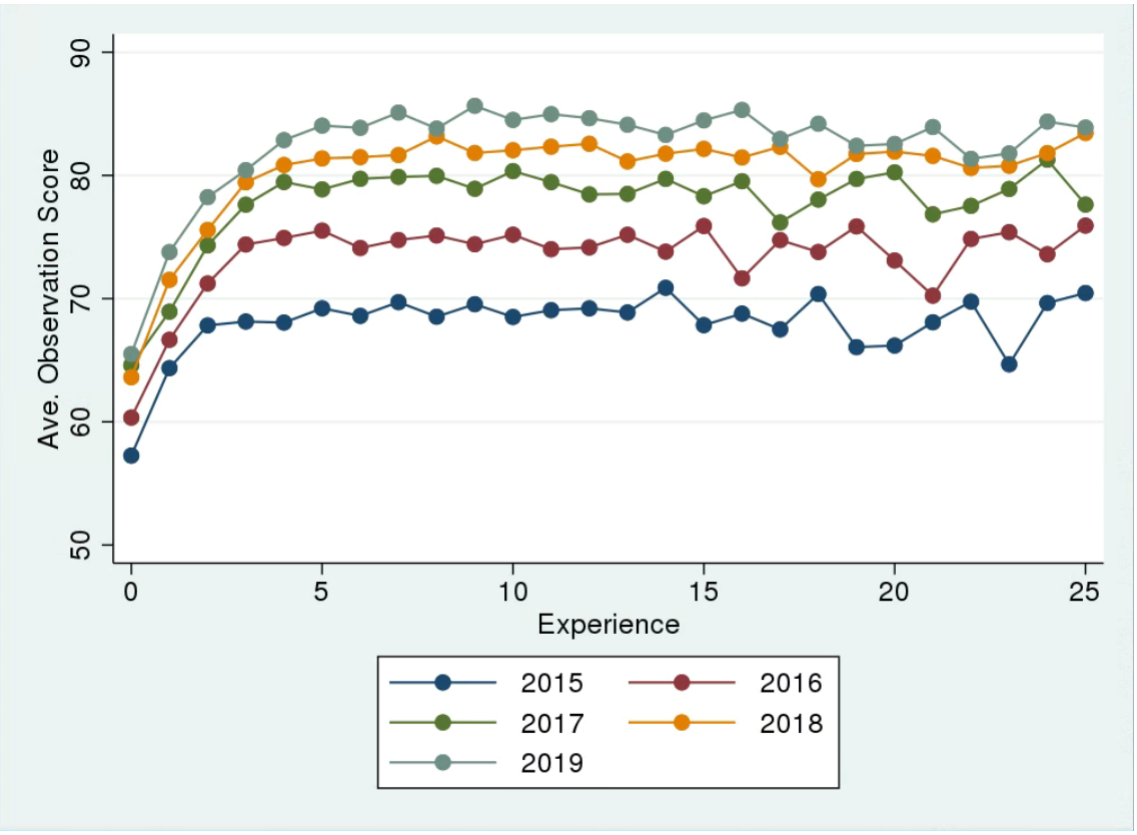


Figure 3: Average observation scores by teacher experience. Each dot represents the average classroom observation score for teachers with that level of experience in each year. Each line represents the pattern of average observation scores for one school year of my sample period.

