The Effects of Teacher Tenure on Productivity and Selection

Kevin Ng

October 19, 2021

Abstract

This study examines the productivity and selection effects of K-12 teacher tenure by leveraging variation from New Jersey's TEACHNJ Act. This law extended the pretenure period from three to four years and allowed districts to dismiss consistently low-performing teachers. I use multiple identification strategies to estimate the productivity effects of tenure across a teacher's career. I evaluate the productivity effects at tenure receipt using a difference-in-differences design, which compares fourth-year tenured and pretenured teachers. At tenure receipt, math value-added declines but English language arts value-added and summative ratings remain unchanged. To estimate the productivity effects later in the career, I use a regression discontinuity design relying on discontinuities in job security around summative rating thresholds. Later in the career, tenure has no impact on productivity. Thus, tenure induces a transitory decline in math value-added without impacting other dimensions of teacher performance. Focusing on the labor market effects, I compare teachers hired before and after TEACHNJ within the same district and experience level. The TEACHNJ Act disproportionately increased male and Black teacher turnover rates. TEACHNJ did not impact the quality of the teacher labor market as measured by value-added, though higher rated teachers often filled new vacancies. Since the TEACHNJ Act only relies on summative ratings to make personnel decisions, this result aligns with a principal-agent model where only one of several measures of performance is used to evaluate the employee.

I am grateful to the New Jersey Department of Education for assistance with the data. The conclusions of this research do not necessarily reflect the opinions or official position of the New Jersey Department of Education or the State of New Jersey. All errors are my own.

1 Introduction

Teacher tenure is a hallmark of the United States public education system; currently, 55.6% of teacher jobs are secured by tenure (National Center for Education Statistics, 2012). Tenure remains contentious because teachers are a core input into the education production function. Proponents argue that tenure attracts high-quality teachers through compensating differentials associated with increased job security. Opponents assert that tenure has become a cost-prohibitive barrier to justified dismissals, permitting teachers to shirk responsibilities while protecting low-quality educators. Much of this debate hinges on how teachers respond to tenure and how it impacts students. I provide novel evidence to address these questions.

Figure 1 shows that tenure reforms remain at the forefront of legislative action, as 49 states have combined to pass 222 tenure laws from 1994 to 2020. The constitutionality of tenure has been challenged in the court systems of California, Minnesota, New Jersey, New York, and North Carolina. These decisions rely on sparse evidence regarding tenure's impact on teacher productivity and selection. A fundamental challenge to identifying the productivity effects of tenure has been the inability to disentangle it from experience, teacher unions, human resources policies, and voluntary turnover.

Prior research has been unable to isolate early career productivity effects of tenure from the returns to experience because states rarely change the time to tenure receipt. This is a core policy variable as longer pretenure periods provide more information about teacher quality but less stability for teachers. Existing cross-sectional estimates that compare tenure across states are confounded with teacher union strength and human resources policies, such as performance pay (Jones, 2015; Roberts, 2018).

The productivity effects also may vary throughout a teacher's career. For example, tenure reduces the incentive to invest in human capital. Negative productivity effects may worsen if teachers reduce on-the-job training as they near retirement and the costs of training exceed the stream of benefits. Alternatively, negative productivity effects may diminish if effort becomes relatively less important as innate ability improves due to experience.

Understanding productivity effects of tenure across a teacher's career is critical to developing optimal tenure policy.

Little research has identified later career effects of tenure on productivity because experienced teachers rarely lose tenure protections. Without standardized tenure dismissal policies, researchers must compare tenured teachers who voluntarily left the district and lost tenure protections to those who stayed in the district. The choice to leave the district could be correlated with unobserved factors that bias the estimates.

In this paper, I use the Teacher Effectiveness and Accountability for the Children of New Jersey (TEACHNJ) Act and teacher-student linked administrative data to isolate the effects of tenure. To estimate productivity effects of tenure at tenure receipt, I leverage the TEACHNJ provision that extended the time to tenure from three to four years. Using a difference-in-differences design, I compare the change in performance of tenured and pretenured teachers with the same amount of experience to estimate early career productivity effects. By comparing teachers within the same state and experience level, I isolate productivity effects of tenure from experience, teacher unions, and human resources policies.

I evaluate the effects of tenure later in the career by using the TEACHNJ clause that allows schools to dismiss tenured teachers who received consecutive summative ratings below 2.65. After the first rating[1] below 2.65, teachers must earn a 2.65 or higher in the subsequent year to guarantee continued employment.[2] Using a continuous rating running variable, I conduct a regression discontinuity design (RDD) comparing teachers on the threshold of job insecurity to estimate the effects of tenure protections on experienced teachers. I use this policy to isolate the effects of tenure from the decision to voluntarily leave the district.

To measure the effects of tenure reform on selection through retention and sorting, I compare teachers hired before and after TEACHNJ within the same district and experience level. I estimate changes in teacher turnover rates and average performance to identify the impact of TEACHNJ on the teacher labor market. In combination with the previous strate-

[1] I use the terms "summative rating" and "rating" interchangeably throughout the paper.

[2] As discussed in Section 2, teachers may be offered a third opportunity to improve in certain situations.

gies, my results provide new insights into how tenure policies affect productivity throughout teachers' careers and how these policies alter the composition of the teacher workforce.

When estimating productivity and selection effects, I use two measures of teacher performance from the New Jersey Department of Education (NJDOE): value-added and summative ratings. This novel dataset allows me to calculate each teacher's annual value-added using a lagged test score model. The NJDOE also provides annual summative ratings, which districts use to make performance-based personnel decisions as required by the TEACHNJ Act. For example, teachers need two ratings of at least 2.65 to receive tenure, while tenured teachers may be dismissed after receiving consecutive ratings below 2.65. Every teacher receives one summative rating each year, even when teaching multiple subjects. The summative ratings combine student tests scores and classroom observations based on an NJDOE approved evaluation rubric. These rubrics capture competencies within general categories, such as lesson planning, classroom management, and professionalism. Summative ratings measure components of performance that are distinct from value-added, such as the development of non-cognitive skills. However, the scoring rubrics offer scope for evaluator discretion. Since tests provide less opportunity for subjectivity, I primarily rely on the value-added estimates.

I estimate productivity effects at tenure receipt using a difference-in-differences framework. I compare teachers hired before TEACHNJ with a three-year pretenure period to those hired after TEACHNJ with a four-year pretenure period. If tenure causes declines in productivity, the tenured fourth-year teachers hired before TEACHNJ would experience a decrease in productivity relative to the pretenured fourth-year teachers hired after TEACHNJ. The relative performances would realign in the fifth year when both sets of teachers have tenure.

The difference-in-differences design relies on standard assumptions that the effects of experience on performance are not changing for those hired before and after TEACHNJ. This assumption has two components: no confounding factors at tenure receipt and identical relative returns to experience, which are analogous to parallel trends. I create a model of tenure to elucidate potential threats to identification and suggest tests of underlying

assumptions. As discussed in Section 2, TEACHNJ contained other provisions that increased the difficulty to receive tenure and weakened tenure protections. According to the model, these components would bias the results if they changed performance standards. I provide evidence of similar standards by demonstrating comparable retention rates before and after TEACHNJ. Thus, I find no evidence that the other components of TEACHNJ bias the estimates. In fact, the results are robust to a variety of specifications that account for these potential confounding factors. My equations also include teacher fixed effects that account for level shifts in performance, though differential returns to experience between the groups could remain problematic. I show direct evidence of identical relative returns to experience in the first three years before and after TEACHNJ, which supports my empirical design.

At tenure receipt, teachers experience a decline of 0.033 standard deviations of math value-added but no change in English language arts (ELA) value-added. Tenure also has no impact on summative ratings. The 95% confidence intervals rule out decreases larger than 0.015 ELA standard deviations and 0.046 rating points on a scale of 1.00 to 4.00. The effects are concentrated among wealthier schools with few minority students and high proficiency rates. Focusing on the math value-added decline, partial equilibrium estimates from Chetty et al. (2014) equate this shock to a $237 present value loss per student. However, these point estimates remain smaller than the returns to the first four years of experience, which are 0.079 standard deviations. In the short-run, replacing teachers due to these productivity effects would remain counterproductive on average, unless schools can find novice teachers whose initial math value-added is at least 0.046 standard deviations higher (the difference between the productivity effects of tenure and the returns to experience). The long-run effects of this reform or alternative policies, such as the elimination of tenure, depend on the persistence of the effects. If the productivity effects are temporary, these policies will generate few gains with potentially large negative impacts on selection into teaching. I estimate later career productivity effects to evaluate whether these shocks linger.

I estimate the effects of tenure on more experienced teachers by leveraging dismissal

threats. TEACHNJ allows districts to dismiss tenured teachers who earn consecutive ratings below 2.65. By removing job security, districts eliminate the benefits of tenure among experienced teachers. After receiving a rating below 2.65, tenured teachers have a full year to improve and regain job security. This environment creates discontinuities in job security among similarly-rated teachers. I use an RDD relying on a continuous summative rating running variable to evaluate these productivity effects.

The RDD shows that teachers on the margin of job insecurity do not increase their productivity when receiving a dismissal threat. As a result, the negative productivity effects of tenure diminish later in the career. Given the temporary nature of the math value-added decline, any weakening of tenure protections would have limited scope to improve productivity. Although policymakers would like to eliminate any negative productivity shocks, they also must consider the labor market impacts of these tenure reforms.

Finally, I estimate the selection and retention effects of TEACHNJ by comparing teachers hired before and after the law with the same amount of experience. This strategy identifies the labor market effects of weakening tenure protections, though the policy variation does not allow me to estimate the effects of completely eliminating tenure. Since weakening tenure protections (without pay increases) would make teaching jobs less desirable, teachers may exit the profession for more lucrative opportunities. Tenure reforms also may increase dismissal rates among low-performing teachers and increase average teacher quality. In fact, TEACHNJ increased turnover among "effective" and "ineffective" teachers by 4 and 22 percentage points, respectively. While the law removed low-performing teachers, it is important to consider the consequences of this turnover on diversity.

TEACHNJ disproportionately increased male and Black teacher turnover, as they received lower average summative ratings. However, these teachers produced similar value-added to female and White teachers. Although value-added and ratings capture different components of teacher effectiveness, male and Black teachers only performed less effectively along the one dimension of performance with scope for subjectivity: summative ratings. In

fact, biases may influence these ratings because male and Black teachers receive lower ratings when paired with principals of other genders and races. Specifically, the disparities increase by 0.007–0.009 points when male and Black teachers are evaluated by female and White principals, respectively. As a result, tenure reforms tied to subjective evaluation criteria may have unintended consequences on teachers who are underrepresented in the profession. This effect on diversity also impacts students because academic achievement improves when male and Black students are paired with teachers of the same gender (Dee, 2007) or race (Gershenson et al., 2018; Dee, 2004; Egalite et al., 2015).

Despite increases in turnover, TEACHNJ had no effect on average teacher value-added. However, the law attracted new teachers whose summative ratings were 0.021 points higher. Since TEACHNJ included tenure reforms tied to ratings, the ratings of new hires responded accordingly. Average performance improved along the dimension that dictated personnel decisions (ratings) but remained unchanged along other dimensions (value-added). This result aligns with a principal-agent model where only one of several measures of performance is used to evaluate the employee (Holmstrom & Milgrom, 1991; Baker, 2002).

Overall, my results show that tenure reforms offer a tradeoff between productivity and diversity. Tenure generates temporary declines in math value-added, while the TEACHNJ Act disproportionately increased male and Black teacher turnover. Given the unintended consequences on diversity along with limited productivity and selection effects, the efficacy of tenure reforms to improve teacher performance is limited.

This paper contributes to the literature by estimating the productivity effects of tenure. Understanding these effects is critical to making personnel decisions and deciding between less incentivized tenured teachers and inexperienced novices. Few papers have estimated these productivity effects because it is difficult to disentangle tenure receipt from non-linear returns to experience (Kraft & Papay, 2015; Wiswall, 2013). Alternative cross-sectional analyses comparing tenure receipt among teachers with the same experience across states are confounded with varying teacher union strength and human resources policies (Jones, 2015;

Roberts, 2018). Using the TEACHNJ Act, I overcome these sources of bias by comparing teachers within the same state and experience level. These models only rely on standard difference-in-differences and RDD assumptions, which I provide evidence to support.

This study additionally is related to research that estimates the productivity effects of dismissal threats. These analyses estimate the impact of dismissal threats on the performance (Dee & Wyckoff, 2015) and absenteeism (Jacob, 2013) of non-tenured teachers. I add to the literature by evaluating the productivity effects among experienced, tenured teachers.

This paper also estimates the labor market effects of tenure reforms. Previous research finds short-term improvements in average teacher quality by weakening tenure protections (Carruthers et al., 2018; Loeb et al., 2015; Anderson et al., 2019; Rodriguez, 2018; Kraft, 2015; Goldhaber et al., 2016). These policies increase turnover among low-performing teachers, which has been shown to improve student achievement (Adnot et al., 2017). However, these policies may eventually cause high-quality teachers to sort into other professions (Strunk et al., 2017). The negative impacts on student performance are exacerbated when accounting for disruptions to the teaching staff, such as lost teaching experience, due to increased turnover (Ronfeldt et al., 2013; Hanushek et al., 2016; Sorensen & Ladd, 2020). To offset these negative effects, several structural papers find that tenure elimination needed to be accompanied by large salary increases (Rothstein, 2015; Johnston, 2018). While I estimate similar effects on teacher quality and retention, I advance the literature by leveraging multiple dimensions of performance. This allows me to measure distortions in performance when job security becomes tied to summative ratings but not value-added.

Finally, this paper estimates the effect of tenure reforms on diversity by considering differential turnover rates by teacher characteristics. While previous work finds male and Black teachers earn lower ratings (Bailey et al., 2016; Drake et al., 2019; Sartain & Steinberg, 2020; Chi, 2021), my analysis connects disproportionate increases in turnover to these rating disparities.

2 TEACHNJ Act

This study relies on variation from the TEACHNJ Act, which passed on August 6, 2012 as a means to improve student achievement. The law lengthened the pretenure period from three years to four years, introduced a teacher mentor program, reformed teacher evaluation criteria, and modified the tenure appeals process. The TEACHNJ Act was New Jersey's largest teacher tenure reform since it passed comprehensive tenure laws in 1909. The teacher union immediately notified its members about the law, which includes 97.6% of New Jersey public school teachers (National Center for Education Statistics, 2012).

The TEACHNJ Act extended the pretenure period from three to four years for those hired on or after August 6, 2012.[3] To receive tenure, teachers hired after TEACHNJ must receive at least two "effective" ratings in their second to fourth years. For those hired before TEACHNJ, each district designed its own requirements for tenure receipt. Teachers are notified about these tenure decisions by May 15 of their final pretenure year.

During their first year of teaching, novice teachers hired after TEACHNJ receive a mentor to provide feedback, support, and opportunities for modeling. The mentor must be an experienced, "effective" teacher who completed a mentor training program.

Teacher evaluation criteria[4] evolved from a two-tiered system (satisfactory or unsatisfactory) to a four-tiered system (highly effective, effective, partially effective, or ineffective). Previously, evaluation criteria varied by school. Now, evaluations rely on a combination of Teacher Practice, Student Growth Objectives (SGO), and median Student Growth Percentiles (mSGP).[5] Teachers rated as ineffective or partially effective must create a Corrective Action Plan (CAP) with their supervisors. Each CAP includes specific demonstrable goals, timelines for improvement, and responsibilities of the teacher and school.

Tenured teachers rated ineffective or partially effective for consecutive years may receive

[3] Appendix Section A.1 shows that three-year tenure clocks are common throughout the United States.

[4] In Appendix Section A.2, I discuss the implementation of this evaluation system reform.

[5] Summative ratings for grades 4 to 8 ELA and grades 4 to 7 math teachers rely on Teacher Practice, SGO, and mSGP, while ratings for other teachers only depend on Teacher Practice and SGO.

a charge of inefficiency. The teacher may be offered a third opportunity if the second rating is partially effective. However, the teacher will always receive a charge of inefficiency if they receive a third consecutive partially effective or ineffective rating. The teacher's employment is then subject to an arbitration process of no more than 48 days. Previously, there was no time limit for the process.

3 Data

I use the NJDOE's teacher-student linked administrative test score data from 2012 to 2018. These math and ELA tests include the New Jersey Assessment of Skills and Knowledge (NJASK) for Grades 3 to 8 from 2012 to 2014, the High School Proficiency Assessment (HSPA) for grades 11 to 12 from 2012 to 2014, and the Partnership for Assessment of Readiness for College and Careers (PARCC) exam for grades 3 to 11 from 2015 to 2018.

These data also include student gender, race, Free or Reduced-Price Lunch (FRPL) eligibility, English Language Learner (ELL) status, and special education status. In addition, the dataset contains teacher gender, race, and experience.[6] The data lack tenure status information, though districts dismiss nearly every teacher who is not offered tenure after the pretenure period. Therefore, I mark tenure receipt as an indicator for remaining in the same district for four years if hired before TEACHNJ or five years if hired after TEACHNJ.

I calculate annual value-added using the following regression:

$$A_{ijgst} = \alpha A_{it-1} + \beta X_{it} + \eta C_{it} + \lambda S_{it} + \Theta_{jt} + \varepsilon_{ijgst} \tag{1}$$

where A_{ijgst} is the test score of student i in teacher j's grade g class in school s and year t.[7] I control for the student's previous year test score (A_{it-1}), as well as student, classroom, and school characteristics. The student variables (X_{it}) include gender, race, FRPL eligibility,

[6] Table A1 provides summary statistics for students (first column) and teachers (second column). These statistics match expectations given New Jersey's demographic composition and national proficiency rates.

[7] Each grade-year exam is standardized to have mean 0 and standard deviation 1.

ELL status, and special education status. The classroom controls (C_{it}) are class size and aggregated student controls. School covariates (S_{it}) include urbanicity[8], enrollment, racial composition, and percentage of FRPL eligible. Value-added is measured annually by Θ_{jt}.

Summative ratings from 2014 to 2018 capture components of performance that are distinct from value-added. Figure 2 shows the correlation between value-added and summative ratings in Panels B and C are much weaker than the correlation between math and ELA value-added in Panel A. In fact, the correlation coefficient between math and ELA value-added (0.522) is over four times larger than those between value-added and summative ratings (0.118–0.128).

These ratings measure performance using a weighted average of Teacher Practice, SGO, and mSGP. In Teacher Practice, supervisors observe classes using an NJDOE approved rubric. These rubrics evaluate various teaching competencies, such as lesson planning, classroom management, and professionalism. Each district designs its own SGO and scores them by the percentage of students passing the objective. Grades 4 to 8 ELA and grades 4 to 7 math teachers rely on mSGP, which are based on state assessments. The mSGP differ from value-added because they only account for previous test scores rather than a variety of student, classroom, and school characteristics.[9]

Table A2 shows the weighting schemes for 2014 and 2017–2018 (first two columns), as well as 2015–2016 (last two columns).[10] Ratings primarily rely on Teacher Practice with some weight placed on test scores. The odd columns record the weights for subjects that partially rely on state tests. The even columns show the weights for other subjects. Based on these weights, teachers receive a summative rating between 1.00 and 4.00. These ratings place teachers into one of four categories with minimum thresholds included in parentheses: ineffective (1.00), partially effective (1.85), effective (2.65), and highly effective (3.50).

[8] I merge urbanicity data from the National Center for Education Statistics (2018) using the crosswalk from the New Jersey Department of Education (2017).

[9] Betebenner (2011) provides a detailed description of the Student Growth Percentile methodology.

[10] In 2015 and 2016, the NJDOE placed less weight on mSGP to give educators time to acclimate to the new PARCC assessments (Shulman, 2016).

4 Productivity Effects at Tenure Receipt

To estimate the productivity effects of tenure at tenure receipt, I use the following difference-in-differences framework:

$$y_{jtc} = \gamma ten_{jt} + \sum_{\tau=1}^{T} \delta_\tau \mathbf{1}(exp_{jt} = \tau) + \psi_j + \nu_c + \mu_t + \varepsilon_{jtc}. \quad (2)$$

In equation (2), y_{jtc} is teacher j's performance in year t at school c as measured by value-added or summative ratings.[11] I include an indicator for tenure status (ten_{jt}), as well as experience ($\mathbf{1}(exp_{jt} = \tau)$) and teacher ($\psi_j$) fixed effects. I only include school (ν_c) and year (μ_t) fixed effects in the rating regressions. The coefficient of interest (γ) would be negative if tenure eliminates some effort incentives. I restrict the analysis to math and ELA teachers with non-missing value-added in years 2 and 3 to guarantee multiple observations of pretenure performance.[12] I cluster standard errors at the school level to account for autocorrelation in the residuals generated by each principal's implementation of TEACHNJ.[13]

4.1 Productivity Effects at Tenure Receipt Assumptions

Equation (2) relies on the assumption that the effects of experience on performance are not changing for those hired before and after TEACHNJ. This assumption has two components: no confounding factors at tenure receipt and identical relative returns to experience, which are analogous to parallel trends. Since the difference-in-differences model compares teachers hired before and after TEACHNJ, the other components of the policy could confound the estimated effect. As a result, γ in equation (2) would capture the productivity effects of tenure in addition to differences in tenure receipt standards and the mentor program. Section 4.2 addresses the core of these confounding factors. These components may also generate

[11] I simplify the value-added model to one step as described in Appendix Section A.3.

[12] The rating samples also are limited to teachers with non-missing ratings in years 2 and 3. In Table A3, I record the number of teacher observations remaining after restricting the sample. Limiting the sample to teachers with non-missing value-added in years 2 and 3 has the largest effect on sample size.

[13] The standard errors are not sensitive to clustering at the district level.

differences in the relative returns to experience across cohorts that would bias the estimated γ. Section 4.3 addresses this concern.

These assumptions are much weaker than those used in previous studies. Typically, experience and tenure receipt are collinear, making it impossible to isolate the productivity effects of tenure when using fixed effects to flexibly control for non-linear returns to experience. Instead, these studies must rely on strong functional form assumptions regarding these returns. Cross-sectional variation using cross-state comparisons are confounded with teacher unions[14] and human resources policies, such as performance pay (Jones, 2015; Roberts, 2018). Leveraging the TEACHNJ Act, I overcome the fundamental challenge to identifying the productivity effects of tenure: the inability to disentangle it from experience, teacher unions, human resources policies, and voluntary turnover.

4.2 No Confounding Factors at Tenure Receipt

To test for confounding factors at tenure receipt, it is necessary to address the other components of TEACHNJ. First, I consider the tenure dismissal and appeals reform. Since all observations occur after 2012, every teacher in the sample encountered these changes. Thus, equation (2) estimates the effects of the new, weaker tenure protections rather than the old, stronger protections. The standardized tenure removal procedure and the streamlined appeals process can only bias my estimates if teachers reacted differently to the law over time. This is unlikely because the union immediately informed teachers about the law.

The other potentially confounding aspects of the reform are standardized tenure receipt and the mentor program. These components differentially impacted those hired before and after TEACHNJ, so I develop a model of tenure to elucidate this threat to identification and suggest tests of underlying assumptions.

[14] According to National Center for Education Statistics (2012), tenure and union membership are correlated, as 63.5% of teacher union members have tenure. However, they remain distinct, as 33.3% of non-union members also have tenure. Cowen and Strunk (2015) find much of the teacher union literature suggests null to slightly negative effects on student achievement (Hoxby, 1996; Marianno & Strunk, 2018). Similarly, Lovenheim and Willén (2019) find teacher unions have negative effects on student labor market outcomes.

4.2.1 Model

Let a_t be the teacher's ability in experience year t. Suppose a_t is an exogenously given, increasing, and concave function of t that may vary by initial ability. Assume student test scores in year t are given by $p_t = a_t + e_t$, where e_t is teacher effort. Define teacher utility as a strictly concave function ($u(a_t, e_t) \equiv u_t(e_t)$) where e_t^* is its unique global maximum.[15]

I add an employment contract where retention depends on previous performance. Suppose teachers may be dismissed for poor annual performance. Teachers without tenure are dismissed if $p_t < n$, while teachers with tenure are dismissed if $p_t < y$. Since tenured teachers have greater job security, let $n > y$. Also, assume teachers are offered tenure in year T if $\sum_{\tau=1}^{T} \frac{p_\tau}{T} \geq r$ and dismissed otherwise. In words, teachers must perform above a minimum average level during their pretenure years to receive tenure in year T. For the tenure receipt requirements to impact performance, $r > n$. Otherwise, any teacher dismissed at tenure receipt would have already been dismissed in an earlier year.[16]

In year t, teachers are offered continued employment if:[17]

$$m_t = \prod_{\tau=1}^{T} \mathbf{1}(p_\tau \geq n) \mathbf{1}\left(\sum_{\tau=1}^{T} \frac{p_\tau}{T} \geq r \right) \prod_{\tau=T+1}^{t-1} \mathbf{1}(p_\tau \geq y).$$

Suppose teachers have an outside option that produces utility level $f(a_t, t)$ for each of the remaining periods. Teachers who are dismissed or quit after year t receive utility $\sum_{\tau=t+1}^{\overline{T}} f(a_\tau, \tau)$, where $\overline{T}$ is the retirement experience level.

I find a closed form solution to the model in Appendix Section A.4. The utility function dictates a single period optimal level of effort. Teachers deviate from this optimal effort level if they would otherwise fail to meet the performance standards and the future stream of

[15] I model performance and utility as additively separable, though the results are similar when I allow initial ability (a_1) and effort (e_t) to interact multiplicatively.

[16] For computational simplicity, this setup varies slightly from the TEACHNJ Act. The law required teachers to receive two effective ratings prior to tenure receipt. In practice, 40% of ineffective and partially effective pretenured teachers immediately lost their jobs even if it was several years prior to tenure receipt. Thus, the combination of n and r are a close proxy for both the law and its actual implementation.

[17] Pretenure teacher job security ($t \leq T$) only relies on the first product, while teacher job security in experience year $T + 1$ only relies on the first two products.

utility from teaching exceeds that from the outside option. As a result, high-quality teachers exert the single period optimal level of effort (e_t^*) in every period and maintain job security. Teachers near the tenure receipt threshold increase their early career effort to receive tenure. Lower quality teachers meet minimal annual performance standards but fail to receive tenure. The lowest-quality teachers leave the profession during the pretenure period.

4.2.2 Potential Biases and Testable Implications

When mapping TEACHNJ's standardized tenure receipt to increases in the tenure receipt requirements (r), the model identifies r as a potential source of bias.[18] These requirements may be more stringent for those hired after TEACHNJ, which would increase effort and performance in the first three years relative to the fourth year. The teacher fixed effects may overestimate teacher quality for those hired after TEACHNJ and attribute some of the fourth-year pretenure effort to innate ability. In the difference-in-differences design, this will underestimate the pretenured group's effort relative to that of the tenured group and bias the estimated productivity effects upward.

To directly test for changes in minimum performance standards, Figure 3 shows retention rates by experience year for those hired before (solid black) and after (hollow red) TEACHNJ when controlling for ratings.[19] Those hired after TEACHNJ were retained less frequently, though the estimates are statistically indistinguishable from one another.[20] This suggests the performance standards became slightly stricter.

Next, I directly test for bias by simulating the model based on these implied stricter performance standards. Initial teacher ability follows a standard normal distribution and ability increases by one standard deviation over 24 years before retirement. Using backward induction, I solve the model and set performance standards to reflect retention rates in

[18] I connect the standardized removal and streamlined appeals to increases in annual standards (n and y). However, n and y would not bias the estimates because they did not differentially change for those hired before and after TEACHNJ.

[19] I control for ratings rather than value-added because ratings dictate personnel decisions.

[20] Relative to other pretenure years, the first year retention rate for those hired after TEACHNJ is higher because TEACHNJ ignores first year ratings for retention decisions.

New Jersey. I set the annual and tenure receipt standards to match actual turnover rates. Otherwise, the values of n, y, and r do not have a simple interpretation. I set $n = 2.7$, $y = 2.2$, and $r = 3.05$, while I set $T = 3$ to reflect the three-year pretenure period before TEACHNJ. These parameters result in 30% of new hires failing to receive tenure. After TEACHNJ, I do not change n and y because the standardized tenure removal procedure and streamlined appeals process remained unchanged for those hired before and after TEACHNJ. Instead, I set $T = 4$ and $r = 5.45$ after TEACHNJ to simulate the worst case bounds of 54% of new hires failing to receive tenure.[21]

Using the simulated data and these worst case parameters, I estimate equation (2). The results are presented in Table 1. First, I simulate the data when only extending T from three to four years. I show the results in Column (1), which estimates the productivity effects of tenure. Second, I simulate the data when changing T from three to four years and increasing the tenure receipt standards. In Column (2), I present the estimated productivity effects of tenure that are biased by the change in performance standards. The estimated effects in Column (2) are only biased upward by 0.012 units relative to Column (1). This difference is less than 0.4% of the true value, so the bias is unlikely to meaningfully impact the estimates.

In fact, the increased performance standards theoretically have limited scope to bias the results. The estimated productivity effects primarily capture the difference between the performance standards and innate ability of those who remain in the district for four years.[22] The increased standards will induce some lower quality, marginal teachers to leave the profession prior to tenure receipt. As a result, the large performance declines associated with removing incentives from these lower quality teachers would not impact the estimates. In turn, some higher quality educators will become marginal teachers who may not receive tenure. Since the new marginal educators have higher innate ability, the estimated productivity declines following tenure receipt capture the difference between the new higher standards and the more talented marginal teachers. This difference is similar in size to the effects calculated

[21] Appendix Section A.4.1 provides a complete description of the data simulation.
[22] The difference also considers the optimal single period effort level.

in Column (1) based on lower standards and less talented marginal teachers.

4.3 Identical Relative Returns to Experience

Equation (2) also relies on identical relative returns to experience before and after TEACHNJ. The returns to experience may vary for those hired before and after the law due to standardized tenure receipt and the mentor program. This would bias the estimated effect. To evaluate this assumption, I estimate the returns to experience using the following equation:

$$y_{jtc} = \sum_{\tau=2}^{5} \beta_\tau \mathbf{1}(exp_{jt} = \tau) * (1 - post_{jt}) + \sum_{\tau=1}^{5} \delta_\tau \mathbf{1}(exp_{jt} = \tau) * post_{jt} + \psi_j + \nu_c + \mu_t + \varepsilon_{jtc}. \quad (3)$$

In this model, y_{jtc} is the annual performance measure (value-added or summative rating), while $post_{jt}$ is an indicator for being hired after TEACHNJ.[23] I include teacher fixed effects (ψ_j) in all specifications. I only include school (ν_c) and calendar year (μ_t) fixed effects in the rating regressions. I limit the sample to teachers in their first five years of teaching to focus on the pretenure period and tenure receipt. If $\delta_1 = \delta_2 = \delta_3$, then I have evidence of identical relative returns to experience for those hired before and after TEACHNJ. In Section 4.4, I graph β_τ and δ_τ from equation (3) and connect them to my main results.

4.4 Productivity Effects at Tenure Receipt Results

Following equation (3), Figure 4 shows the event study graphs for math value-added (Panel A), ELA value-added (Panel B), and summative ratings (Panel C). The solid black dots measure the returns to experience before TEACHNJ (β_τ), while the hollow red dots show the returns after TEACHNJ (δ_τ). The relative returns to each experience year remain nearly identical, except for math value-added in year 4 when those hired before TEACHNJ receive

[23] Teachers may be hired before and after TEACHNJ if they switch districts, so $post_{jt}$ may vary within teachers. However, all results are robust to making $post_{jt}$ time invariant by restricting the sample to teachers in their first district (not shown).

tenure and appear to experience a productivity decline. The identical relative returns to the first three years of experience provide additional evidence that standardized tenure receipt and the mentor program do not threaten identification.

Figure 4 suggests a decline in math value-added for tenured teachers hired before TEACHNJ in year 4 but no change in the other performance measures. I corroborate these findings in Table 2 where I estimate equation (2) using math value-added (Panel A), ELA value-added (Panel B), and summative ratings (Panel C). In Column (1), I estimate the productivity effects at tenure receipt among all teachers. Math value-added declines by 0.033 standard deviations, while ELA value-added and ratings remain unchanged. In fact, the 95% confidence intervals rule out productivity declines larger than 0.015 ELA standard deviations and 0.046 rating points. As seen in brackets, the point estimates are -0.106, 0.025, and -0.057 teacher standard deviations of math value-added, ELA value-added, and summative rating points, respectively.[24] The productivity effects are smaller than the returns to the first four years of experience, which are 0.079 standard deviations, 0.043 standard deviations, and 0.19 points for math value-added, ELA value-added, and ratings, respectively.[25]

The smaller effects on ELA relative to math in Table 2 match common patterns in the literature (Taylor & Tyler, 2012; Hanushek & Rivkin, 2010; Aucejo et al., 2019; Biasi, 2018; Roth, 2017; Ost, 2014; Nagler et al., 2015; Wiswall, 2013). Previous research has offered multiple explanations for this trend. First, math is primarily learned in the classroom, whereas out of school exposure to ELA is quite common. As a result, teachers have more influence over math performance than ELA performance (Jackson et al., 2014). Second, reading tests are not as sensitive to teacher effort (Kane & Staiger, 2012). Since teachers have more control over math scores than ELA scores, the negative productivity effects of tenure manifest in math value-added declines.

[24] In all tables relying on performance as the dependent variable, the main effects are measured in student test score standard deviations. I also include a standardized estimate of the effect in brackets by dividing the coefficient by the standard deviation of teacher performance in the sample.

[25] These values were calculated by regressing the performance of teachers hired after TEACHNJ on experience and teacher fixed effects. These teachers received tenure in year five, so the returns to experience are not confounded with tenure receipt in the first four years.

I also consider additional threats to identification due to selection into tenure, occupation sorting, and value-added bias. Since lower quality teachers are not offered tenure, the model identifies the effects of tenure by using below average pretenured teachers as a comparison group. I can mitigate this selection problem by estimating equation (2) with a sample that only includes "high-quality" teachers who are likely to receive tenure. Since TEACHNJ relies on ratings to make personnel decisions,[26] I define "high-quality" teachers as those whose third year performance exceeds the twenty-fifth percentile of third year ratings for eventually tenured teachers. Column (1) of Table 3 shows the estimated productivity effects for "high-quality" teachers are quite similar to those found in Table 2. This provides evidence that equation (2) accounts for this selection bias. These results also show that teachers who are likely to receive tenure still experience negative productivity effects at tenure receipt.

Another threat to identification is occupation sorting in response to TEACHNJ. Figure A2 shows similar trends in education bachelor's (Panel A) and master's (Panel B) degrees awarded in New Jersey colleges (dashed red and dashed and dotted blue) relative to other United States colleges (solid black). The supply of new teachers appears to be unaffected by the passage of TEACHNJ. To further reduce the likelihood of sorting by teachers-in-training, I limit the sample to teacher hired before 2013. As college seniors or certified teachers when TEACHNJ passed in 2012, these prospective teachers had little time to change occupations.[27] Column (2) of Table 3 shows the results are robust to samples with limited scope for sorting.[28]

In addition, value-added biases may arise if teachers have systematically different class compositions before and after receiving tenure (Rothstein, 2017). Tenured teachers may receive preferential treatment if districts allow them to select students who are most likely

[26] Figure A1 plots tenure receipt rates by pretenure performance. Ratings are strong predictors of tenure receipt (Panel C), while math (Panel A) and ELA (Panel B) value-added are weak predictors.

[27] Although introduced to the senate on February 6, 2012, the bill had little media coverage prior to June 2012. Since teacher certification programs require several semesters of coursework, it is unlikely that teachers-in-training anticipated the policy and changed majors that late in their college careers.

[28] Sorting into other states and charter schools also threatens identification. However, cross-state sorting is limited by licensure requirements and pension structures (Goldhaber et al., 2015). Also, tenure concerns would not cause public school teachers to switch to charter schools, which have weaker tenure protections.

to improve. To evaluate this hypothesis, I estimate the following regression:

$$C_{jt} = \gamma ten_{jt} + \sum_{\tau=1}^{T} \beta_\tau \mathbf{1}(exp_{jt} = \tau) + \varepsilon_{jt}. \tag{4}$$

In this model, C_{jt} includes class size, class demographics (gender, race, FRPL, ELL, and special education composition), and an indicator for switching grade levels. The coefficient of interest (γ) identifies the effect of receiving tenure on the outcome. If $\gamma \neq 0$, tenured teachers may be manipulating their class rosters.