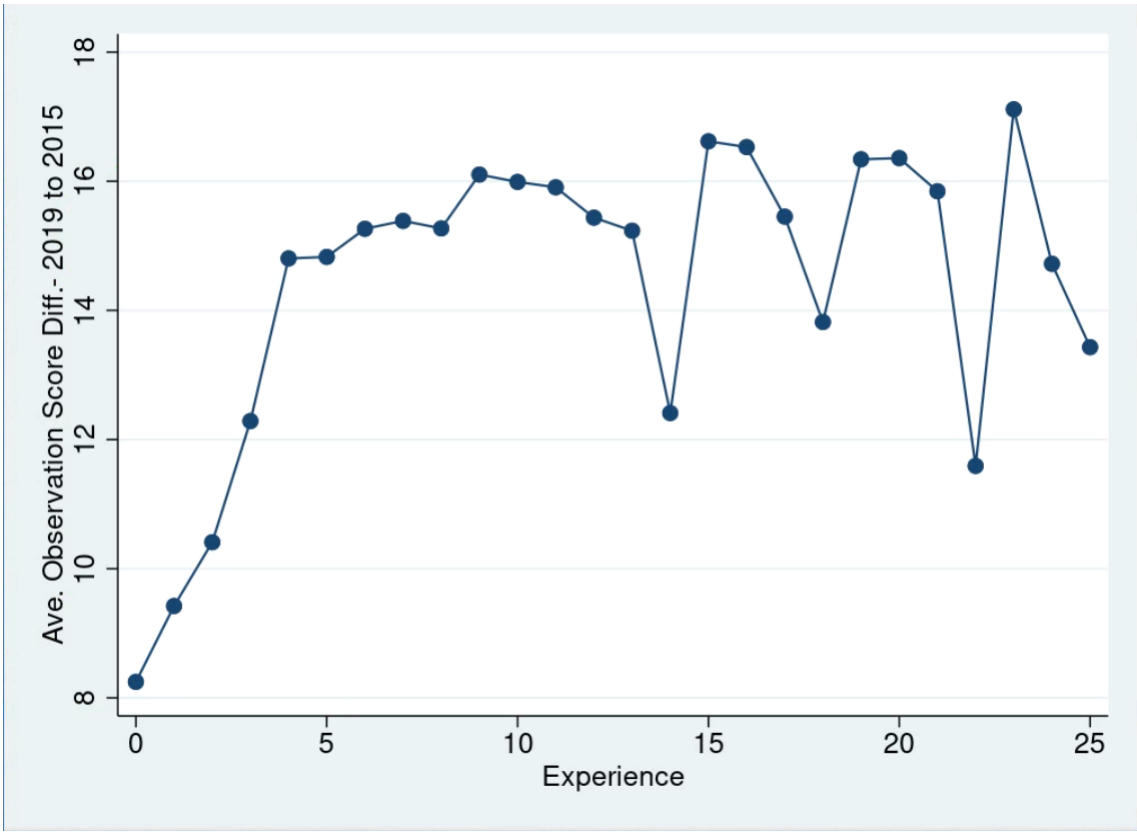


Figure 4: The difference between average observation scores in 2018-2019 and 2014-2015 for teachers in each year with a given level of experience. Each dot represents the difference in the average classroom observation score for teachers with that level of experience between the two years.

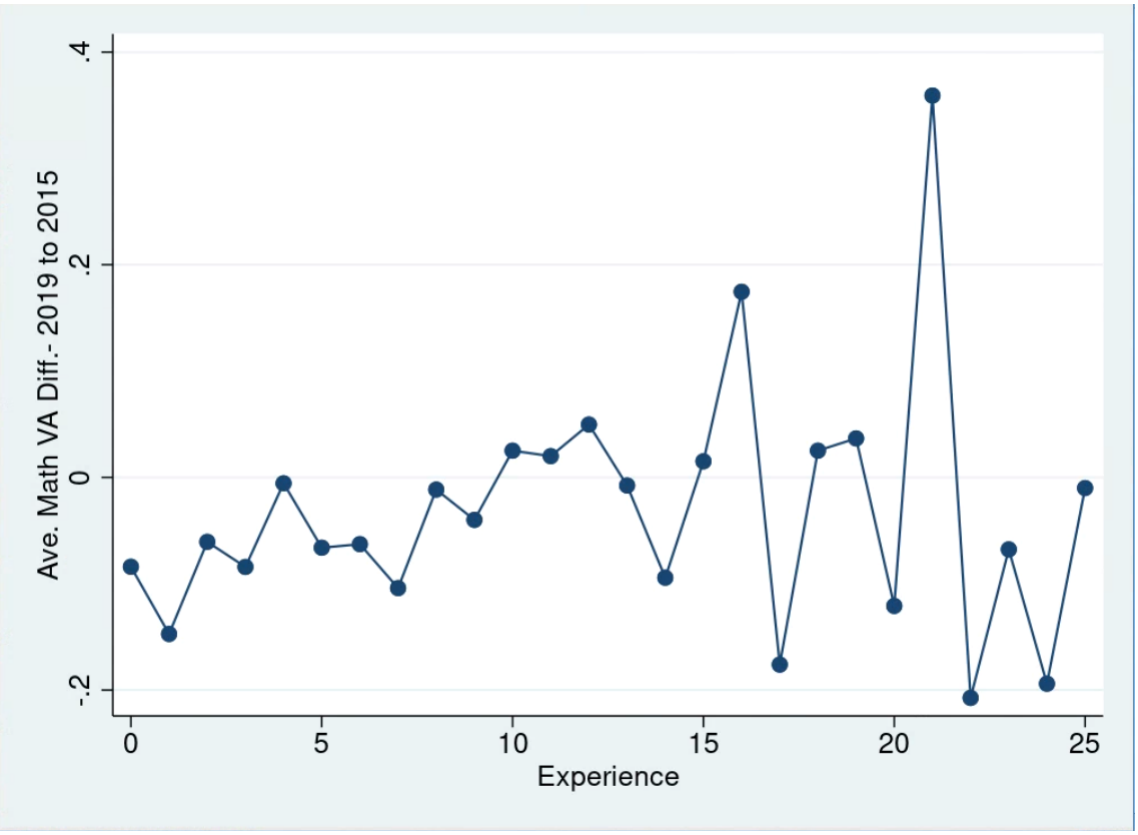


Figure 5: The difference between the average math value-added in 2018-2019 and 2014-2015 for teachers in each year with a given level of experience. Each dot represents the difference in the average math value-added score for teachers with that level of experience between the two years.

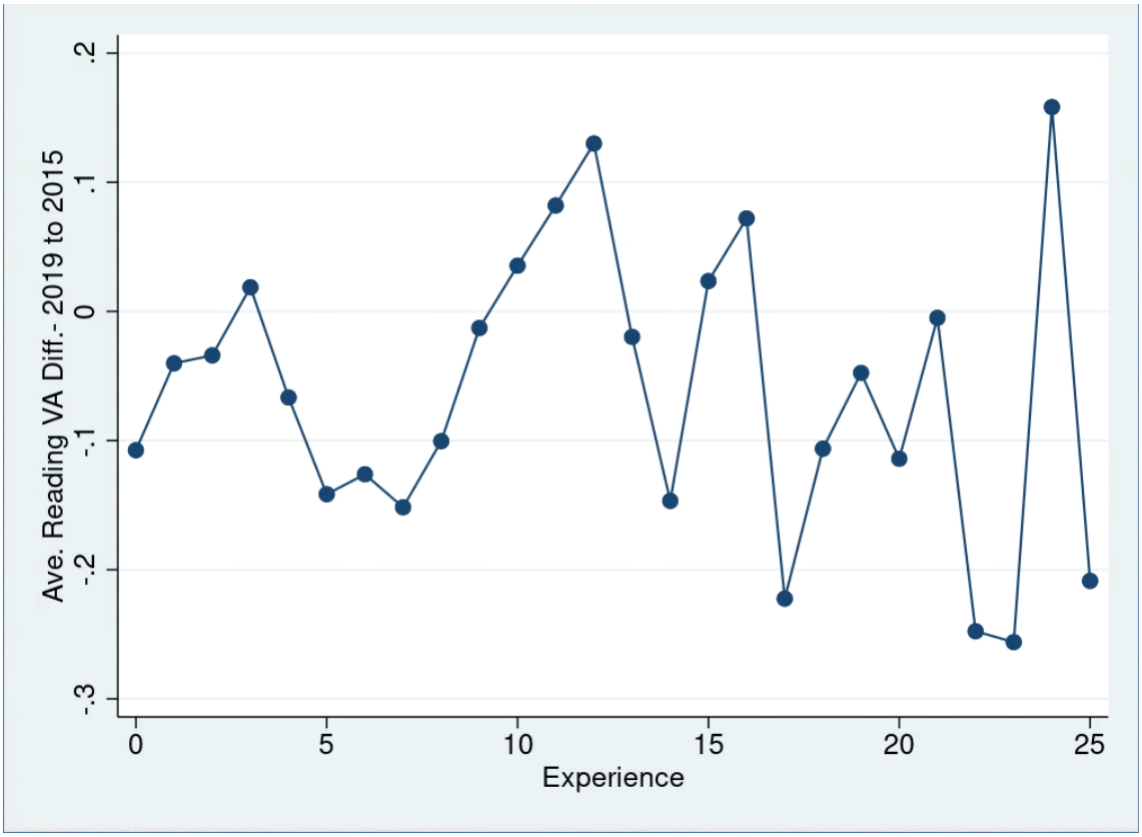


Figure 6: The difference between the average reading value-added in 2018-2019 and 2014-2015 for teachers in each year with a given level of experience. Each dot represents the difference in the average reading value-added score for teachers with that level of experience between the two years.