Using equation (4), Table A4 shows tenured teachers switch grades less frequently, have fewer Black students, and more FRPL eligible students. The value-added models control for student race and FRPL eligibility but do not account for grade switching. Previous research finds that switching grades leads to short-term value-added declines (Ost, 2014). This would bias tenured teachers' value-added and the estimated effect of tenure upward.

To test for bias due to grade switching, I reestimate equation (2) with an indicator for whether a teacher switches grades in a given year. This allows level shifts in performance due to grade switching. Column (3) of Table 3 shows the math value-added estimate declines to -0.040 standard deviations but remains statistically indistinguishable from the main estimates in Table 2. I find no evidence that differences in class composition bias the results.

Value-added also may be biased by the transition from the NJASK and HSPA to the PARCC in 2015. To test for bias, I limit the value-added analysis to only years in which the PARCC test was administered.[29] Column (4) of Table 3 shows the difference-in-differences results are robust to limiting the sample to PARCC years.

The estimated effects are robust to a variety of specifications that account for selection into tenure, occupation sorting, and value-added bias. When considering policy implications, the 0.033 standard deviation decline in math value-added from Table 2 is economically meaningful, though practically small. Using partial equilibrium estimates from Chetty et al.

[29] I cannot estimate the effects of tenure with a sample of only NJASK and HSPA tests because no teacher hired after TEACHNJ had four years of experience prior to 2014. As a result, I do not have any fourth year pretenured teachers to serve as a comparison group.

(2014), this fourth-year decline equates to a present value loss of about $237 per student.[30]

As discussed in Rothstein (2015), there are multiple ways to overcome this productivity decline. First, districts could increase dismissal rates to remove additional lower-performing teachers at tenure receipt. To fill the vacancies, districts would replace these teachers with new hires. However, test scores will only rise in the short-run if districts replace these tenured teachers with novices of higher initial quality. Specifically, math value-added increases by 0.079 standard deviations between years 1 and 4, while tenure induces a productivity decline of 0.033 standard deviations. As a result, districts would have to hire a novice whose initial performance is at least 0.046 standard deviations (0.15 teacher standard deviations) higher than the teacher they replace. This type of reform would be challenging to implement because it is very difficult to identify teacher quality based on observable characteristics at the point of hire (Hanushek, 1997; Buddin & Zamarro, 2009). In the long-run, the change in quality remains ambiguous and depends on the persistence of the effects.

Second, districts could eliminate tenure protections. While this policy would eliminate these negative productivity effects, the efficacy of this reform also depends on the persistence of these declines and the resulting effects on selection into teaching. However, either policy could also improve teacher quality if the productivity effects at tenure receipt demonstrate significant heterogeneity. In Section 4.5, I consider this heterogeneity.

4.5 Productivity Effects at Tenure Receipt Heterogeneity

Though I find weak productivity effects of tenure, these estimates may vary by teacher, school, and student characteristics. If certain teachers are particularly responsive to tenure, directed tenure reforms may remain effective. To conduct this analysis, I reestimate equation (2) by interacting the independent variables with these traits.

First, I investigate which types of teachers are most responsive to tenure. Columns

[30] Chetty et al. (2014) estimates a one standard deviation increase in value-added for one grade generates a present value gain of $7,000 per student. I scale this estimate by the 4% of the teacher workforce that has four years of experience, the 0.106 teacher standard deviation decline in math value-added, and the eight grades for which I can calculate value-added.

(2)–(4) of Table 2 show no differences in value-added by gender. However, the difference in summative ratings across genders is statistically significant at the 5% level, as shown by the p-value from an F-test of equality in Column (4). This finding aligns with previous teacher incentive research, which shows men are more responsive than women to performance pay (Jones, 2013) and increased pressure to remain employed due to weak outside options (Nagler et al., 2015). In addition, Niederle and Vesterlund (2007) find men prefer working in a competitive environment relative to women. These factors would facilitate a performance decline after tenure receipt when the competitive environment to receive tenure vanishes.

Next, I estimate the early career productivity effects by race in Columns (5)–(8) of Table 2. The results are similar, though Hispanic teachers experience no change in math value-added and an increase in ELA value-added. The differences rely on fewer than 300 teachers, so I interpret them cautiously.

The effects may be concentrated among certain subsets of teachers that are particularly sensitive to value-added. Specifically, standardized test scores only affect the ratings of math teachers from grades 4 to 7 and ELA teachers from grades 4 to 8. In Column (5) of Table 3, I reestimate equation (2) using these high stakes subject-grades.[31] Although the future employment of these teachers is directly tied to test scores, their productivity effects remain nearly identical to the average effect of all teachers in Table 2.

I then evaluate which types of schools are most impacted by the productivity effects of tenure in Table 4. I consider indicators for having at least 20% Black (Columns (1)–(3)), Hispanic (Columns (4)–(6)), and FRPL eligible (Columns (7)–(9)) students. In Table 5, I conduct a similar analysis using 50% math (Columns (1)–(3)) and ELA (Columns (4)–(6)) proficiency rates as thresholds. Although I find no statistically significant differences in summative ratings, the negative value-added effects are concentrated among wealthier schools with few minority students and high proficiency rates. These differences are statistically significant for the percentage of Hispanic, FRPL eligible, and math proficient students.

[31] Students in other grades take these exams but their scores do not impact their teachers' ratings.

To test the robustness of these results, I estimate which types of students are most impacted by the productivity effects of tenure. I only consider value-added because I cannot separate ratings by student characteristics. Table 6 shows the results by student gender (Columns (1)–(3)), race (Columns (4)–(7)), and FRPL eligibility (Column (8)–(10)). Male students suffer more than female students when paired with a tenured teacher. This finding reflects the results by teacher gender in Table 2. Similarly, Hispanic and FRPL eligible students experience less of a decline in their teachers' math and ELA value-added relative to their counterparts. These results align with school characteristics in Table 4.

From a policy perspective, tenure is often criticized for protecting low-performing teachers in impoverished districts. However, these results show the negative effects are actually concentrated in wealthier, higher-achieving schools with few minority students. This suggests tenure reforms should target these schools. However, almost all of the impacts remain smaller than the returns to experience in the first four years.[32] As discussed in Section 4.4, the efficacy of alternative policies depend on the persistence of the performance declines. In Section 5, I estimate later career productivity effects to evaluate whether these shocks linger.

5 Productivity Effects Later in Career

In this section, I use an RDD to estimate the productivity effects of tenure on experienced teachers. In combination with the results from Section 4, I estimate variation in these effects throughout a teacher's career. This heterogeneity may arise if tenure reduces the incentive to invest in human capital. As a result, negative productivity effects may worsen when educators reduce on-the-job training as they near retirement and the costs of training exceed the stream of benefits. Alternatively, negative productivity effects may diminish if effort becomes relatively less important as innate ability improves due to experience. For example, experienced teachers may only need to refine lesson plans rather than construct

[32] The only point estimate that is larger than early career returns to experience (0.043 ELA standard deviations) is the ELA value-added productivity effects for schools with few FRPL eligible students (-0.050 standard deviations). However, the two values are statistically indistinguishable from each other.

new ones, so marginal changes in incentives may have little impact on their productivity.

The RDD relies on the TEACHNJ provision that allows districts to dismiss tenured teachers who received consecutive low summative ratings. After the first low rating, teachers must earn a high rating in the subsequent year to keep their jobs.[33] Thus, I estimate the effects of dismissal threats by comparing the performance of teachers in year t whose year $t-1$ summative rating fell just above or below the effective rating threshold. In addition to the dismissal threat, teachers receiving a low rating must develop a Corrective Action Plan (CAP) with their supervisors. Little research has investigated improvement plans, such as the CAP, though evaluations can be effective tools for professional development among low-performing teachers (Taylor & Tyler, 2012). I assume the CAP weakly improves performance, so the estimated effects of dismissal threats serve as upper bounds.

First, I evaluate turnover rates conditional on previous low ratings to illustrate the strength of the dismissal threats. Ex-ante, I do not expect dismissal rates to increase by 100% at the discontinuity. As discussed in Section 2, some teachers earning consecutive low ratings are retained because TEACHNJ allows districts to offer a third opportunity to teachers earning partially effective ratings in the second year. This flexibility is particularly important for schools that struggle to fill vacancies. Other teachers are retained when arbitrators rule in favor of them after receiving charges of inefficiency. In addition, some teachers earning effective ratings leave the district voluntarily.

Panel A of Figure 5 plots the fraction of tenured teachers returning by previous year rating conditional on a low rating two years prior. Since teachers receiving consecutive low annual ratings may be dismissed, there should be a discontinuity at the effective threshold (2.65) among teachers receiving dismissal threats. For these teachers, anyone scoring lower than 2.65 may be dismissed, while those scoring at or above 2.65 should be retained. However, the aforementioned factors result in a discontinuity of 10.1 percentage points with a p-value

[33] To simplify notation, I define partially effective and ineffective ratings as low ratings, while effective and highly effective ratings are defined as high ratings. This classification is consistent with the TEACHNJ Act, as two effective or highly effective (high) ratings are required to receive tenure, while tenured teachers can be dismissed for consecutive partially effective or ineffective (low) ratings.

of 0.051.[34] In comparison, Panel B of Figure 5 shows the fraction of teachers returning by previous summative rating given a high rating two years prior. These teachers had not yet received a dismissal threat, so they all should be retained. The point estimate for the discontinuity is 0.8 percentage points and statistically indistinguishable from 0. Threats of dismissal increase turnover rates at the effective threshold by 9.3 percentage points (from 0.8 to 10.1 percentage points) among tenured teachers.

In Panel A of Figure 5, the dismissal threats appear weak. However, the TEACHNJ Act requires teacher summative ratings to remain confidential, while performance-related dismissals are public information. Teachers know the unconditional number of performance-related dismissals but they do not know the dismissal rates of teachers earning low ratings. As a result, teachers would recognize the dismissal threats as credible when observing any performance-related dismissal in conjunction with the language of the law.[35]

In fact, the percentage of New Jersey teachers who believed they would be dismissed for sustained poor performance increased from 31% to 43% after TEACHNJ passed (Callahan & Sadeghi, 2015). This 12 percentage point increase represents a large shift in teachers' beliefs. Historically, tenured teachers were almost never dismissed for poor performance. For example, over a 7 year period before TEACHNJ, only 3 performance-related tenure hearings appeared in courts with 2 upheld. In comparison, New Jersey had 74 performance-related teacher tenure arbitration hearings with 40 dismissals upheld from 2012 to 2019.[36] This trend is common throughout the United States where only 0.1% of tenured teachers are dismissed for poor performance (National Center for Education Statistics, 2012). Since so few tenured teachers are dismissed, tenure reforms are unlikely to universally change teachers' beliefs about threats of dismissal. However, the threat of dismissal in New Jersey

[34] I calculate the discontinuity by using a linear spline above and below the threshold with the optimal bandwidth developed by Calonico et al. (2014).

[35] Teachers receiving low ratings may voluntarily share these ratings with colleagues and the union. However, they have little incentive to disclose their poor performance, especially if they are trying to find work in another school district.

[36] I manually recorded these values by reviewing every performance-related arbitration decision available on the New Jersey government website between 2001 and 2008, as well as 2012 and 2019 (Department of Education, 2020). The website did not have court data from 2009 to 2012.

remains relatively strong.

Given the credibility of the dismissal threats, I formally introduce the RDD. In the RDD, the running variable is the summative rating ($S_{j(t-1)} \in [1, 4]$) of teacher j in year $t-1$. A tenured teacher faces a dismissal threat in year t after receiving a low rating in year $t-1$. Thus, treatment in year t (D_{jt}) is defined as follows:

$$D_{jt} = \mathbf{1}(S_{j(t-1)} < 2.65).$$

Using equation (1) to calculate each teacher's annual value-added, I estimate the following model for tenured teachers:

$$y_{jt} = \gamma D_{jt} + m(S_{j(t-1)}) + \sum_{\tau=4}^{T} \delta_\tau \mathbf{1}(exp_{jt} = \tau) + \varepsilon_{jt}. \tag{5}$$

The dependent variable (y_{jt}) is the annual performance measure (value-added or summative rating) of teacher j in year t. I include experience fixed effects ($\mathbf{1}(exp_{jt} = \tau)$) to account for non-linear returns to experience and use a linear spline above and below the threshold to estimate $m(S_{j(t-1)})$. The coefficient of interest (γ) estimates the effect of dismissal threats on performance. Based on the bandwidth selection method developed by Calonico et al. (2014), I use bandwidths of 0.193, 0.411, and 0.185 for math value-added, ELA value-added, and summative ratings, respectively.

To estimate causal effects, I assume teachers above and below the threshold perform similarly in the absence of dismissal threats (Imbens & Lemieux, 2008). In Appendix Section A.5, I show the results of several balance tests to provide evidence that supports this assumption.

5.1 Productivity Effects Later in Career Results

Using equation (5), Table 7 presents the productivity effects of tenure later in the career. In Column (1), I find no effect of dismissal threats on math value-added (Panel A), ELA

value-added (Panel B), or summative ratings (Panel C). The 95% confidence intervals rule out increases larger than 0.073 math standard deviations, 0.030 ELA standard deviations, and 0.050 rating points. Figure 6 corroborates these findings as the binned scatterplots for math value-added (Panel A), ELA value-added (Panel B), and summative ratings (Panel C) reveal no discontinuity at the threshold. As discussed in Section 5, the estimated null effects represent upper bounds because the CAP should weakly improve performance. Columns (2)–(8) of Table 7 show no evidence of heterogeneity by teacher gender and race.[37] Table A5 shows the results are robust to various bandwidths, except when they exceed 0.5 points. These differences are likely due to bias from extremely large bandwidths.[38]

These results show experienced teachers do not respond to dismissal threats. Surprisingly, these dismissal threats do not generate rating improvements even though the teacher's continued employment depends on these ratings.[39] Thus, removing tenure protections from these experienced teachers generates no gains.

These results suggest that the negative productivity effects from Section 4.4 do not persist throughout teachers' careers. Thus, the $237 present value loss per student based on a one year productivity shock is a reasonable estimate for the costs of tenure.

Referring back to alternative policies, a reform that replaces lower-performing tenured teachers with new hires becomes even less attractive. In Section 4.4, I showed that districts must hire novices whose initial performance is at least 0.046 standard deviations higher than the teachers they replace to generate short-run test score improvements. Long-run changes were ambiguous if the negative productivity effects persisted. However, the productivity effects are transitory, while the accrued experience effects are permanent. Once productivity recovers, the comparison between new hires and tenured teachers relies solely on innate

[37] I also find consistent null effects when testing for heterogeneity by high stakes subjects, school composition, and student characteristics (not shown).

[38] The results also are robust to including district fixed effects (not shown).

[39] These estimates differ from Dee and Wyckoff (2015). In Appendix Section A.6, I explain how the summative rating effects may vary by the credibility of the dismissal threats and teacher experience. I find that New Jersey teachers facing dismissal threats are much more experienced than those in Washington D.C. The more experienced teachers appear to be less responsive to dismissal threats.

ability and the returns to experience without differencing out any productivity effects. This comparison will often favor more experienced teachers, even when considering the largest effects from Section 4.5 in wealthy, higher-achieving schools with few minority students. Alternative policies, such as the elimination of tenure, will also generate few gains since productivity recovers later in the career. However, tenure reforms also impact selection into teaching and retention. In Section 6, I estimate these effects on the teacher labor market.

6 Effects on Teacher Retention and Average Quality

While the previous sections found a temporary decline in math value-added at tenure receipt that dissipated over time, this section evaluates the effects of the tenure reform on the teacher labor market. TEACHNJ may create vacancies, as more stringent requirements increase teacher turnover and alter sorting patterns. These sorting patterns may contribute to inequalities as the teacher workforce becomes less diverse. In addition, weaker tenure protections may remove low-quality teachers, although reduced compensating differentials also may induce high-quality teachers to leave the profession for more lucrative opportunities. As a result, TEACHNJ has a theoretically ambiguous effect on average teacher quality.

When evaluating the labor market impacts of TEACHNJ, I focus on the increased pretenure time and standardized tenure requirements because the teacher-student linked data start in 2012. Every observation occurs after TEACHNJ passed, when the standardized evaluation criteria and streamlined tenure removal process had already been universally implemented. In this analysis, I compare teachers hired before and after TEACHNJ.[40]

6.1 Retention Effects of TEACHNJ

TEACHNJ made tenure receipt more difficult. This would increase both the number of pretenure dismissals and the voluntary attrition rate for those hired after TEACHNJ. Fewer

[40] The strategy also captures effects from the mentor program, though it likely had far less impact on teacher sorting than tenure reforms that directly impacted job security.

teachers will meet the tenure receipt standards, while more teachers will voluntarily quit due to reduced compensating differentials associated with the more arduous process. To evaluate this hypothesis, I estimate the following regression:

$$y_{jtd} = \gamma post_{jt} + \sum_{\tau=1}^{T} \beta_\tau \mathbf{1}(exp_{jt} = \tau) + \nu_d + \varepsilon_{jtd}. \tag{6}$$

In this model, y_{jtd} is an indicator for teacher j leaving district d after year t. I estimate the effects of being hired after TEACHNJ using $post_{jt}$. I include experience ($\mathbf{1}(exp_{jt} = \tau)$) and district ($\nu_d$) fixed effects to account for differences in turnover rates by experience level and district. I use district fixed effects because I define turnover as leaving the district rather than moving between schools within a district.[41] As a result, I cluster standard errors at the district level.

In this regression, I cannot simultaneously control for experience, year, and cohort because they are nearly collinear. I must include controls for cohort because the coefficient of interest compares teachers hired before and after TEACHNJ. Thus, I can control for either experience or year fixed effects. If I control for experience and exclude year fixed effects, I must assume that turnover rates are invariant across calendar years in my sample. If I control for calendar years and exclude experience fixed effects, I must assume that turnover rates are invariant across years of experience. To test which assumption is more reasonable, I plot turnover rates by experience (Panel A) and calendar year (Panel B) in Figure A3. I find large variation in turnover rates across experience years and little variation across calendar years, so I use experience fixed effects.