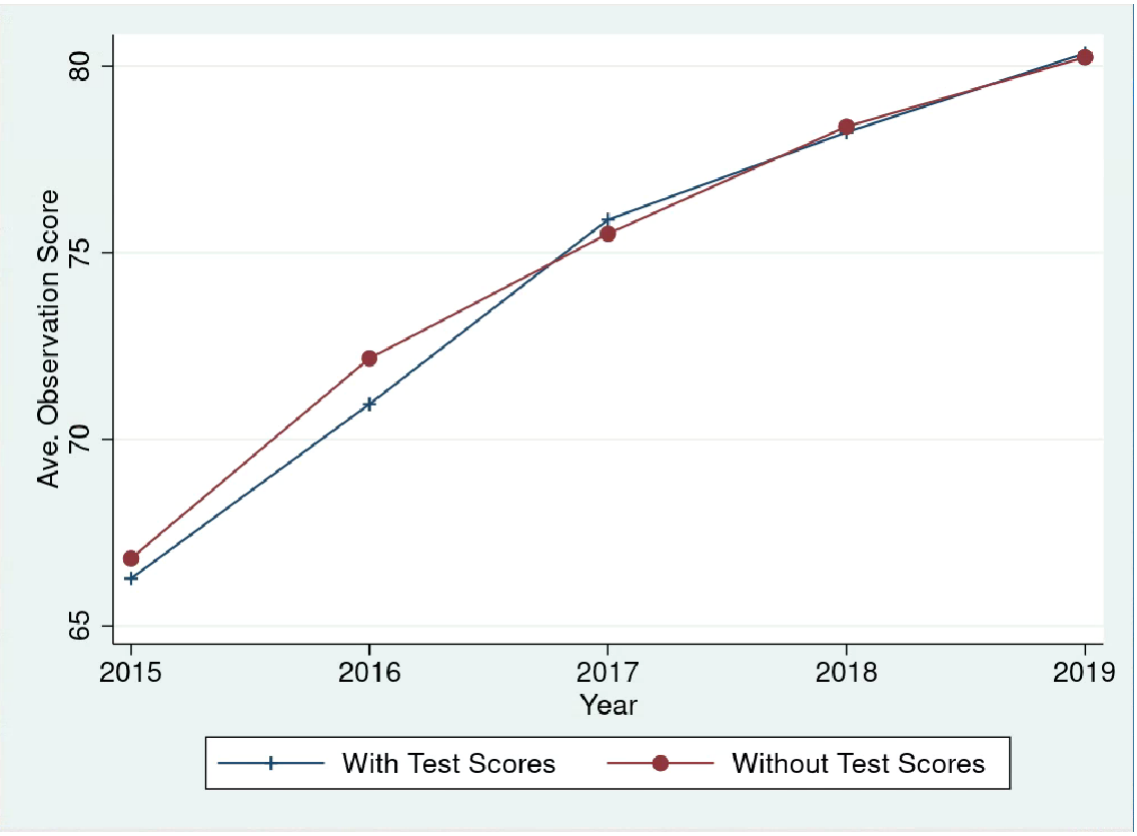
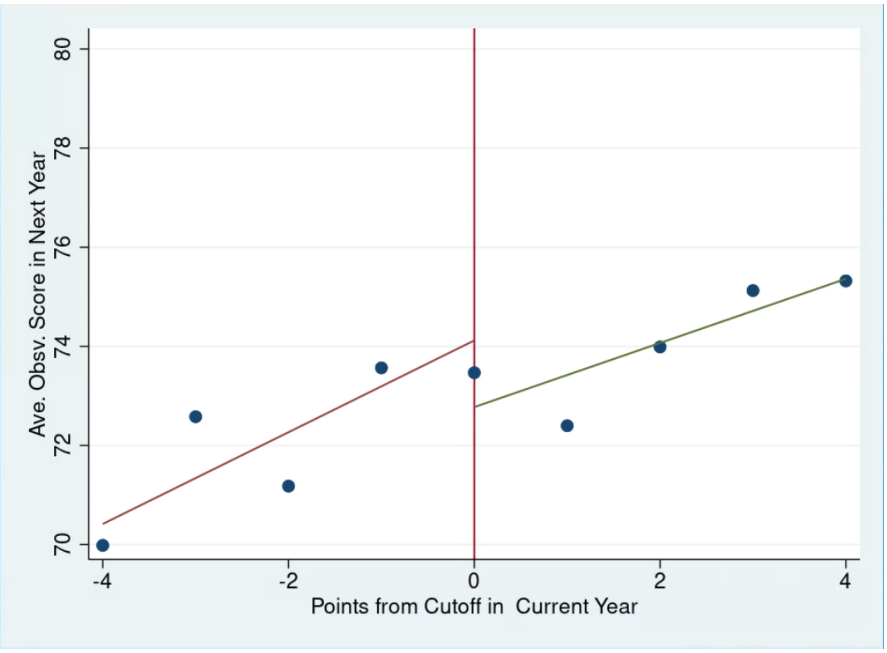
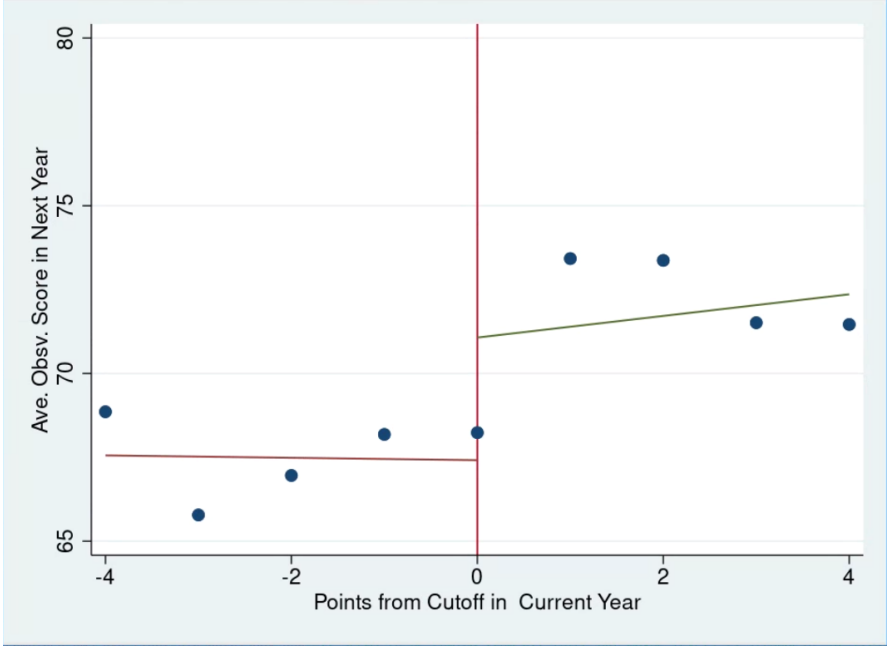


Figure 7: The trends in observation score means for teachers with and without student test scores. The red line represents teachers without student test scores and thus were not included in the principal penalty calculation. The blue line represents the trend for teachers with their own student test scores and were included.

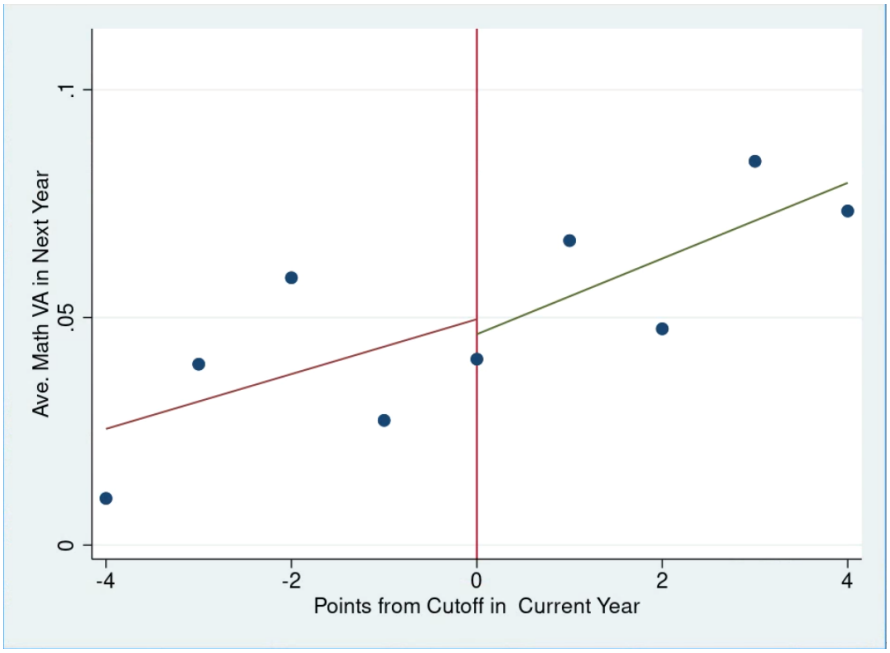


(a) With Financial Incentive

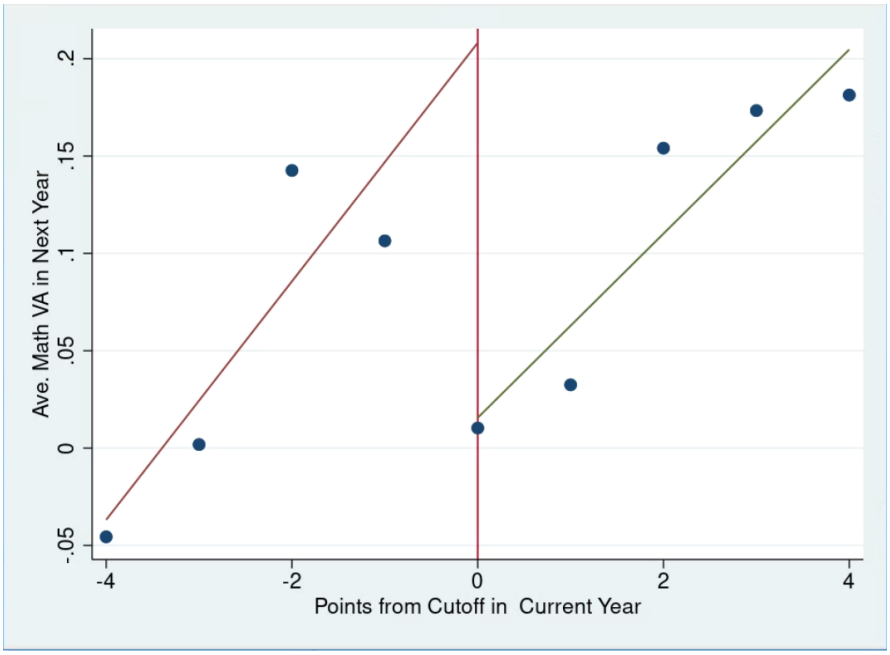


(b) Without Financial Incentive

Figure 8: Discontinuity plots of classroom observations scores in the following year around a compensation cutoff in the year prior for (a) teachers with a financial incentive to achieve the higher compensation level and (b) teachers without such an incentive. Estimates are pooled for every year of my sample period. Each dot represents the average observation score for the teachers at that points-distance from the cutoff in the year prior.