Table 8 reports estimates of equation (6). In Column (1), the reforms increased overall teacher turnover by 6.3 percentage points. As seen in the row labeled "Mean", 12.6% of pretenured teachers leave their district each year. Thus, the reforms dramatically increased overall turnover rates among these inexperienced teachers.

[41] In fact, the district, rather than the school, employs the teacher. Teachers retain tenure protections when moving between schools within a district but lose these protections when switching districts.

While TEACHNJ increased overall teacher turnover, it is important to identify which teachers are leaving. Teachers may voluntarily leave the district following TEACHNJ due to reduced compensating differentials or involuntarily leave the district as they fail to meet the more arduous standards. Although I cannot precisely measure the source of turnover, Table 9 provides a proxy for it by dividing the sample of teachers into those receiving high ratings (Panel A) and low ratings (Panel B). Since teachers do not face performance-related dismissals if they receive high summative ratings, Panel A provides a proxy for voluntary attrition among effective teachers.[42] In comparison, Panel B measures both involuntary and voluntary turnover among teachers receiving low ratings. Column (1) shows TEACHNJ increased the proxy for voluntary attrition by 3.8 percentage points and turnover among low-performing teachers by 21.4 percentage points.[43] Therefore, TEACHNJ primarily increased turnover among low-performing teachers, though there also was a meaningful rise in turnover among effective teachers.

Changes in turnover rates also may vary by teacher characteristics. Some teachers may be disproportionately impacted by TEACHNJ leading to increased turnover. In Columns (2)–(8) of Table 8, I evaluate differential turnover rates by interacting the independent variables in equation (6) with teacher gender and race. Columns (2)–(4) show the effects are larger for men (7.6 percentage points) than women (6 percentage points), while Columns (5)–(8) show higher increases in turnover among Black teachers (8.6 percentage points) than other teachers (about 6 percentage points). Columns (4) and (8) show these differences are statistically significant.[44] The increased turnover among male and Black teachers could be a result of voluntary attrition or involuntary dismissals. In Panel A of Table 9, the proxy for voluntary

[42] In fact, it is very difficult to identify voluntary and involuntary turnover in any dataset. For example, some low-performing teachers may appear to voluntarily leave the district if they knew that they would be dismissed shortly afterwards. As a result, comparing turnover among low- and high-performing teachers provides a reasonable proxy for these measures.

[43] During the sample period, pretenure turnover rates among effective and ineffective teachers hovered near 10% and 33%–42%, respectively.

[44] Table 8 differs from Figure 3 because Table 8 does not control for summative ratings, while Figure 3 is limited to teachers in their first districts. However, the overall pattern of results in Table 8 is robust to this sample variation (not shown).

attrition for male and Black teachers increased by 4.5 and 4.3 percentage points, respectively. However, female, White, and Hispanic teachers only experienced 3.5–3.7 percentage point increases in the proxy for voluntary attrition. Similarly, Panel B shows turnover among low-performing male and Black teachers increased by 22.7 and 26.6 percentage points, respectively, while other teachers only experienced 19.9–20.5 percentage point increases. The differences among low-performing teachers are not statistically significant; however, the point estimates suggest male and Black teachers encountered disproportionately higher turnover rates due to poor performance.

Male and Black teachers also were disproportionately affected by TEACHNJ because they consistently received lower summative ratings. Figure A4 shows the cumulative distribution functions of the ratings by teacher gender (Panel E) and race (Panel F). The distribution of ratings for female (dashed red) and White (dashed red) teachers first order stochastically dominates the distribution for male (dashed and dotted blue) and Black (solid black) teachers, respectively. This finding matches previous research that has documented similar racial and gender gaps in teacher evaluation scores (Bailey et al., 2016; Drake et al., 2019; Sartain & Steinberg, 2020; Chi, 2021). Since TEACHNJ includes provisions to dismiss teachers earning low ratings, the lower scoring male and Black teachers faced greater turnover rates. These teachers' lower ratings may be due to a variety of factors including sorting patterns, differences in performance, and evaluation bias.

Male and Black teachers may sort into different schools than other teachers. To evaluate whether this mechanism could generate the differential turnover rates by teacher characteristics, Table 10 presents estimates of equation (6) by school size, poverty level (Panel A), racial composition (Panel B), and proficiency rates (Panel C). TEACHNJ increased turnover similarly across school types with differences in point estimates that are no larger than 0.6 percentage points. None of the differences are statistically significant at even the 10% level.[45] Given similar changes in turnover rates across school characteristics, the in-

[45] In addition, the mean turnover rates are similar across school types.

crease in turnover among male and Black teachers is unlikely to be attributable to differences in school attributes.

Next, I test for differences in performance along other dimensions of teacher quality. I plot the cumulative distribution functions of value-added by gender and race in Figure A4 (Panels A–D). While male and Black teachers consistently receive lower summative ratings, I find similar distributions of value-added by gender and race. In fact, relative to White teachers, Black teachers have slightly higher average value-added. Although value-added and ratings capture distinct components of teacher performance, it is surprising that male and Black teachers perform so much worse along the only potentially subjective dimension.

Given similar value-added by gender and race, summative rating biases may contribute to the rating discrepancies. For example, supervisors may offer more lenient ratings to teachers of the same gender or race. As a result, teachers from other groups may receive lower ratings and encounter increased risk of dismissals. To evaluate this hypothesis, I estimate the following model:

$$rate_{jtp} = \gamma group_j + \delta rate_{j(t-1)} + \sum_{\tau=1}^{T} \beta_\tau \mathbf{1}(exp_{jt} = \tau) + \xi_p + \varepsilon_{jtp}. \tag{7}$$

In equation (7), $rate_{jtp}$ is the summative rating of teacher j in year t who is assigned to principal p, while $group_j$ is an indicator for teacher gender or race. I include experience ($\mathbf{1}(exp_{jt} = \tau)$) and principal ($\xi_p$) fixed effects to account for returns to experience and differential evaluation standards by principal. I control for prior year summative rating ($rate_{j(t-1)}$) and interact each variable with the principal's gender or race to test for differences in rating standards by group membership. Table 11 presents estimates of γ where column headers define principal characteristics. This table shows that male (Panel A) and Black (Panel B) teachers consistently receive lower ratings but the differences are even larger when paired with an out-of-group principal.[46] For example, male teachers earn ratings that are 0.028 points lower than female teachers when paired with a male principal. However, their

[46] Hispanic teachers also receive lower ratings than White teachers but the differences are smaller.

ratings drop by an additional 0.007 points to −0.035 when paired with a female principal. Similarly, Black teachers earn ratings that are 0.024 points lower when paired with Black principals. This disparity increases to 0.033 points when paired with White principals. These differences across principal characteristics are statistically significant at the 5% level for both groups. The increased out-of-group rating disparities suggest evaluation biases contribute to the lower ratings for male and Black teachers.

The increased turnover for male and Black teachers is particularly problematic for male and Black students. Gershenson et al. (2018) find Black students' graduation and college enrollment rates increased when paired with Black teachers. Other papers show test score improvements when male and Black students were assigned to teachers of their own gender (Dee, 2007) and race (Dee, 2004; Egalite et al., 2015). Similarly, Dee (2005, 2007), Ehrenberg et al. (1995), and Gershenson et al. (2016) find teachers had worse perceptions of out-of-group students. Many of these papers also suggest that simply increasing the number of male and Black teachers could have symmetric negative impacts on female and White students. However, Table A1 shows that male and Black teachers are underrepresented relative to the size of the corresponding student populations. About 52% of students are male, while only 20% of teachers are male. Similarly, about 20% of students are Black, whereas only 8% of teachers are Black.[47] With already limited access to in-group teachers, male and Black students may be disproportionately harmed by increased in-group turnover when implementing tenure reforms relying on subjective evaluations.

6.2 Labor Market Effects: Teacher Quality

TEACHNJ also may impact average teacher performance. Tenure reforms increase dismissals among low-performing teachers; however, reduced compensating differentials from weakened job security and longer pretenure periods may increase attrition among high-performing teachers. In addition, TEACHNJ standardized tenure receipt and removal, such that they

[47] These gender and racial disparities are prevalent throughout the United States (National Center for Education Statistics, 2020a, 2020b).

only rely on summative ratings rather than value-added. As a result, schools may alter hiring practices to focus on the rating dimension, while the law incentivizes potentially highly rated teachers to sort into the profession.

To estimate these labor market effects, I reestimate equation (6) and replace y_{jtd} with value-added and summative ratings. As discussed earlier, I cannot estimate the effects of being hired after TEACHNJ while simultaneously controlling for both experience and year fixed effects. To prevent this collinearity, I omit year fixed effects because average value-added does not drift over time[48] and the summative rating returns to the first five years of experience are over twice as large as their drift across calendar years.

Table 12 presents the results. There is no change in value-added and a 0.021 point increase in summative ratings following TEACHNJ's passage. The 95% confidence intervals rule out declines larger than 0.009 math and 0.011 ELA standard deviations.

These results show that teaching candidates who perform well on ratings disproportionately filled vacancies following the TEACHNJ Act. However, these candidates who received high ratings were not necessarily more effective as measured by value-added. As a result, teacher labor market quality improved along the dimension that dictated personnel decisions (ratings), while quality remained unchanged along other dimensions of performance (value-added). This result aligns with a principal-agent model where only one of several measures of performance is used to evaluate the employee (Holmstrom & Milgrom, 1991; Baker, 2002).

7 Conclusion

At tenure receipt, I estimate negative productivity effects of tenure on math value-added with no impact on ELA value-added or summative ratings. Since the productivity effects are smaller than early career returns to experience, replacing teachers due to these productivity concerns will likely cause short-term declines in student performance. The long-term gains from increasing dismissal rates or eliminating tenure are limited by the transitory nature of

[48] This is a mechanical effect of the annual test score standardization.

the negative productivity effects.

Focusing on retention effects, male and Black teachers are disproportionately affected by the tenure reforms of TEACHNJ. These teachers tend to earn lower summative ratings leading to greater turnover when switching to more stringent accountability standards. Despite lower ratings, male and Black teachers have similar value-added as their female and White counterparts. While these reforms remove the lowest rated teachers, they may unintentionally reduce the diversity of the teacher workforce. Worse yet, summative ratings with subjective components may vary by gender and race leading to disproportionate effects. Policymakers need to consider these unintended consequences when enacting future tenure reforms tied to subjective evaluation criteria.

TEACHNJ created vacancies that were filled by teachers who received higher summative ratings. However, these new hires produced similar value-added to previous cohorts. The discrepancies between the value-added and summative rating results demonstrate the importance of carefully analyzing these performance metrics. If research can link summative ratings to future student outcomes, similar to Chetty et al. (2014) for value-added, these rating gains may prove to be useful. Without this clear relationship, schools may be manipulating their hiring practices to improve along a less meaningful metric.

Given these effects, the efficacy of tenure reforms to improve teacher performance is limited. Other policies may be more effective instruments to increase teacher productivity. For example, the 0.033 standard deviation decline in math value-added is less than one-third of the magnitude of the 0.11 standard deviation improvement associated with increased teaching feedback (Taylor & Tyler, 2012). In addition, increased feedback generated improvements that persisted over several years without any evidence of negative labor market effects.

Overall, both sides of the tenure debate make valid points. Tenure induces a temporary math value-added productivity decline; however, tenure reforms also impact selection into teaching and retention. These reforms have unintended consequences on diversity, which may harm male and Black students.

References

Adnot, M., Dee, T., Katz, V., & Wyckoff, J. (2017). Teacher turnover, teacher quality, and student achievement in dcps. *Educational Evaluation and Policy Analysis*, *39*(1), 54–76.

Anderson, K. P., Cowen, J., & Strunk, K. (2019). The impact of teacher labor market reforms on student achievement: Evidence from Michigan. *Education Policy Innovation Collaborative (EPIC) Working Paper*, *1*.

Aucejo, E., Romano, T., & Taylor, E. (2019). Does evaluation distort teacher effort and decisions? Quasi-experimental evidence from a policy of retesting students.

Bailey, J., Bocala, C., Shakman, K., & Zweig, J. (2016). Teacher demographics and evaluation: A descriptive study in a large urban district. *Institute of Education Sciences*.

Baker, G. (2002). Distortion and risk in optimal incentive contracts. *Journal of Human Resources*, 728–751.

Betebenner, D. (2011). A technical overview of the student growth percentile methodology: Student growth percentiles and percentile growth projections/trajectories. *National Center for the Improvement of Educational Assessment*.

Biasi, B. (2018). The labor market for teachers under different pay schemes. *National Bureau of Economic Research*.

Buddin, R., & Zamarro, G. (2009). Teacher qualifications and student achievement in urban elementary schools. *Journal of Urban Economics*, *66*(2), 103–115.

Callahan, K., & Sadeghi, L. (2015). Teacher perceptions of the value of teacher evaluations: New Jersey's AchieveNJ. *International Journal of Educational Leadership Preparation*, *10*(1), 46–59.

Calonico, S., Cattaneo, M., & Titiunik, R. (2014). Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica*, *82*(6), 2295–2326.

Carruthers, C., Figlio, D., & Sass, T. (2018). Did tenure reform in Florida affect student test scores? *Evidence Speaks Reports*, *2*(52), 1–14.

Chetty, R., Friedman, J., & Rockoff, J. (2014). Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. *American Economic Review*, *104*(9), 2633–79.

Chi, O. (2021). A classroom observer like me: The effects of race-congruence and gender-congruence between teachers and raters on observation scores. *Brown University Ed-WorkingPaper*.

Cowen, J., & Strunk, K. (2015). The impact of teachers' unions on educational outcomes: What we know and what we need to learn. *Economics of Education Review*, *48*, 208–223.

Dee, T. (2004). Teachers, race, and student achievement in a randomized experiment. *Review of Economics and Statistics*, *86*(1), 195–210.

Dee, T. (2005). A teacher like me: Does race, ethnicity, or gender matter? *American Economic Review*, *95*(2), 158–165.

Dee, T. (2007). Teachers and the gender gaps in student achievement. *Journal of Human Resources*, *42*(3), 528–554.

Dee, T., & Wyckoff, J. (2015). Incentives, selection, and teacher performance: Evidence from IMPACT. *Journal of Policy Analysis and Management*, *34*(2), 267–297.

Department of Education. (2020). New Jersey School Law. *State of New Jersey*. Retrieved from `https://www.state.nj.us/education/legal/`.

Drake, S., Auletto, A., & Cowen, J. (2019). Grading teachers: Race and gender differences in low evaluation ratings and teacher employment outcomes. *American Educational Research Journal*, *56*(5), 1800–1833.

Egalite, A., Kisida, B., & Winters, M. (2015). Representation in the classroom: The effect of own-race teachers on student achievement. *Economics of Education Review*, *45*, 44–52.

Ehrenberg, R., Goldhaber, D., & Brewer, D. (1995). Do teachers' race, gender, and ethnicity matter? Evidence from NELS88 (No. w4669). *National Bureau of Economic Research*.

Gershenson, S., Hart, C., Hyman, J., Lindsay, C., & Papageorge, N. (2018). The long-run impacts of same-race teachers.

Gershenson, S., Holt, S., & Papageorge, N. (2016). Who believes in me? The effect of student–teacher demographic match on teacher expectations. *Economics of Education Review*, *52*, 209–224.

Goldhaber, D., Grout, C., Holden, K., & Brown, N. (2015). Crossing the border? Exploring the cross-state mobility of the teacher workforce. *Educational Researcher*, *44*(8), 421–431.

Goldhaber, D., Hansen, M., & Walch, J. (2016). Time to tenure: Does tenure reform affect teacher absence behavior and mobility? *National Center for Analysis of Longitudinal Data in Education Research*.

Hanushek, E. (1997). Assessing the effects of school resources on student performance: An update. *Educational Evaluation and Policy Analysis*, *19*(2), 141–164.

Hanushek, E., & Rivkin, S. (2010). Generalizations about using value-added measures of teacher quality. *American Economic Review*, *100*(2), 267–71.

Hanushek, E., Rivkin, S., & Schiman, J. (2016). Dynamic effects of teacher turnover on the quality of instruction. *Economics of Education Review*, *55*, 132–148.

Holmstrom, B., & Milgrom, P. (1991). Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. *Journal of Law, Economics, and Organization*, *7*, 24.

Hoxby, C. (1996). How teachers' unions affect education production. *The Quarterly Journal of Economics*, *111*(3), 671–718.

Imbens, G., & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, *142*(2), 615–635.

Jackson, C., Rockoff, J., & Staiger, D. (2014). Teacher effects and teacher-related policies. *Annual Review of Economics*, *6*(1), 801–825.

Jacob, B. (2013). The effect of employment protection on teacher effort. *Journal of Labor*

Economics, *31*(4), 727–761.

Johnston, A. (2018). Teacher utility, separating equilibria, and optimal compensation: Evidence from a discrete-choice experiment. *NBER Economics of Education Coference*.

Jones, M. (2013). Teacher behavior under performance pay incentives. *Economics of Education Review*, *37*, 148–164.

Jones, M. (2015). How do teachers respond to tenure? *IZA Journal of Labor Economics*, *4*(1), 8.

Kane, T., & Staiger, D. (2012). Gathering feedback for teachers: Combining high-quality observations with student surveys and achievement gains. *Bill and Melinda Gates Foundation*.

Kraft, M. (2015). Teacher layoffs, teacher quality, and student achievement: Evidence from a discretionary layoff policy. *Education Finance and Policy*, *10*(4), 467–507.

Kraft, M., & Papay, J. (2015). Productivity returns to experience in the teacher labor market: Methodological challenges and new evidence on long-term career improvement. *Journal of Public Economics*, *130*, 105–119.

Lee, D., & Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of Economic Literature*, *48*(2), 281–355.

Loeb, S., Miller, L., & Wyckoff, J. (2015). Performance screens for school improvement: The case of teacher tenure reform in New York City. *Educational Researcher*, *44*(4), 199–212.

Lovenheim, M., & Willén, A. (2019). The long-run effects of teacher collective bargaining. *American Economic Journal: Economic Policy*, *11*(3), 292–324.

Marianno, B., & Strunk, K. (2018). The bad end of the bargain? Revisiting the relationship between collective bargaining agreements and student achievement. *Economics of Education Review*, *65*, 93–106.

McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics*, *142*(2), 698–714.