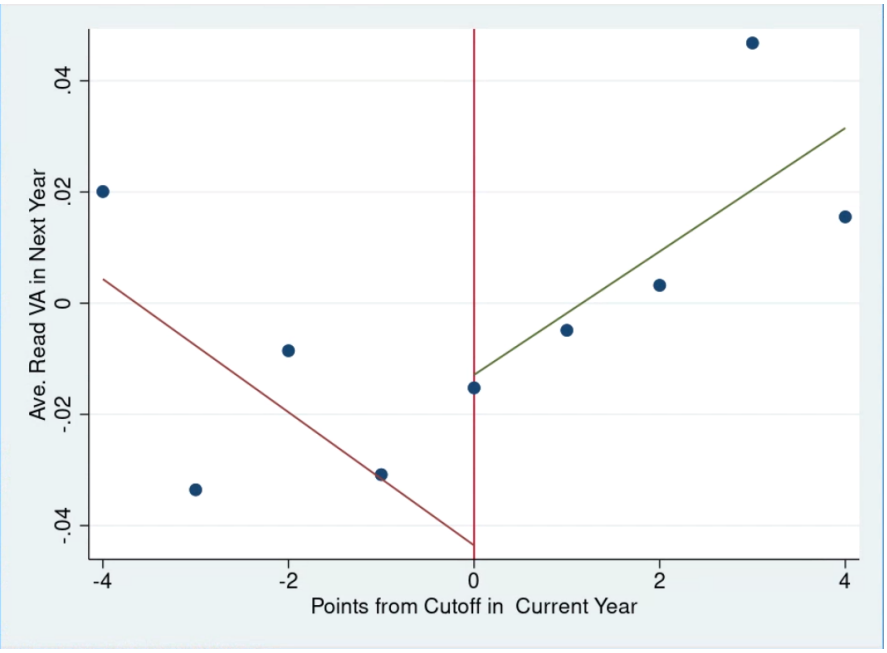


(a) With Financial Incentive

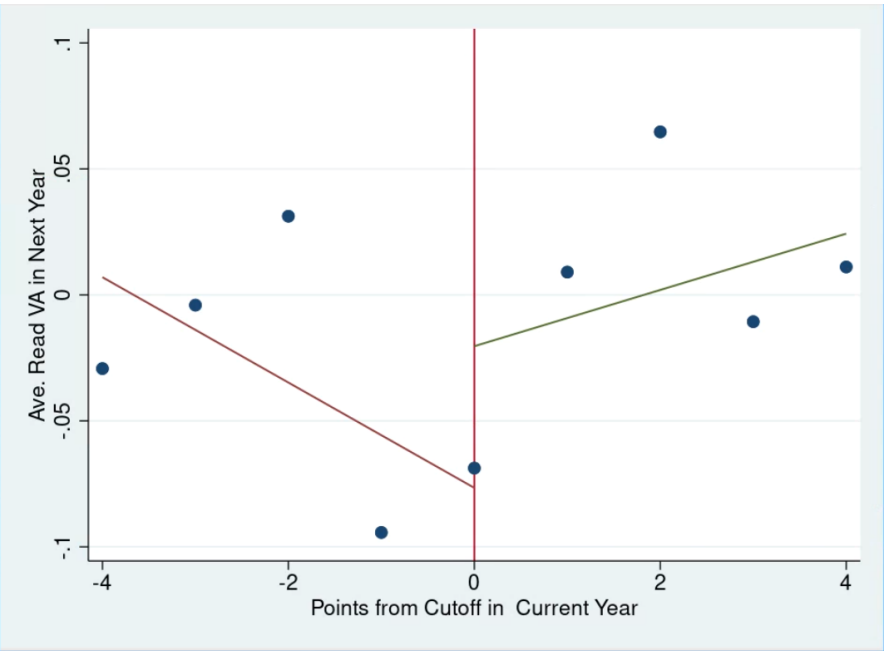


(b) Without Financial Incentive

Figure 9: Discontinuity plots of math value-added in the following year around a compensation cutoff in the year prior for (a) teachers with a financial incentive to achieve the higher compensation level and (b) teachers without such an incentive. Estimates are pooled for every year of my sample period. Each dot represents the average math value-added for the teachers at that points-distance from the cutoff in the year prior.