Nagler, M., Piopiunik, M., & West, M. (2015). Weak markets, strong teachers: Recession at career start and teacher effectiveness. *National Bureau of Economic Research*.

National Center for Education Statistics. (2012). School and Staffing Survey.

National Center for Education Statistics. (2018). School Locations and Geoassignments. Retrieved from `https://nces.ed.gov/programs/edge/Geographic/SchoolLocations`.

National Center for Education Statistics. (2020a). Characteristics of Public School Teachers. Retrieved from `https://nces.ed.gov/programs/coe/indicator_clr.asp`.

National Center for Education Statistics. (2020b). Racial/Ethnic Enrollment in Public Schools. Retrieved from `https://nces.ed.gov/programs/coe/indicator_cge.asp`.

New Jersey Department of Education. (2017). New Jersey School Directory. Retrieved from `https://homeroom5.doe.state.nj.us/directory/`.

Niederle, M., & Vesterlund, L. (2007). Do women shy away from competition? Do men compete too much? *The Quarterly Journal of Economics*, *122*(3), 1067–1101.

Ost, B. (2014). How do teachers improve? the relative importance of specific and general human capital. *American Economic Journal: Applied Economics*, *6*(2), 127–51.

Roberts, M. (2018). *Essays on regime change and education policy reform* (Unpublished doctoral dissertation). University of Kansas.

Rodriguez, L. (2018). *An examination of teacher tenure reform in Tennessee: Turnover, Performance, and Sense-Making* (Unpublished doctoral dissertation).

Ronfeldt, M., Loeb, S., & Wyckoff, J. (2013). How teacher turnover harms student achievement. *American Educational Research Journal*, *50*(1), 4–36.

Roth, J. (2017). Union reform and teacher turnover: Evidence from Wisconsin's Act 10. *Harvard Kennedy School*.

Rothstein, J. (2015). Teacher quality policy when supply matters. *American Economic Review*, *105*(1), 100–130.

Rothstein, J. (2017). Measuring the impacts of teachers: Comment. *American Economic*

Review, *107*(6), 1656–84.

Sartain, L., & Steinberg, M. (2020). What explains the race gap in teacher performance ratings? Evidence from Chicago Public Schools. *Educational Evaluation and Policy Analysis*.

Shulman, P. (2016). Announcement of evaluation weights for 2016-17. *New Jersey Department of Education*. Retreived from: `https://www.nj.gov/education/broadcasts/2016/AUG/31/15215/AchieveNJ%20Weight%20Memo.pdf`.

Sorensen, L., & Ladd, H. (2020). The hidden costs of teacher turnover. *AERA Open*, *6*(1).

State of New Jersey Department of Education. (2017). 2015-16 Educator Evaluation Implementation Report.

Strunk, K., Barrett, N., & Lincove, J. (2017). When tenure ends: The short-run effects of the elimination of Louisiana's teacher employment protections on teacher exit and retirement. *Education Research Alliance Technical Report*.

Taylor, E., & Tyler, J. (2012). The effect of evaluation on teacher performance. *American Economic Review*, *102*(7), 3628–51.

Thomsen, J. (2020). State legislation: Teaching quality - tenure or continuing contract. *Education Commission of the States*. Retrieved from `https://www.ecs.org/state-education-policy-tracking/`.

Wiswall, M. (2013). The dynamics of teacher quality. *Journal of Public Economics*, *100*, 61–78.

Tables

Table 1: Tenure Receipt and Length Reform Estimates Worst Case Scenario

	(1)	(2)
	Tenure Length	Tenure Receipt and Length
Tenure Receipt	-3.085***	-3.073***
	(0.008)	(0.010)
Obs	257,410	257,410

Notes: This table shows γ coefficients from equation (2) when using the model defined in Section 4.2.1. The estimates are parameterized to depict the worst case scenario 6% annual decrease in retention attributable to TEACHNJ (see Figure 3). The parameterization is described in Appendix Section A.4.1. Column (1) shows the effect when only extending the time to tenure from three to four years. Column (2) shows the effect when extending the time to tenure and increasing the tenure receipt requirements.

Standard errors in parentheses.

* p<0.10, ** p<0.05, *** p<0.01

Table 2: Productivity Effects at Tenure Receipt

Panel A: Math Value-Added

	(1) All	(2) Male	(3) Female	(4) P-Value	(5) White	(6) Black	(7) Hispanic	(8) P-Value
Tenure	-0.033***	-0.034	-0.032***	0.935	-0.033***	-0.047*	0.000	0.493
	(0.011)	(0.022)	(0.011)		(0.012)	(0.028)	(0.030)	
	[-0.106]	[-0.110]	[-0.104]		[-0.108]	[-0.151]	[0.001]	
Num Schools	1,867	1,023	1,765		1,822	425	381	
Num Teachers	3,199	651	2,548		2,796	196	215	
Num Students	514,487	122,780	391,707		448,306	31,860	34,930	

Panel B: ELA Value-Added

	(1) All	(2) Male	(3) Female	(4) P-Value	(5) White	(6) Black	(7) Hispanic	(8) P-Value
Tenure	0.007	0.010	0.008	0.945	-0.001	0.041	0.080**	0.095
	(0.011)	(0.037)	(0.011)		(0.012)	(0.035)	(0.040)	
	[0.025]	[0.036]	[0.027]		[-0.003]	[0.143]	[0.278]	
Num Schools	1,895	854	1,832		1,830	494	480	
Num Teachers	3,562	505	3,057		3,069	245	261	
Num Students	645,807	90,858	554,949		560,965	42,216	44,549	

Panel C: Summative Ratings

	(1) All	(2) Male	(3) Female	(4) P-Value	(5) White	(6) Black	(7) Hispanic	(8) P-Value
Tenure	-0.018	-0.118***	-0.002	0.015	-0.008	-0.137*	-0.066	0.187
	(0.014)	(0.044)	(0.016)		(0.016)	(0.082)	(0.047)	
	[-0.057]	[-0.367]	[-0.006]		[-0.024]	[-0.426]	[-0.204]	
Num Schools	1,568	1,568	1,568		1,484	225	225	
Num Teachers	3,871	671	3,200		3,371	229	282	
Obs	13,960	2,420	11,540		12,241	775	976	

Notes: This table shows γ from equation (2) for the performance measure listed in the panel title. Only Panel C includes school and year fixed effects. Column (1) shows the effect on all teachers. Columns (2) and (3) show the effect by interacting each independent variable with teacher gender. Column (4) provides the p-value from an F-test of equality for the coefficients. Columns (5)–(8) are defined similarly for teacher race. For race, the F-test evaluates whether the coefficients for Black and Hispanic teachers are jointly different from the coefficient for White teachers.

Standard errors in parentheses and clustered at the school level. Performance units rescaled to standard deviation 1 in the dataset are included in brackets.

* p<0.10, ** p<0.05, *** p<0.01

Table 3: Productivity Effects at Tenure Receipt Robustness Tests

Panel A: Math Value-Added

	(1) High Quality	(2) No Sort	(3) Grade Switch	(4) PARCC	(5) High Stakes
Tenure	-0.037***	-0.041***	-0.040***	-0.038***	-0.032***
	(0.013)	(0.013)	(0.010)	(0.011)	(0.011)
	[-0.119]	[-0.134]	[-0.130]	[-0.124]	[-0.104]
Num Schools	1,618	1,569	1,832	1,827	1,528
Num Teachers	2,013	1,850	3,199	3,120	2,569
Obs	334,481	339,216	397,131	388,186	331,446

Panel B: ELA Value-Added

	(1) High Quality	(2) No Sort	(3) Grade Switch	(4) PARCC	(5) High Stakes
Tenure	0.005	0.005	0.001	-0.009	0.008
	(0.012)	(0.014)	(0.014)	(0.011)	(0.012)
	[0.018]	[0.017]	[0.004]	[-0.031]	[0.028]
Num Schools	1,662	1,617	1,865	1,850	1,572
Num Teachers	2,290	2,074	3,562	3,467	3,099
Obs	430,883	426,977	491,214	478,726	564,811

Panel C: Summative Ratings

	(1) High Quality	(2) No Sort	(3) Grade Switch	(4) PARCC	(5) High Stakes
Tenure	-0.015	-0.006	NA	NA	-0.021
	(0.015)	(0.019)	NA	NA	(0.017)
	[-0.047]	[-0.018]	NA	NA	[-0.064]
Num Schools	1,345	1,141	NA	NA	1,246
Num Teachers	2,812	1,936	NA	NA	2,899
Obs	10,383	7,648	NA	NA	9,673

Notes: This table shows γ from equation (2) for the performance measure listed in the panel title. Only Panel C includes school and year fixed effects. Column (1) shows the effect on teachers whose year 3 summative rating exceeds the 25^{th} percentile of eventually tenured teachers. Column (2) shows the effect on teachers hired by 2013. Column (3) shows the effect when including an indicator for switching grades. Column (4) shows the effect when restricting the sample to PARCC tests. Column (5) shows the effect on grades 4 to 7 math and 4 to 8 ELA teachers.

Standard errors in parentheses and clustered at the school level. Performance units rescaled to standard deviation 1 in the dataset are included in brackets.

* p<0.10, ** p<0.05, *** p<0.01

Table 4: Productivity Effects at Tenure Receipt by School Characteristics

Panel A: Math Value-Added

	(1) Black: 20%+	(2) <20%	(3) P-Value	(4) Hispanic: 20%+	(5) <20%	(6) P-Value	(7) FRPL: 20%+	(8) <20%	(9) P-Value
Tenure	-0.026	-0.043***	0.460	-0.001	-0.058***	0.006	-0.011	-0.062***	0.020
	(0.018)	(0.013)		(0.014)	(0.015)		(0.012)	(0.018)	
	[-0.084]	[-0.138]		[-0.004]	[-0.189]		[-0.035]	[-0.201]	
Num Schools	942	1,566		1,160	1,357		1,427	951	
Num Teachers	1,021	2,326		1,431	1,919		2,063	1,262	
Obs	147,591	362,219		218,622	291,188		311,242	198,568	

Panel B: ELA Value-Added

	(1) Black: 20%+	(2) <20%	(3) P-Value	(4) Hispanic: 20%+	(5) <20%	(6) P-Value	(7) FRPL: 20%+	(8) <20%	(9) P-Value
Tenure	0.018	-0.002	0.414	0.053***	-0.037**	0.000	0.038***	-0.050***	0.000
	(0.021)	(0.014)		(0.017)	(0.015)		(0.015)	(0.018)	
	[0.064]	[-0.008]		[0.186]	[-0.128]		[0.133]	[-0.176]	
Num Schools	946	1,582		1,155	1,408		1,444	975	
Num Teachers	1,086	2,623		1,565	2,181		2,272	1,455	
Obs	163,238	477,281		255,192	385,327		375,390	265,129	

Panel C: Summative Ratings

	(1) Black: 20%+	(2) <20%	(3) P-Value	(4) Hispanic: 20%+	(5) <20%	(6) P-Value	(7) FRPL: 20%+	(8) <20%	(9) P-Value
Tenure	-0.049	-0.010	0.271	-0.033	-0.010	0.450	-0.019	-0.019	0.994
	(0.031)	(0.016)		(0.023)	(0.018)		(0.021)	(0.020)	
	[-0.151]	[-0.031]		[-0.101]	[-0.031]		[-0.061]	[-0.060]	
Num Schools	403	1,138		608	933		903	638	
Num Teachers	1,085	2,938		1,647	2,390		2,383	1,647	
Obs	3,533	10,362		5,561	8,334		8,235	5,660	

Notes: This table shows γ from equation (2) for the performance measure listed in the panel title. Only Panel C includes school and year fixed effects. Column (1) shows the effect for schools with at least 20% Black students, while Column (2) shows the effect for schools with less than 20% Black students. Column (3) provides the p-value from an F-test of equality for the coefficients. Columns (4)–(9) are defined similarly for school Hispanic and FRPL compositions.

Standard errors in parentheses and clustered at the school level. Performance units rescaled to standard deviation 1 in the dataset are included in brackets.

* p<0.10, ** p<0.05, *** p<0.01

Table 5: Productivity Effects at Tenure Receipt by School Proficiency

Panel A: Math Value-Added

	(1) Math Prof: 50%+	(2) <50%	(3) P-Value	(4) ELA Prof: 50%+	(5) <50%	(6) P-Value
Tenure	-0.051***	-0.010	0.056	-0.045***	-0.007	0.108
	(0.013)	(0.017)		(0.013)	(0.019)	
	[-0.166]	[-0.031]		[-0.144]	[-0.024]	
Num Schools	1,492	1,038		1,632	846	
Num Teachers	2,150	1,227		2,461	868	
Obs	333,238	175,600		393,503	115,335	

Panel B: ELA Value-Added

	(1) Math Prof: 50%+	(2) <50%	(3) P-Value	(4) ELA Prof: 50%+	(5) <50%	(6) P-Value
Tenure	-0.015	0.040**	0.024	-0.001	0.014	0.600
	(0.014)	(0.020)		(0.013)	(0.026)	
	[-0.052]	[0.140]		[-0.004]	[0.049]	
Num Schools	1,501	1,025		1,664	845	
Num Teachers	2,438	1,295		2,777	925	
Obs	444,206	194,578		509,792	128,992	

Panel C: Summative Ratings

	(1) Math Prof: 50%+	(2) <50%	(3) P-Value	(4) ELA Prof: 50%+	(5) <50%	(6) P-Value
Tenure	-0.014	-0.031	0.600	-0.012	-0.053	0.315
	(0.017)	(0.027)		(0.016)	(0.038)	
	[-0.045]	[-0.096]		[-0.039]	[-0.165]	
Num Schools	1,019	507		1,189	337	
Num Teachers	2,624	1,449		3,067	941	
Obs	9,176	4,690		10,850	3,016	

Notes: This table shows γ from equation (2) for the performance measure listed in the panel title. Only Panel C includes school and year fixed effects. Column (1) shows the effect for schools with at least 50% math proficient students, while Column (2) shows the effect for schools with less than 50% math proficient students. Column (3) provides the p-value from an F-test of equality for the coefficients. Columns (4)–(6) are defined similarly for ELA proficiency rates.

Standard errors in parentheses and clustered at the school level. Performance units rescaled to standard deviation 1 in the dataset are included in brackets.

* p<0.10, ** p<0.05, *** p<0.01

Table 6: Productivity Effects at Tenure Receipt by Student Characteristics

Panel A: Math Value-Added

	(1) Male	(2) Female	(3) P-Value	(4) White	(5) Black	(6) Hispanic	(7) P-Value	(8) FRPL	(9) Not FRPL	(10) P-Value
Tenure	-0.055***	-0.010	0.000	-0.047***	-0.030*	-0.007	0.046	-0.014	-0.048***	0.019
	(0.011)	(0.012)		(0.014)	(0.018)	(0.014)		(0.012)	(0.013)	
	[-0.178]	[-0.034]		[-0.153]	[-0.098]	[-0.022]		[-0.046]	[-0.157]	
Num Schools	1,761	1,737		1,737	1,737	1,737		1,676	1,721	
Num Teachers	3,198	3,186		3,186	3,186	3,186		3,085	3,156	
Obs	263,745	250,742		283,497	88,682	142,308		205,879	308,608	

Panel B: ELA Value-Added

	(1) Male	(2) Female	(3) P-Value	(4) White	(5) Black	(6) Hispanic	(7) P-Value	(8) FRPL	(9) Not FRPL	(10) P-Value
Tenure	-0.016	0.030**	0.000	-0.019	0.007	0.056***	0.002	0.040**	-0.014	0.004
	(0.012)	(0.013)		(0.013)	(0.020)	(0.019)		(0.016)	(0.013)	
	[-0.055]	[0.105]		[-0.067]	[0.023]	[0.195]		[0.138]	[-0.050]	
Num Schools	1,797	1,782		1,782	1,782	1,782		1,730	1,769	
Num Teachers	3,558	3,541		3,541	35,41	3,541		3,429	3,507	
Obs	332,165	313,642		375,751	98,289	171,767		244,551	401,256	

Notes: This table shows γ from equation (2) for the performance measure listed in the panel title. Column (1) shows the effect for male students, while Column (2) shows the effect for female students. Column (3) provides the p-value from an F-test of equality for the coefficients. Columns (4)–(10) are defined similarly for student race and FRPL eligibility.

Standard errors in parentheses and clustered at the school level. Performance units rescaled to standard deviation 1 in the dataset are included in brackets.

* p<0.10, ** p<0.05, *** p<0.01

Table 7: Dismissal Threat Effects

Panel A: Math Value-Added

	(1) All	(2) Male	(3) Female	(4) P-Value	(5) White	(6) Black	(7) Hispanic	(8) P-Value
Dismissal Threat	0.003	-0.037	0.020	0.509	-0.013	0.032	0.099	0.672
	(0.035)	(0.078)	(0.037)		(0.041)	(0.079)	(0.137)	
	[0.009]	[-0.120]	[0.065]		[-0.041]	[0.104]	[0.319]	
Obs	1,742	556	1,181		1,282	324	132	

Panel B: ELA Value-Added

	(1) All	(2) Male	(3) Female	(4) P-Value	(5) White	(6) Black	(7) Hispanic	(8) P-Value
Dismissal Threat	-0.018	-0.014	-0.018	0.937	-0.047*	0.036	0.103	0.121
	(0.024)	(0.045)	(0.028)		(0.027)	(0.050)	(0.090)	
	[-0.061]	[-0.050]	[-0.065]		[-0.163]	[0.127]	[0.361]	
Obs	8,134	1,575	6,554		5,879	1,537	721	

Panel C: Summative Ratings

	(1) All	(2) Male	(3) Female	(4) P-Value	(5) White	(6) Black	(7) Hispanic	(8) P-Value
Dismissal Threat	0.008	0.010	0.008	0.958	0.017	0.017	-0.073	0.360
	(0.021)	(0.036)	(0.026)		(0.026)	(0.049)	(0.059)	
	[0.026]	[0.032]	[0.025]		[0.052]	[0.052]	[-0.226]	
Obs	6,039	1,999	4,035		4,389	1,126	526	

Notes: This table shows γ from equation (5). I use a linear spline above and below the threshold to estimate $m(S_{j(t-1)})$. Based on the bandwidth selection method developed by Calonico et al. (2014), I use bandwidths of 0.193, 0.411, and 0.185 for math value-added (Panel A), ELA value-added (Panel B), and summative ratings (Panel C), respectively. I limit the sample to tenured teachers within these bandwidths. Column (1) shows the effect on all teachers. Columns (2) and (3) show the effect by interacting each independent variable in equation (5) with teacher gender. Column (4) provides the p-value from an F-test of equality for the coefficients. Columns (5)–(8) are defined similarly for teacher race. For race, the F-test evaluates whether the coefficients for Black and Hispanic teachers are jointly different from the coefficient for White teachers.

Standard errors in parentheses and clustered at the school level. Performance units rescaled to standard deviation 1 in the dataset are included in brackets.

* p<0.10, ** p<0.05, *** p<0.01

Table 8: Effect of TEACHNJ on Turnover Rates by Teacher Characteristics

	(1) All	(2) Male	(3) Female	(4) P-Value	(5) White	(6) Black	(7) Hispanic	(8) P-Value
Post	0.063***	0.076***	0.060***	0.000	0.061***	0.086***	0.060***	0.022
	(0.003)	(0.005)	(0.003)		(0.003)	(0.011)	(0.007)	
Mean	0.126	0.129	0.125		0.124	0.144	0.135	
Obs	673,601	140,347	533,220		578,442	48,754	47,058	

Notes: This table shows γ from equation (6). Column (1) shows the effect on all teachers. Columns (2) and (3) show the effect by interacting each independent variable in equation (6) with teacher gender. Column (4) provides the p-value from an F-test of equality for the coefficients. Columns (5)–(8) are defined similarly for teacher race. For race, the F-test evaluates whether the coefficients for Black and Hispanic teachers are jointly different from the coefficient for White teachers. The row titled "Mean" provides average turnover rates among pretenured teachers.

Standard errors in parentheses and clustered at the district level.

* p<0.10, ** p<0.05, *** p<0.01

Table 9: Effect of TEACHNJ on Turnover Rates by Summative Rating

Panel A: Effective and Highly Effective Teachers

	(1) All	(2) Male	(3) Female	(4) P-Value	(5) White	(6) Black	(7) Hispanic	(8) P-Value
Post	0.038***	0.045***	0.035***	0.038	0.037***	0.043***	0.037***	0.818
	(0.003)	(0.005)	(0.003)		(0.003)	(0.011)	(0.008)	
Mean	0.100	0.098	0.101		0.100	0.092	0.109	
Obs	424,522	93,883	330,603		370,325	25,468	28,966	

Panel B: Ineffective and Partially Effective Teachers

	(1) All	(2) Male	(3) Female	(4) P-Value	(5) White	(6) Black	(7) Hispanic	(8) P-Value
Post	0.214***	0.227***	0.205***	0.661	0.199***	0.266***	0.204**	0.340
	(0.050)	(0.073)	(0.047)		(0.048)	(0.069)	(0.097)	
Mean	0.396	0.423	0.381		0.420	0.353	0.332	
Obs	6,702	2,265	4,311		4,192	1,701	704	

Notes: This table shows γ from equation (6) by teacher summative rating. Panel A shows the effect for effective and highly effective teachers, while Panel B shows the effect for ineffective and partially effective teachers. Column (1) shows the effect on all teachers. Columns (2) and (3) show the effect by interacting each independent variable in equation (6) with teacher gender. Column (4) provides the p-value from an F-test of equality for the coefficients. Columns (5)–(8) are defined similarly for teacher race. For race, the F-test evaluates whether the coefficients for Black and Hispanic teachers are jointly different from the coefficient for White teachers. The row titled "Mean" provides average turnover rates among pretenured teachers.

Standard errors in parentheses and clustered at the district level.

* p<0.10, ** p<0.05, *** p<0.01

Table 10: Effects of TEACHNJ on Turnover by School Characteristics

Panel A: School Size and Poverty

	(1) 300+ Students	(2) <300 Students	(3) P-Value	(4) 20%+ FRPL	(5) <20% FRPL	(6) P-Value
Post	0.063***	0.061***	0.761	0.065***	0.060***	0.323
	(0.003)	(0.005)		(0.004)	(0.003)	
Mean	0.123	0.143		0.124	0.128	
Obs	562,793	87,250		389,256	260,679	

Panel B: School Racial Composition

	(1) 20%+ Black	(2) <20% Black	(3) P-Value	(4) 20%+ Hisp	(5) <20% Hisp	(6) P-Value
Post	0.064***	0.062***	0.795	0.064***	0.062***	0.749
	(0.006)	(0.003)		(0.005)	(0.003)	
Mean	0.134	0.122		0.124	0.127	
Obs	196,497	453,543		260,371	389,670	

Panel C: School Proficiency Rates

	(1) 50%+ Math Prof	(2) <50% Math Prof	(3) P-Value	(4) 50%+ ELA Prof	(5) <50% ELA Prof	(6) P-Value
Post	0.064***	0.061***	0.520	0.065***	0.059***	0.330
	(0.003)	(0.004)		(0.003)	(0.005)	
Mean	0.126	0.123		0.127	0.120	
Obs	326,364	273,896		405,127	195,012	

Notes: This table shows γ from equation (6). Panel A shows effects by school size and poverty levels. Panel B shows effects by school racial composition. Panel C shows effects by school proficiency rates. The column headers define the school characteristics. Columns (3) and (6) provide the p-value from an F-test of equality for the coefficients. The row titled "Mean" provides average turnover rates among pretenured teachers.

Standard errors in parentheses and clustered at the district level.

* p<0.10, ** p<0.05, *** p<0.01

Table 11: Teacher Demographics and Summative Ratings

Panel A: Gender

	(1) All Principals	(2) Male Principals	(3) Female Principals	(4) P-Value
Male Teacher	-0.031***	-0.028***	-0.035***	0.011
	(0.001)	(0.001)	(0.002)	
	[-0.096]	[-0.088]	[-0.110]	
Obs	319,693	183,926	135,766	

Panel B: Race

	(1) All Principals	(2) White Principals	(3) Black Principals	(4) Hispanic Principals	(5) P-Value
Black Teacher	-0.029***	-0.033***	-0.024***	-0.042***	0.017
	(0.002)	(0.003)	(0.003)	(0.007)	
	[-0.091]	[-0.102]	[-0.073]	[-0.129]	
Hispanic Teacher	-0.009**	-0.010**	0.005	-0.021*	0.111
	(0.004)	(0.004)	(0.010)	(0.011)	
	[-0.028]	[-0.030]	[0.015]	[-0.064]	
Obs	302,716	246,867	39,584	16,333	

Notes: This table shows γ from equation (7). In these regressions, $group_j$ defines teacher gender (Panel A) or race (Panel B). Column (1) shows the effect on all teachers. The remaining column headers define the effect when paired with a principal of a given gender or race. The final column provides the p-value from an F-test of equality for the coefficients.

Standard errors in parentheses and clustered at the school level. Performance units rescaled to standard deviation 1 in the dataset are included in brackets.

* p<0.10, ** p<0.05, *** p<0.01

Table 12: Tenure Extensive Margin Effects

	(1) Math VA	(2) ELA VA	(3) Ratings
Post	0.001	-0.003	0.021***
	(0.005)	(0.004)	(0.003)
	[0.002]	[-0.010]	[0.065]
Obs	29,463	33,078	129,457

Notes: This table shows γ from equation (6). I limit the sample to teachers in their first five years in a particular district. The column headers define the performance measure.

Standard errors in parentheses and clustered at the school level. Performance units rescaled to standard deviation 1 in the dataset are included in brackets.

* p<0.10, ** p<0.05, *** p<0.01

Figures

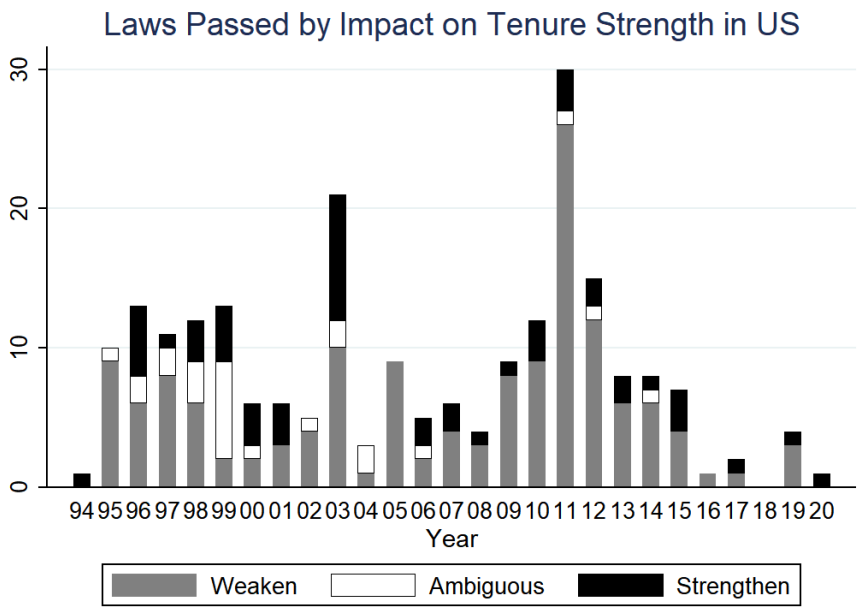
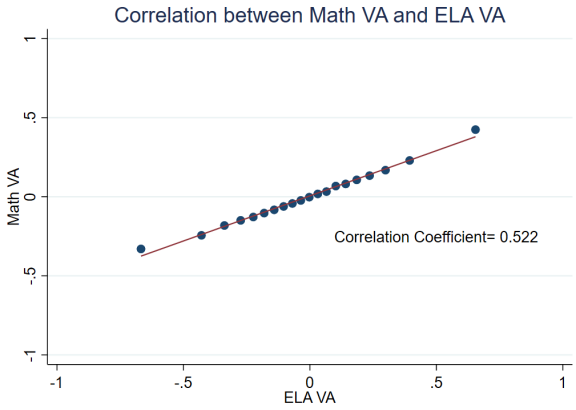


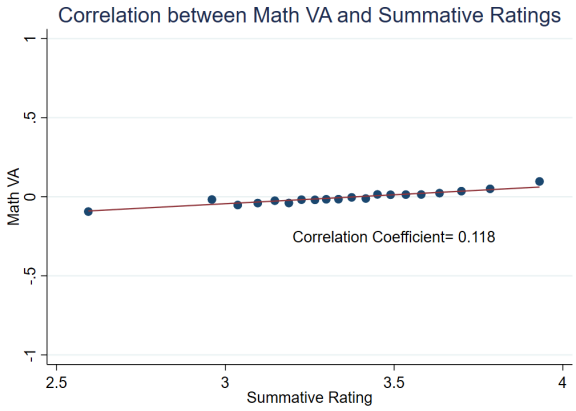
Figure 1: Laws Passed by Impact on Tenure Strength in the US

Notes: This figure records the number of tenure laws passed in the United States. The x-axis defines the year from 1994 to 2020, while the y-axis counts the number of laws. The laws are divided by their impact on tenure. Some examples of weakened tenure protections include extended pretenure periods and streamlined tenure dismissal policies. Some examples of strengthened tenure protections include shortened pretenure periods and narrowed scope for contract non-renewals. Laws that strengthen and weaken different components of tenure are listed as ambiguous. In total, this figure records 222 tenure laws across 49 states. This figure uses data from the Education Commission of the States.

Panel A: Math Value-Added and ELA Value-Added



Panel B: Math Value-Added and Summative Ratings



Panel C: ELA Value-Added and Summative Ratings

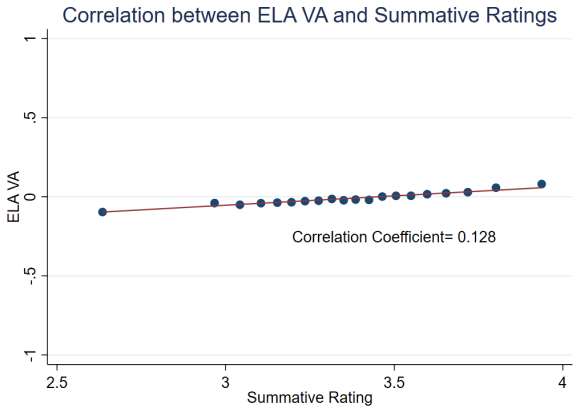


Figure 2: Correlation Between Performance Measures

Notes: This figure shows the relationship between performance measures. The x-axis and y-axis variables are labeled in each graph. The x-axis records the average value in 20 equal-sized bins, while the y-axis records the average value within that bin of the x-axis variable. The graphs include lines of best fit and correlation coefficients.

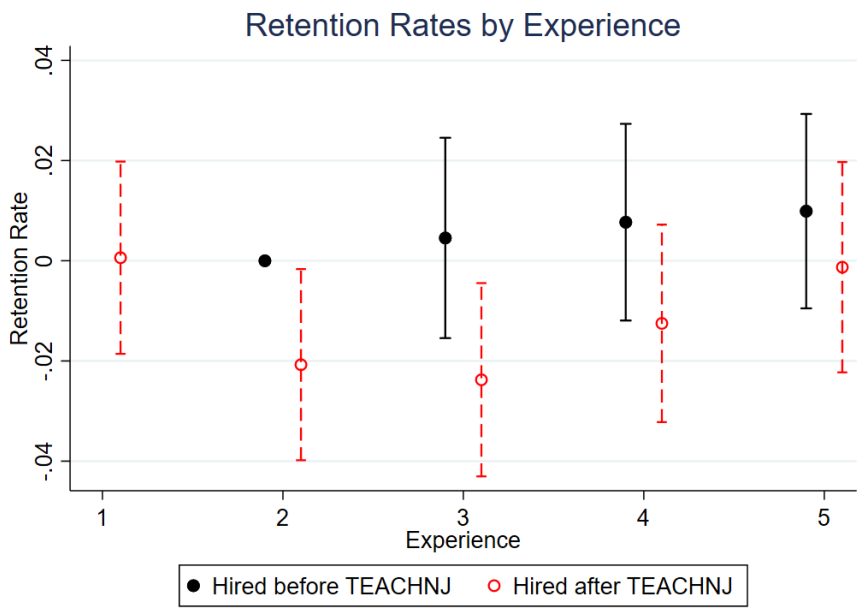
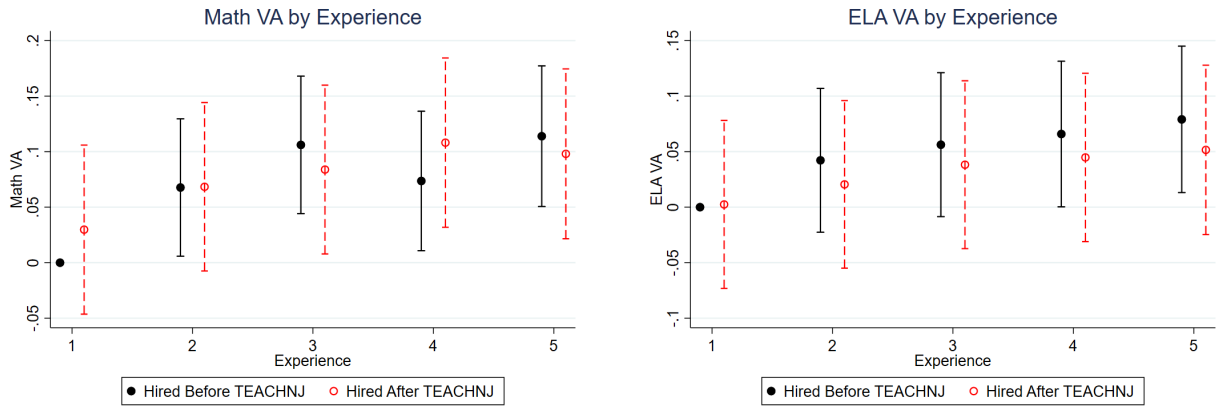


Figure 3: Retention Rates by Tenure Regime

Notes: This figure shows retention rates for those hired before and after TEACHNJ in their first district. I regress retention on experience fixed effects interacted with being hired before or after the law. I control for summative ratings to account for differences in teacher quality. I plot the estimated retention rates and 95% confidence intervals. The x-axis records experience, while the y-axis records the retention rates relative to second year teachers hired before TEACHNJ. This value is 0.916 when using the average summative rating in this sample. Year 1 retention for those hired before TEACHNJ is omitted because I do not have year 1 summative ratings for them. The solid black dots show the estimates for those hired before TEACHNJ. The hollow red dots show the estimates for those hired after TEACHNJ.

Panel A: Math Value-Added *Panel B: ELA Value-Added*



Panel C: Summative Ratings

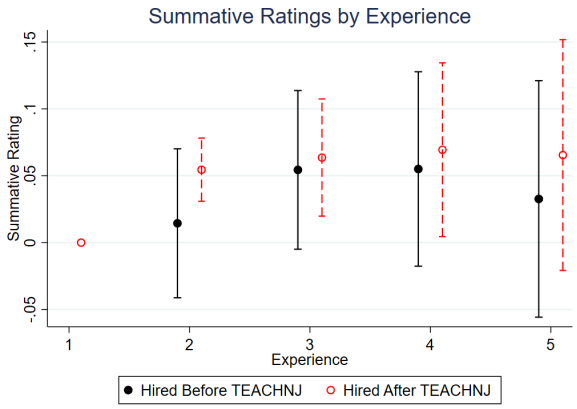


Figure 4: Performance by Experience and Tenure Regime

Notes: This figure shows the returns to experience separately for those hired before and after TEACHNJ. The panel title defines the performance measure. I plot the coefficients and 95% confidence intervals using equation (3). The x-axis records years of experience, while the y-axis records the performance level relative to the omitted group. For the value-added graphs, I omit the year 1 performance of those hired before TEACHNJ. These values are -0.10 and -0.07 standard deviations for math and ELA value-added, respectively. Since I do not have year 1 summative ratings for those hired before TEACHNJ, I use the year 1 performance for those hired after TEACHNJ as the omitted group for the summative rating graph. This value is 3.24 summative rating points. The solid black dots show the estimates for those hired before TEACHNJ, while the hollow red dots show the estimates for those hired after TEACHNJ.