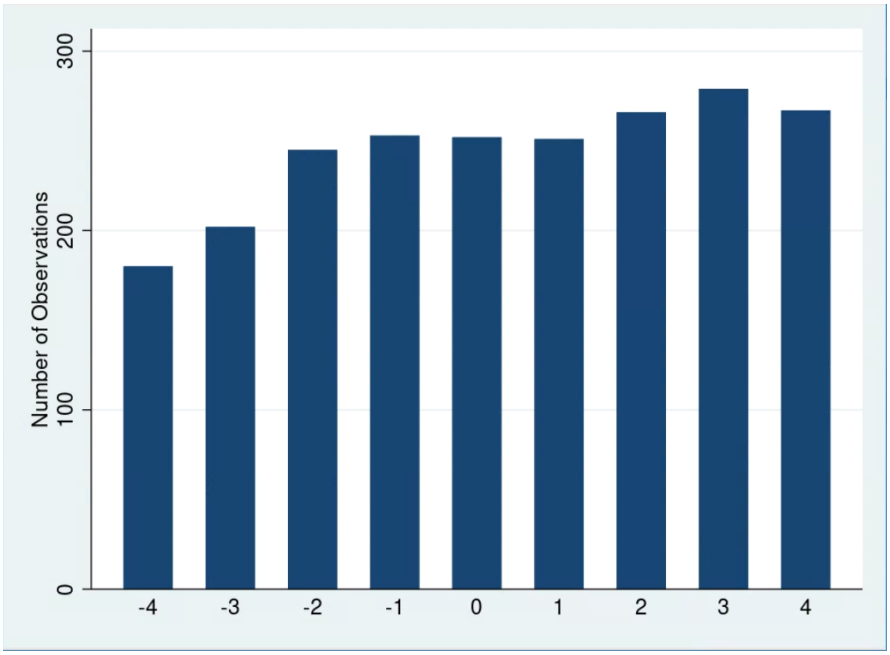


(a) With Financial Incentive

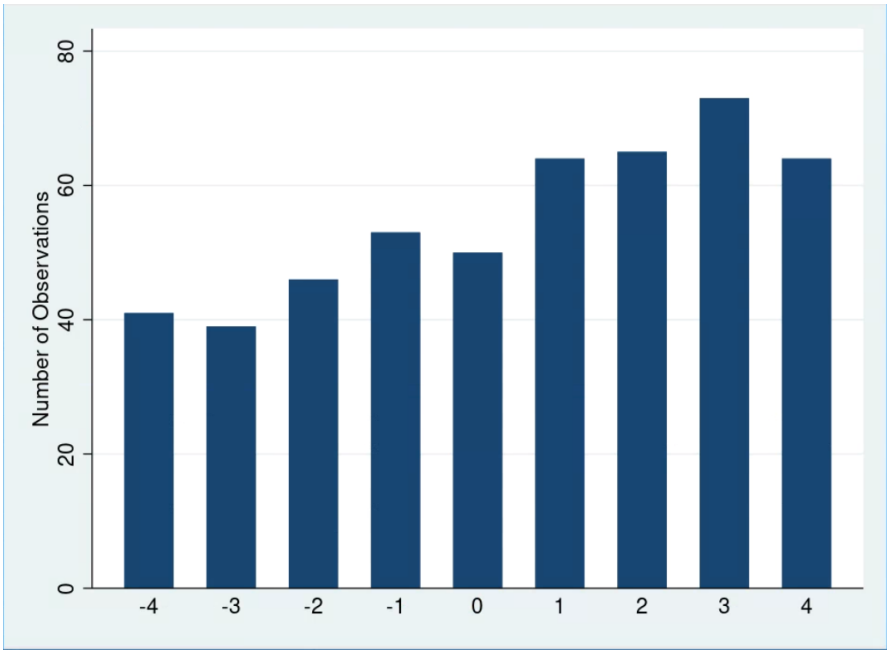


(b) Without Financial Incentive

Figure 10: Discontinuity plots of reading value-added in the following year around a compensation cutoff in the year prior for (a) teachers with a financial incentive to achieve the higher compensation level and (b) teachers without such an incentive. Estimates are pooled for every year of my sample period. Each dot represents the average reading value-added for the teachers at that points-distance from the cutoff in the year prior.



(a) With Financial Incentive



(b) Without Financial Incentive

Figure 11: Plots of the number of teacher observations around a compensation cutoff for (a) teachers with a financial incentive to achieve the higher compensation bin and (b) teachers without such an incentive. Observations are pooled for every year of my sample period. The height of each bar represents the number of teachers at that distance from a cutoff.

Tables

Table 1: Summary Statistics for Teachers

	Full Sample	Value-Added Sample
Observation Score	73.18	73.47
	(16.12)	(16.39)
# of Observations	8.033	8.082
	(2.591)	(2.624)
Experience	10.07	8.894
	(9.932)	(9.184)
Female	0.690	0.782
	(0.462)	(0.413)
Advanced Degree	0.296	0.281
	(0.457)	(0.449)
White	0.318	0.273
	(0.466)	(0.446)
Black or Af. American	0.362	0.398
	(0.481)	(0.490)
Hispanic or Lantine	0.274	0.289
	(0.446)	(0.453)
Teacher-Year Observations	29612	9611

Summary statistics represented here are for the school years 2014-2015 to 2017-2018. Standard deviations in parentheses. # of Observations represents the average number of classroom observations for a teacher in each year. Experience represents average years of professional experience. Advanced Degree represents the fraction holding a Master's degree or higher.

Table 2: Mean and Variance of Observation Scores by Year

Year	Mean	Std. Dev.	Fraction 66+	Fraction 95+
2015	66.42	15.89	0.56	0.04
2016	71.31	15.58	0.68	0.06
2017	75.75	15.41	0.75	0.10
2018	78.29	15.72	0.79	0.15
2019	80.30	15.68	0.82	0.20

Each row describes classroom observations for one school year. Fraction 66+ signifies the fraction of teachers that received an average score of 66 or higher, representing the fraction of teacher who received an average metric score in the two highest categories. Fraction 95+ represents the fraction of teachers with scores of at least 95 points out of 100.

Table 3: Distribution of Observation Score Growth by Year

Year	Mean	10%	25%	50%	75%	90%	Fract. Neg. Growth
2016	5.94	-9	-1	6	13	21	0.24
2017	5.25	-8	-1	5	12	20	0.28
2018	3.43	-9	-2	3	10	17	0.32
2019	3.20	-9	-2	2	9	17	0.32

Each row represents the change in teachers' observation scores from last school year. Fract. Neg. Growth represents the fraction of teachers who received any decrease in observation score from the year prior.

Table 4: Correlations Between Classroom Observation Scores and Value-Added

Year	Math	Reading
2015	0.216	0.174
2016	0.270	0.203
2017	0.192	0.132
2018	0.252	0.185
2019	0.300	0.213

Each row represents the Pearson's correlation coefficient between classroom observations and either math or reading value-added in each school year.

Table 5: Predicted Evaluations on Student Achievement

	Math Test Score		
	(1)	(2)	(3)
	2017	2018	2019
Pred. Obsv.	0.112***	0.120***	0.111***
	(0.0157)	(0.0183)	(0.0154)
Observations	41660	40421	43559
	Reading Test Score		
	(1)	(2)	(3)
	2017	2018	2019
Pred. Obsv.	0.0721***	0.0803***	0.0730***
	(0.0101)	(0.00944)	(0.00896)
Observations	47050	46340	47919

Standard errors in parentheses.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Each regression is run at the student level. Standard errors are clustered at the teacher level. Each regression estimates the math or reading test score for students in that year based on the predicted classroom observation score for that student's teacher, based on observations from the two prior years. Each regression also controls for student demographics including student sex, race and ethnicity, free and reduced price lunch status, special education status and limited English proficiency, as well as prior year test score, absences and disciplinary infractions.

Table 6: Difference in Differences of Observation Score by Penalty Inclusion

	(1)	(2)
	Ave. Obsv. Score	Ave. Obsv. Score
2016	-0.773	-0.738
	(0.479)	(0.474)
2017	0.879*	0.791*
	(0.467)	(0.461)
2018	0.296	0.244
	(0.471)	(0.465)
2019	0.285	0.269
	(0.472)	(0.466)
Teacher Controls	N	Y
Observations	48564	48564

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Each regression is run at the teacher-year level. Standard errors are robust. Difference in differences estimates show the difference in average classroom observation scores between teachers who were and were not included in the principal penalty calculation. Teacher controls in the second column include teacher years of experience and whether or not a teacher has a Master's degree or higher.