Panel A: Dismissal Threat *Panel B: No Dismissal Threat*

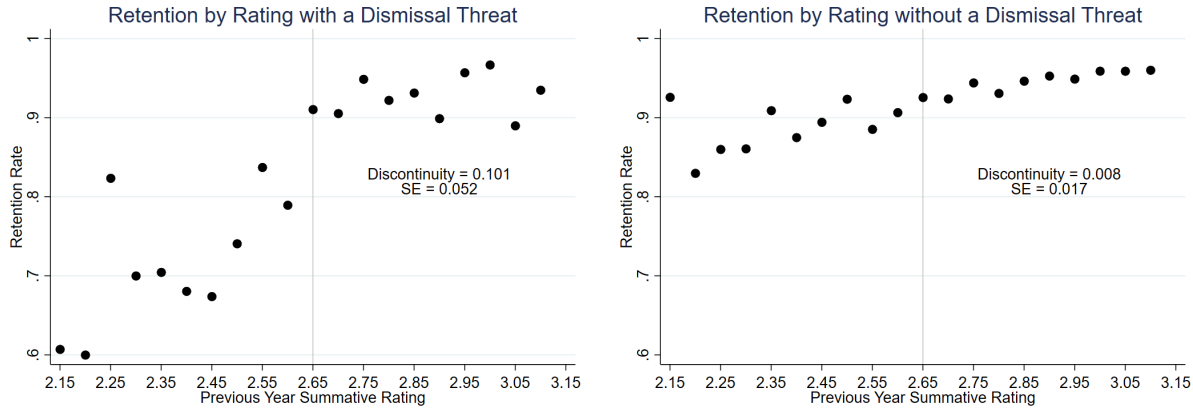
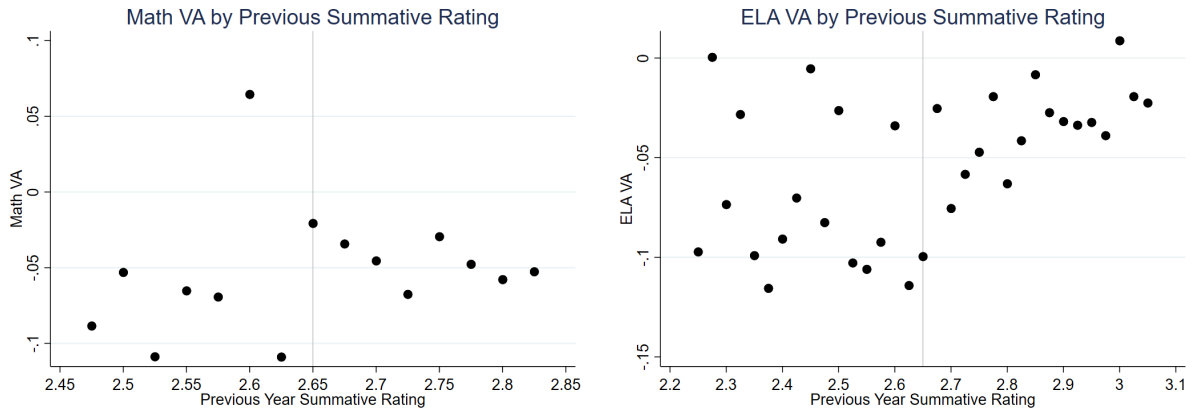


Figure 5: Fraction Returning by Summative Rating and Dismissal Threat

Notes: This figure shows retention rates for teachers based on their previous year summative rating. I limit the sample to tenured teachers. The x-axis records the teacher's previous year summative rating, while the y-axis measures the retention rates. Panel A includes teachers who received a partially effective or ineffective rating in the previous year, while Panel B includes teachers who received an effective or highly effective rating in the previous year. The scatterplot shows average retention in 0.05 summative rating point bins. In each graph, I include the estimated discontinuity and standard error generated using a linear spline above and below the threshold with the optimal bandwidth developed by Calonico et al. (2014). These bandwidths are 0.331 (Panel A) and 0.250 (Panel B) summative rating points.

Panel A: Math Value-Added *Panel B: ELA Value-Added*



Panel C: Summative Ratings

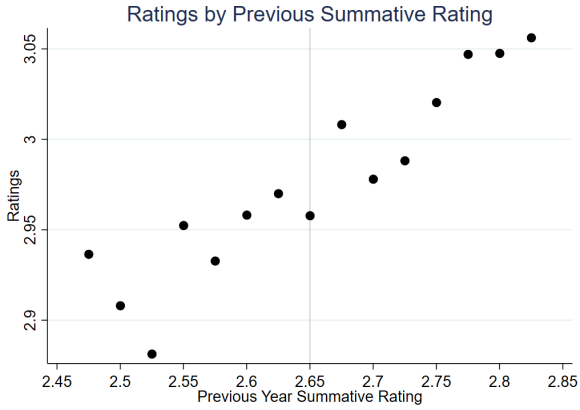


Figure 6: RDD Results

Notes: This figure shows RDD graphs for each of the performance measures using equation (5). Based on the bandwidth selection method developed by Calonico et al. (2014), I use bandwidths of 0.193, 0.411, and 0.185 for math value-added (Panel A), ELA value-added (Panel B), and summative ratings (Panel C), respectively. I limit the sample to tenured teachers within these bandwidths. The x-axis records the teacher's previous year summative rating, while the y-axis measures the performance measure. The scatterplot shows average performance in 0.025 summative rating point bins.

A Appendix

A.1 Relevance to Other States

While each state has a unique set of tenure policies, many components of New Jersey's tenure laws are common throughout the country (Thomsen, 2020). Specifically, the three- to four-year tenure length reform remains quite relevant because 32 states have three-year tenure clocks, while 4 states have four-year tenure clocks. In addition, the justifications for tenure dismissal in New Jersey include inefficiency, incapacity, or unbecoming conduct, which are quite common across states. The similarity in tenure laws across states extends this study's external validity.

A.2 New Jersey Summative Rating Implementation

Teacher summative ratings were carefully implemented in New Jersey following the passage of the TEACHNJ Act (State of New Jersey Department of Education, 2017). These ratings provided greater score differentiation than the previous two-tier rating system. In addition, teacher summative ratings have improved over time, which may be attributable to clearer expectations for good teaching, additional opportunities for feedback, and the use of data to improve teacher practice.

The NJDOE also provided districts with the autonomy to implement these systems. While this provided local control, it also allowed the distributions of teacher effectiveness to vary by district. In some districts, nearly every teacher received a highly effective rating. In other districts of similar sizes and student populations, teacher summative ratings are centered around effective ratings and distributed more normally. My specifications account for this variation by including either school, district, or principal fixed effects.

A.3 Difference-in-Differences One-Step VA Model

For value-added, I estimate a modified version of equation (2) using the following one-step model:

$$A_{ijgst} = \gamma ten_{jt} + \sum_{\tau=1}^{T} \delta_\tau \mathbf{1}(exp_{jt} = \tau) + \alpha A_{it-1} + \beta X_{it} + \eta C_{it} + \lambda S_{it} + \psi_j + \varepsilon_{ijgst}. \quad (8)$$

In this regression, A_{ijgst} is the test score of student i in teacher j's grade g class in school s and year t. I control for the student's previous year test score (A_{it-1}), as well as student, classroom, and school characteristics. The student controls (X_{it}) include gender, race, FRPL eligibility, ELL status, and special education status. The classroom variables (C_{it}) are class size and aggregated student controls. School covariates (S_{it}) include urbanicity, enrollment, school racial composition, and percentage of FRPL eligible. The model also controls for teacher fixed effects (ψ_j); ε_{ijgst} is a mean-zero error term.

A.4 Model Solution

I define a value function ($V_t(e_t)$) recursively for teachers retained in experience year t as:

$$V_t(e_t) = u_t(e_t) + emp_{t+1} V_{t+1} + (1 - emp_{t+1}) \sum_{\tau=t+1}^{\overline{T}} f(a_\tau, \tau).$$

In this equation, $u_t(e_t)$ is the utility from effort level e_t, while emp_{t+1} is an indicator for continued employment. The function $f(a_\tau, \tau)$ calculates the annual value of the outside option given ability a_t. Since teachers have perfect foresight, they may decide to quit prior to year t if:

$$V_t < \sum_{\tau=t}^{\overline{T}} f(a_\tau, \tau).$$

Thus, $emp_{t+1} = 1$ if the teacher meets the performance requirements ($m_{t+1} = 1$) and

the remaining value from teaching exceeds the value from the outside option ($V_{t+1} \geq \sum_{\tau=t+1}^{\overline{T}} f(a_\tau, \tau)$).

First, I solve for tenured teachers where $t \geq T$. Since these teachers still have their jobs, I know that $\prod_{\tau=1}^{T} \mathbf{1}(p_\tau \geq n)\left(\sum_{\tau=1}^{T} \frac{p_\tau}{T} \geq r\right) \prod_{\tau=T}^{t-1} \mathbf{1}(p_\tau \geq y) = 1$. In this equation, p_t is the teacher's performance, while n, y, and r are the annual pretenure, annual tenure, and tenure receipt requirements, respectively. In this section, I assume the utility from teaching exceeds that from the outside option.[49] Then, the value function simplifies to:

$$V_t(e_t) = u_t(e_t) + \mathbf{1}(p_t \geq y)V_{t+1} + \mathbf{1}(p_t < y) \sum_{\tau=t+1}^{\overline{T}} f(a_\tau, \tau).$$

Since job security only depends on the current year performance, e_t only impacts V_t through $u_t(e_t)$ and $\mathbf{1}(p_t \geq y)$. I take the first order condition with respect to effort and evaluate several cases. Without considering future employment, I find $e_t = e_t^*$ (which I call the single period optimal level of effort) and $p_t = a_t + e_t^*$. If I assume $p_t \geq y$, then the solution remains $e_t = e_t^*$ because the optimal single period utility level ensures an offer of continued employment. Even if teachers decide to quit after year t, e_t^* must maximize utility for period t or perfect foresight would have caused them to quit in a previous year.

Otherwise, $p_t < y$. In this case, the optimal solution is either $e_t = e_t^*$ or $e_t = y - a_t$. Any values between e_t^* and $y - a_t$ would not be optimal because the teacher would still not receive an offer of continued employment. However, they would move further away from the single period optimal solution. Since $u_t(e_t)$ is strictly concave and e_t^* is its unique global maximum, moving further away from e_t^* would result in less utility. Similarly, any values above $y - a_t$ would not gain any additional benefit of future employment but would reduce utility as they move further away from the single period optimal solution.

The optimal solution would be $e_t = e_t^*$ if:

[49] When simulating the model in Appendix Section A.4.1, I relax this assumption and incorporate quitting behavior.

$$V_t(e_t^*) \geq V_t(y - a_t)$$

$$u_t(e_t^*) + \sum_{\tau=t+1}^{\overline{T}} f(a_\tau, \tau) \geq u_t(y - a_t) + V_{t+1}$$

$$V_{t+1} \leq u_t(e_t^*) - u_t(y - a_t) + \sum_{\tau=t+1}^{\overline{T}} f(a_\tau, \tau).$$

Otherwise, the solution would be $e_t = y - a_t$. Thus, for tenured teachers:

$$e_t = \begin{cases} e_t^*, & e_t^* \geq y - a_t \text{ or } V_{t+1} \leq u_t(e_t^*) - u_t(y - a_t) + \sum_{\tau=t+1}^{\overline{T}} f(a_\tau, \tau) \\ y - a_t, & \text{otherwise.} \end{cases}$$

Tenured teachers continue to exert the single period optimal level of effort if it meets the annual performance standards or if the future value of employment is sufficiently low relative to additional effort. Otherwise, teachers increase their effort to maintain job security by meeting the annual performance standards.

For pretenured teachers, I can write the problem for $t < T$ as:

$$\begin{aligned} V_t(e_t) =& u_t(e_t) + \mathbf{1}(p_t \geq n)\Bigg(u_t(e_{t+1}) + \cdots \Bigg(u_t(e_T) + \mathbf{1}(p_T \geq n)\mathbf{1}\Big(\sum_{\tau=1}^{T} \frac{p_\tau}{T} \geq r \Big) V_{T+1} \\ &+ \mathbf{1}\Big(\sum_{\tau=1}^{T} \frac{p_\tau}{T} < r \Big) \sum_{\tau=T+1}^{\overline{T}} f(a_\tau, \tau) \Bigg) \cdots + \mathbf{1}(p_{t+1} < n) \sum_{\tau=t+2}^{\overline{T}} f(a_\tau, \tau) \Bigg) + \mathbf{1}(p_t < n) \sum_{\tau=t+1}^{\overline{T}} f(a_\tau, \tau). \end{aligned}$$

I solve the model in the following cases:

1. The single period optimal level of effort (e_t^*) produces a sufficiently high rating to receive tenure. This occurs when $e_t^* + \overline{a} \geq r$.

   (a) The initial performance given effort level e_t^* is sufficient to maintain employment. This occurs when $a_1 + e_t^* \geq n$.[50]

[50] Since a_t is increasing in t, $a_1 + e_t^* > n$ implies $a_t + e_t^* > n$ for all t.

(b) The performance given effort level e_t^* is insufficient to maintain employment in the first $\bar{t}$ years. This occurs when $a_t + e_t^* < n$ for $t \in \{1, \ldots, \bar{t}\}$.[51]

2. The single period optimal level of effort (e_t^*) does not produce a sufficiently high rating to receive tenure. This occurs when $e_t^* + \overline{a} < r$.

   (a) The initial performance given effort level e_t^* is sufficient to maintain employment into year two. This occurs when $a_1 + e_t^* \geq n$.

   (b) The performance given effort level e_t^* is insufficient to maintain employment in the first $\bar{t}$ years. This occurs when $a_t + e_t^* < n$ for $t \in \{1, \ldots, \bar{t}\}$.

      i. The minimum level of effort needed to earn tenure is sufficient to maintain employment. This occurs when $r - \overline{a} \geq n - a_t$ for all $t \in \{1, \ldots, T\}$, which is equivalent to $r - \overline{a} \geq n - a_1$, since a_t is strictly increasing in t.

      ii. The minimum level of effort needed to earn tenure is insufficient to maintain employment in the first $\widehat{t}$ periods. This occurs when $r - \overline{a} < n - a_t$ for $t \in \{1, \ldots, \widehat{t}\}$.

In the first case, suppose $e_t^* + \overline{a} \geq r$ and $a_1 + e_t^* \geq n$. The solution to the problem is e_t^* because the optimal single period effort level is sufficient to maintain employment.

In the next case, suppose $e_t^* + \overline{a} \geq r$ and $a_t + e_t^* < n$ for $t \in \{1, \ldots, \bar{t}\}$. If the cost of increased effort is sufficiently low, the teacher will exert effort level $e_t = n - a_t$ for $t \in \{1, \ldots, \bar{t}\}$. However, after period $\bar{t}$, the effort level returns to e_t^* because it is the single period optimal level and sufficient to receive tenure. Specifically, I assumed $e_t^* + \overline{a} \geq r$. Since $n - a_t > e_t^*$ for $t \in \{1, \ldots, \bar{t}\}$, effort levels $n - a_t$ for $t \in \{1, \ldots, \bar{t}\}$ and e_t^* for period $t \in \{\bar{t} + 1, \ldots, T\}$ produce a sufficient level of performance for tenure receipt.

For $t \in \{1, \ldots, \bar{t}\}$, teachers prefer $e_t = e_t^*$ to $e_t = n - a_t$ if:

[51] It is not possible that $n - a_t > r - \overline{a}$ for all $t \leq T$ because $r > n$ and $a_t \geq \overline{a}$ for some t by definition of $\overline{a}$.

$$V_t(e_t^*) \geq V_t(n - a_t)$$

$$u_t(e_t^*) + \sum_{\tau=t+1}^{\overline{T}} f(a_\tau, \tau) \geq u_t(n - a_t) + V_{t+1}$$

$$V_{t+1} \leq u_t(e_t^*) - u_t(n - a_t) + \sum_{\tau=t+1}^{\overline{T}} f(a_\tau, \tau).$$

For the next case, suppose $e_t^* + \overline{a} < r$ and $a_1 + e_t^* \geq n$. To receive tenure, teachers need $\sum_{\tau=1}^{T} \frac{e_t}{T} \geq r - \overline{a}$. Since there is no discounting and the utility curve is concave and decreasing near the global optimum (e_t^*), teachers evenly distribute effort throughout the pretenure years, so $e_1 = e_2 = \cdots = e_T$.[52].

Since e_t^* is sufficient to meet the annual retention requirements by assumption, I must compare the utility from $e_t = e_t^*$ for $t \in \{1, \ldots T\}$ to the utility from $e_t = r - \overline{a}$ for $t \in \{1, \ldots T\}$ plus the future utility from keeping their job with tenure (V_{T+1}). This decision should be made in year 1 to optimize, which is feasible due to perfect foresight. Thus, I compare the stream of utility from year 1. This occurs when:

$$V_1(e_t = e_t^* \text{ for } t \in \{1, \ldots, T\}) \geq V_1(e_t = r - \overline{a} \text{ for } t \in \{1, \ldots, T\})$$

$$\sum_{t=1}^{T} u_t(e_t^*) + \sum_{\tau=T+1}^{\overline{T}} f(a_\tau, \tau) \geq V_{T+1} + \sum_{t=1}^{T} u_t(r - \overline{a})$$

$$V_{T+1} \leq \sum_{t=1}^{T} \Big(u_t(e_t^*) - u_t(r - \overline{a}) \Big) + \sum_{\tau=T+1}^{\overline{T}} f(a_\tau, \tau).$$

For the next case, suppose $e_t^* + \overline{a} < r$, $a_t + e_t^* < n$ for $t \in \{1, \ldots, \overline{t}\}$, and $r - \overline{a} \geq n - a_1$. There are three potential options:

1. Provide effort e_t^* and lose the job in the next period if $t \leq \overline{t}$.

[52] I assumed that e_t does not interact with a_t for $t \neq 1$ multiplicatively. The first order condition for e_t does not depend on t and the shape of the utility curve is constant in e_t over time. Consequently, it is optimal to evenly distribute effort throughout the pretenure period

2. Provide effort $n - a_t$ and maintain employment until the next period (only for $t \leq \bar{t}$, otherwise e_t^* is sufficient).