Table 7: Regression Discontinuity Estimates

Panel A: Teachers With Financial Incentive						
	(1)	(2)	(3)	(4)	(5)	(6)
	Obsv. Score	Obsv. Score	Math VA	Math VA	Reading VA	Reading VA
Above Cutoff	-1.663	-1.425	-0.00193	0.00820	0.0352	0.0347
	(1.220)	(1.204)	(0.0523)	(0.0519)	(0.0394)	(0.0393)
Teacher Controls	N	Y	N	Y	N	Y
Observations	2172	2172	1369	1369	1661	1661
Panel B: Teachers Without Financial Incentive						
	(1)	(2)	(3)	(4)	(5)	(6)
	Obsv. Score	Obsv. Score	Math VA	Math VA	Reading VA	Reading VA
Above Cutoff	3.656	3.556	-0.193	-0.174	0.0562	0.0725
	(2.663)	(2.664)	(0.142)	(0.143)	(0.123)	(0.118)
Teacher Controls	N	Y	N	Y	N	Y
Observations	495	495	325	325	382	382

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Each regression is run at the teacher-year level. Standard errors are robust. Regression discontinuity estimates show the difference in average classroom observation scores, math or reading value-added for teachers who just achieved a higher compensation level in the prior year. Panel A presents these estimates for teachers who faced a financial incentive around the cutoff, while Panel B presents these estimates for teachers who had grandfathered pay protection and so did not face a financial incentive. Teacher controls include teacher years of experience and whether or not a teacher has a Master's degree or higher.

Table 8: Regression Discontinuity Estimates-Varied Bandwidths

Panel A: Teachers With Financial Incentive						
	Bandwidth 5			Bandwidth 3		
	(1)	(2)	(3)	(4)	(5)	(6)
	Obsv. Score	Math VA	Reading VA	Obsv. Score	Math VA	Reading VA
Above Cutoff	-1.455	-0.0386	0.0294	-0.772	0.0219	0.00992
	(1.075)	(0.0450)	(0.0347)	(1.403)	(0.0594)	(0.0436)
Teacher Controls	Y	Y	Y	Y	Y	Y
Observations	2610	1624	1937	1731	1089	1297
Panel B: Teachers Without Financial Incentive						
	Bandwidth 5			Bandwidth 3		
	(1)	(2)	(3)	(4)	(5)	(6)
	Obsv. Score	Math VA	Reading VA	Obsv. Score	Math VA	Reading VA
Above Cutoff	3.967	-0.0464	0.0527	1.354	-0.185	0.120
	(2.416)	(0.131)	(0.106)	(3.027)	(0.176)	(0.145)
Teacher Controls	Y	Y	Y	Y	Y	Y
Observations	599	387	444	390	251	302

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Each regression is run at the teacher-year level. Standard errors are robust. Regression discontinuity estimates show the difference in average classroom observation scores, math or reading value-added for teachers who just achieved a higher compensation level in the prior year. Panel A presents these estimates for teachers who faced a financial incentive around the cutoff, while Panel B presents these estimates for teachers who had grandfathered pay protection and so did not face a financial incentive. Teacher controls include teacher years of experience and whether or not a teacher has a Master's degree or higher. Columns 1-3 show estimates for a distance of 5 around a threshold, and Columns 4-6 shows estimates for a distance of 3.

Table 9: Difference in Discontinuity Estimates

	(1) Obsv. Score	(2) Math VA	(3) Reading VA
Above Cutoff W/ Incentive	-5.011**	0.0625	0.0137
	(2.041)	(0.100)	(0.0832)
Teacher Controls	Y	Y	Y
Observations	2672	1689	2050

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Each regression is run at the teacher-year level. Standard errors are robust. Difference in discontinuity estimates show the difference in the discontinuities between teachers with and without a financial incentive around a threshold, on the average classroom observation scores, math or reading value-added for teachers who just achieved a higher compensation level in the prior year. Teacher controls include teacher years of experience and whether or not a teacher has a Master's degree or higher.

Table 10: Regression Discontinuity Estimates-Falsification Tests

Panel A: Teachers With Financial Incentive		
	(1)	(2)
	Experience	Advanced Degree
Above Cutoff	0.169	-0.0631*
	(0.442)	(0.0348)
Observations	2172	2172
Panel B: Teachers Without Financial Incentive		
	(1)	(2)
	Experience	Advanced Degree
Above Cutoff	0.396	0.134
	(1.560)	(0.100)
Observations	495	495

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Each regression is run at the teacher-year level. Standard errors are robust. Regression discontinuity estimates show the difference in professional years of experience and holding a Master's Degree or higher for teachers who just achieved a higher compensation level in the prior year.

Table 11: Observer, Observee Effects on Observation Score

	(1)
	Std. Observation Score
Total Observer Observations	0.000461***
	(0.0000330)
Total Observee Observations	0.0330***
	(0.000757)
Observer Experience	0.0241
	(0.0243)
Observee Observations by Observer	0.0132***
	(0.000543)
Observations	1313466

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Each regression is run at the teacher-observation level. Standard errors are robust. Total Observer Observations represents the total number of observations the observer who rated that teacher in that observation period has done. Total Observee Observations represent the total number of times the observee has been observed. Observer Experience controls for total years of professional experience by that observer. Observee Observations by Observer represents the total number of times an observer has rated that observee.