3. Provide effort $r - \overline{a}$ until period T and receive tenure.

Please note that options (1) and (2) are redundant for periods $t > \bar{t}$, so I focus on option (1) on that interval.

In this case, $e_t = e_t^*$ if the utility from option (1) exceeds the utility from options (2) and (3). For the utility from option (1) to exceed the utility from option (2) in the first $\bar{t}$ periods, I need:

$$V_t(e_t^*) \geq V_t(n - a_t)$$
$$u_t(e_t^*) + \sum_{\tau=t+1}^{\overline{T}} f(a_\tau, \tau) \geq u_t(n - a_t) + \overline{V_{t+1}}$$
$$\overline{V_{t+1}} \leq u_t(e_t^*) - u_t(n - a_t) + \sum_{\tau=t+1}^{\overline{T}} f(a_\tau, \tau)$$

where $\overline{V_{t+1}}$ only provides future teaching value up to tenure receipt, as the performance level is inadequate to earn tenure. For the utility from option (1) to exceed the utility from option (3) in the first $\bar{t}$ periods, I need:

$$V_t(e_t = e_t^*) \geq V_t(e_s = r - \overline{a} \text{ for } s \in \{t, \ldots, T\})$$
$$u_t(e_t^*) + \sum_{\tau=t+1}^{\overline{T}} f(a_\tau, \tau) \geq \sum_{s=t}^{T} u_s(r - \overline{a}) + V_{T+1}$$
$$V_{T+1} \leq u_t(e_t^*) - \sum_{s=t}^{T} u_s(r - \overline{a}) + \sum_{\tau=t+1}^{\overline{T}} f(a_\tau, \tau).$$

For the utility from option (1) to exceed the utility from option (3) in periods $t \in \{\bar{t}, \ldots, T\}$, I need:

$$V_t(e_t = e_t^*) \geq V_t(e_s = r - \overline{a} \text{ for } s \in \{t, \ldots, T\})$$
$$u_t(e_t^*) + \overline{V_{t+1}} \geq \sum_{s=t}^{T} u_s(r - \overline{a}) + V_{T+1}$$
$$V_{T+1} - \overline{V_{t+1}} \leq u_t(e_t^*) - \sum_{s=t}^{T} u_s(r - \overline{a}).$$

Next, I compare the utility from option (2) and (3). The teacher will choose option (2) in periods $t \in \{1, \ldots, \bar{t}\}$ if the following holds:

$$V_t(e_t = n - a_t) \geq V_t(e_s = r - \overline{a} \text{ for } s \in \{t, \ldots, T\})$$
$$u_t(n - a_t) + \overline{V_{t+1}} \geq \sum_{s=t}^{T} u_s(r - \overline{a}) + V_{T+1}$$
$$V_{T+1} - \overline{V_{t+1}} \leq u_t(n - a_t) - \sum_{s=t}^{T} u_s(r - \overline{a}).$$

For the final case suppose $e_t^* + \overline{a} < r$, $a_t + e_t^* < n$ for $t \in \{1, \ldots, \bar{t}\}$, and $r - \overline{a} < n - a_t$ for $t \in \{1, \ldots, \widehat{t}\}$. There are three potential options:

1. Provide effort e_t^* and lose the job in the next period if $t \leq \bar{t}$.

2. Provide effort $n - a_t$ and maintain employment until the next period (only for $t \leq \bar{t}$, otherwise e_t^* is sufficient).

3. Provide effort $n - a_t$ until the period $\overline{\tau}$. For the remaining pretenure years, the teacher provides effort $\frac{T(r-\overline{a}) - \sum_{t=1}^{\bar{t}}(n-a_t)}{T - \overline{\tau} - 1}$. Let $\overline{\tau}$ be defined as the largest $t \in \mathbb{Z}$ such that $\frac{T(r-\overline{a}) - \sum_{t=1}^{\bar{t}}(n-a_t)}{T - \overline{\tau} - 1} < n - a_t$. In this case, the teacher can evenly split his or her effort and still receive tenure.

Again, please note that options (1) and (2) are redundant for periods $t > \bar{t}$, so I focus on option (1) on that interval. In addition, options (2) and (3) are redundant for periods $t \leq \overline{\tau}$,

so I focus on option (2) on that interval.

To calculate the values for option (3), I define $\overline{\tau}$ as the period where I can evenly split remaining effort and still meet the minimum annual requirements. I know that it is optimal to evenly distribute effort across the remaining periods since the utility function is concave and decreasing above e_t^*. Thus, $e_{\overline{\tau}} = e_{\overline{\tau}+1} = \cdots = e_T$. Then, I solve the following equation showing the minimum effort needed to earn tenure after year $\overline{\tau}$ for e_t:

$$\begin{aligned} T(r-\overline{a}) &= \sum_{t=1}^{\overline{\tau}}(n-a_t) + \sum_{t=\overline{\tau}+1}^{T} e_t \\ (T-\overline{\tau}-1)e_t &= T(r-\overline{a}) - \sum_{t=1}^{\overline{t}}(n-a_t) \\ e_t &= \frac{T(r-\overline{a}) - \sum_{t=1}^{\overline{t}}(n-a_t)}{T-\overline{\tau}-1} \end{aligned}$$

where $\overline{\tau}$ occurs as the largest $t \in \mathbb{Z}$ such that $\frac{T(r-\overline{a})-\sum_{t=1}^{\overline{t}}(n-a_t)}{T-\overline{\tau}-1} < n-a_t$. To simplify notation, let $\overline{e} = \frac{T(r-\overline{a})-\sum_{t=1}^{\overline{t}}(n-a_t)}{T-\overline{\tau}-1}$.

The relationship between options (1) and (2) are the same as the previous case, so I focus on the relationship between options (1) and (3) and options (2) and (3).

If $\overline{\tau} < t < \overline{t}$, option (1) is preferred to option (3) when:

$$\begin{aligned} V_t(e_t = e_t^*) &\geq V_t(e_s = \overline{e} \text{ for } s \in \{t, \ldots, T\}) \\ u_t(e_t^*) + \sum_{\tau=t+1}^{\overline{T}} f(a_\tau, \tau) &\geq \sum_{s=t}^{T} u_s(\overline{e}) + V_{T+1} \\ V_{T+1} &\leq u_t(e_t^*) - \sum_{s=t}^{T} u_s(\overline{e}) + \sum_{\tau=t+1}^{\overline{T}} f(a_\tau, \tau). \end{aligned}$$

If $\overline{t} < T < \overline{T}$, option (1) is preferred to option (3) when:

$$V_t(e_t = e_t^*) \geq V_t(e_s = \overline{e} \text{ for } s \in \{t, \ldots, T\})$$

$$u_t(e_t^*) + \overline{V_{t+1}} \geq \sum_{s=t}^{T} u_s(\overline{e}) + V_{T+1}$$

$$V_{T+1} - \overline{V_{t+1}} \leq u_t(e_t^*) - \sum_{s=t}^{T} u_s(\overline{e}).$$

Finally, I must compare the utility from option (2) and (3). The stream of efforts are identical until $\overline{\tau}$ in these two cases, so I only consider $t \in \{\overline{\tau}, \ldots, \bar{t}\}$. The teacher will choose option (2) if the following holds:

$$V_t(e_t = n - a_t) \geq V_t(e_s = \overline{e} \text{ for } s \in \{t, \ldots, T\})$$

$$u_t(n - a_t) + \overline{V_{t+1}} \geq \sum_{s=t}^{T} u_s(\overline{e}) + V_{T+1}$$

$$V_{T+1} - \overline{V_{t+1}} \leq u_t(n - a_t) - \sum_{s=t}^{T} u_s(\overline{e}).$$

Combining these solutions, I find the following effort levels:

$$
e_t = \begin{cases}
n - a_t, & t \le \bar{t} \\
& \text{and } \left\{ \left(e_t^* + \bar{a} \ge r \text{ and } a_t + e_t^* < n \text{ for } t \in \{1, \ldots, \bar{t}\} \text{ and } V_{t+1} > u_t(e_t^*) - u_t(n - a_t) + \sum_{\tau=t+1}^{\overline{T}} f(a_\tau, \tau) \right) \right. \\
& \text{or } \left(e_t^* + \bar{a} < r \text{ and } a_t + e_t^* < n \text{ for } t \in \{1, \ldots, \bar{t}\} \text{ and } r - \bar{a} \ge n - a_1 \right. \\
& \left. \text{and } \overline{V_{t+1}} > u_t(e_t^*) - u_t(n - a_t) + \sum_{\tau=t+1}^{\overline{T}} f(a_\tau, \tau) \text{ and } V_{T+1} - \overline{V_{t+1}} \le u_t(n - a_t) - \sum_{s=t}^{T} u_s(r - \bar{a}) \right) \\
& \text{or } \left(e_t^* + \bar{a} < r \text{ and } a_t + e_t^* < n \text{ for } t \in \{1, \ldots, \bar{t}\} \text{ and } r - \bar{a} < n - a_t \text{ for } t \in \{1, \ldots, \widehat{t}\} \right. \\
& \text{and } \overline{V_{t+1}} > u_t(e_t^*) - u_t(n - a_t) + \sum_{\tau=t+1}^{\overline{T}} f(a_\tau, \tau) \\
& \left. \left. \left. \text{and } \left\{ t \le \overline{\tau} \text{ or } \left(t > \overline{\tau} \text{ and } V_{T+1} - \overline{V_{t+1}} \le u_t(n - a_t) - \sum_{s=t}^{T} u_s(\overline{e}) \right) \right\} \right) \right\} \right) \right\} \\
r - \bar{a}, & t \le T \text{ and } \left\{ \left(e_t^* + \bar{a} < r \text{ and } a_1 + e_t^* \ge n \text{ and } V_{T+1} > \sum_{t=1}^{T} \left(u_t(e_t^*) - u_t(r - \bar{a}) \right) + \sum_{\tau=T+1}^{\overline{T}} f(a_\tau, \tau) \right) \right. \\
& \text{or } \left(e_t^* + \bar{a} < r \text{ and } a_t + e_t^* < n \text{ for } t \in \{1, \ldots, \bar{t}\} \text{ and } r - \bar{a} \ge n - a_1 \right. \\
& \text{and } \left\{ \left(t \le \bar{t} \text{ and } V_{T+1} > u_t(e_t^*) - \sum_{s=t}^{T} u_s(r - \bar{a}) + \sum_{\tau=t+1}^{\overline{T}} f(a_\tau, \tau) \right) \right. \\
& \text{and } V_{T+1} - \overline{V_{t+1}} > u_t(n - a_t) - \sum_{s=t}^{T} u_s(r - \bar{a}) \\
& \left. \left. \left. \text{or } \left(t > \bar{t} \text{ and } V_{T+1} - \overline{V_{t+1}} > u_t(e_t^*) - \sum_{s=t}^{T} u_s(r - \bar{a}) \right) \right\} \right) \right\} \\
\overline{e}, & \overline{\tau} < t \le T \text{ and } e_t^* + \bar{a} < r \text{ and } a_t + e_t^* < n \text{ for } t \in \{1, \ldots, \bar{t}\} \text{ and } r - \bar{a} < n - a_t \text{ for } t \in \{1, \ldots, \widehat{t}\} \\
& \text{and } \left\{ \left(t < \bar{t} \text{ and } V_{T+1} > u_t(e_t^*) - \sum_{s=t}^{T} u_s(\overline{e}) + \sum_{\tau=t+1}^{\overline{T}} f(a_\tau, \tau) \right) \right. \\
& \text{and } V_{T+1} - \overline{V_{t+1}} > u_t(n - a_t) - \sum_{s=t}^{T} u_t(\overline{e}) \Big) \\
& \left. \text{or } \left(t > \bar{t} \text{ and } V_{T+1} - \overline{V_{t+1}} > u_t(e_t^*) - \sum_{s=t}^{T} u_s(\overline{e}) \right) \right\} \\
y - a_t, & t > T \text{ and } e_t^* < y - a_t \text{ and } V_{t+1} > u_t(e_t^*) - u_t(y - a_t) + \sum_{\tau=t+1}^{\overline{T}} f(a_\tau, \tau) \\
e_t^*, & \text{otherwise.}
\end{cases}
$$

Let T be the last pretenure year, $\bar{t}$ be the last year such that $e_t^* < n - a_t$, $\widehat{t}$ be the last year such that $r - \bar{a} < n - a_t$, $\overline{\tau}$ be the last year such that $\frac{T(r-\bar{a}) - \sum_{t=1}^{\bar{t}} (n - a_t)}{T - \overline{\tau} - 1} < n - a_t$, $\overline{e} = \frac{T(r-\bar{a}) - \sum_{t=1}^{\bar{t}} (n - a_t)}{T - \overline{\tau} - 1}$, and $\sum_{\tau=1}^{T} \frac{a_t}{T} = \bar{a}$. Let V_t be the future utility value, while $\overline{V_{t+1}}$ is the future value when dismissed at tenure receipt. The utility from the outside option is $f(a_t, t)$.

A.4.1 Simulating Model

To solve the vector of effort inputs, I use backward induction. In retirement year $\overline{T}$, the optimal level of effort is e_t^* because future employment is irrelevant. Using this utility value, I solve the optimal decision for experience year $\overline{T}-1$. Teachers leave the profession if they are not offered continued employment (when student performance fails to achieve the minimum standards) or quit (when the utility from the outside option exceeds that from teaching). In this case, I replace the future recursive value with the stream of utility from the outside option, so $V_{\overline{T}-1} = u_{\overline{T}-1}(e_t^*) + f(a_T, T)$. I then iterate back to experience year 1 to solve the stream of effort and performance levels.

I simulate data to estimate the model. I suppose initial teacher ability (a_1) follows a standard normal distribution. I assume teachers retire after 25 years because it is the minimum time needed to collect early retirement. Following Figure 2 of Wiswall (2013), I allow ability to increase by one standard deviation over 24 years of experience. Specifically, ability in experience year t is defined as $a_t = a_1 + \sqrt{\frac{t-1}{24}}$. I assume the outside option is defined as $f(a_t, t) = \kappa a_t - \mathrm{e}^{-\mathrm{t}}|\mathrm{s}|$, where $\kappa = 0.3$ and $s \sim N(0, 100)$.

Suppose teacher utility is $u_t(e_t) = bp_t - e_t^2$, where b gives the relative utilities of student test scores and teacher effort. With perfect job security, the teacher's optimal effort is $e^* = \frac{b}{2}$, which does not depend on t. Student test scores are then $p_t = a_t + \frac{b}{2}$.

For Table 1, I set the performance standards to reflect worst case scenario decreases in tenure receipt rates after TEACHNJ. For those hired before TEACHNJ, 30% of teachers left the district before tenure. Figure 3 shows the annual difference in retention between the top of the 95% confidence interval before TEACHNJ and the bottom of the 95% confidence interval after TEACHNJ is about 6 percentage points. This would increase turnover by about 24 percentage points over the four pretenure years and equate to a 54% pretenure turnover rate. Thus, I set $y = 2.2$ and $n = 2.7$ in all models. Before TEACHNJ, I set $T = 3$ and $r = 3.05$. After TEACHNJ, I set $T = 4$ and $r = 5.45$. These values do not have simple interpretations other than they allow the model to match actual tenure receipt rates.

A.5 Productivity Effects Later in Career Assumptions

To estimate causal effects, I assume teachers above and below the threshold perform similarly in the absence of dismissal threats (Imbens & Lemieux, 2008). Specifically, the conditional regression function measuring a teacher's performance without a dismissal threat must be continuous in summative ratings through the threshold. In addition, the conditional distribution function must be continuous through the threshold. I provide three tests to evaluate these smoothness assumptions.

To evaluate covariate balance around the threshold, I use the stacked regression test proposed by Lee and Lemieux (2010). Specifically, I estimate a standard RDD but replace the outcome with a series of predetermined variables. These variables include teacher experience, gender, and race, as well as school characteristics. The school characteristics are school enrollment, as well as the fraction of Black, Hispanic, FRPL, ELL, special education, math proficient, and ELA proficient students. I then estimate the following regression:

$$X_{jtx} = \beta_x S_{j(t-1)x} + \delta_x S_{j(t-1)x} * D_{jtx} + \gamma_x D_{jtx} + \mu_x + \varepsilon_{jtx}.$$

In this regression, I interact each predetermined variable (X_{jtx}) with linear splines ($S_{j(t-1)x}$), treatment indicators (D_{jtx}), and intercepts (μ_x). The coefficients of interest (γ_x) measure the difference in each predetermined variable at the threshold. The smoothness assumptions require $\gamma_x = 0$ for all x, which imply that the covariates do not change discontinuously at the threshold. When conducting a joint hypothesis test that $\gamma_x = 0$ for each predetermined variable, I calculate a p-value of 0.4716. Thus, I fail to reject the null hypothesis and provide evidence of covariate balance across the threshold.

I conduct a second covariate balance test that predicts teacher value-added and summative ratings in year t (y_{jt}) using the aforementioned teacher and school characteristics (X_{jt}). If the predetermined covariates are balanced through the threshold, predicted performance relying on these covariates also should be smooth through the threshold. I regress

performance on predetermined characteristics and plot the predictions in Figure A5. I also include estimated discontinuities using linear splines above and below the threshold with the optimal bandwidth developed by Calonico et al. (2014). Figure A5 shows no discontinuity in the performance measures at the threshold, which provides further evidence of covariate balance near the threshold.

In addition to testing for covariate balance, I evaluate whether there is bunching on either side of the threshold. I conduct the McCrary (2008) running variable density test to evaluate the conditional density smoothness through the threshold. A discontinuity in the conditional density provides evidence of sorting that may bias the results. Figure A6 shows little to no sorting at the threshold rating of 2.65. While more teachers earn effective ratings than partially effective ratings, the density does not change discontinuously at the threshold. In fact, sorting at the threshold is impractical because part of the summative ratings relies on test score performance (see Table A2). Since supervisors and teachers do not have perfect control over these components, intentional sorting is unlikely. In combination, these three tests suggest the smoothness assumptions are reasonable in this context.

A.6 RDD Summative Rating Comparison to Literature

The RDD from Section 5 shows dismissal threats have no impact on productivity. These results differ from Dee and Wyckoff (2015), who find positive productivity effects. However, the results do not contradict each other because the credibility of the threats may vary across regions, while the New Jersey sample is much more experienced than the Washington D.C. sample.[53]

Although the actual credibility of the dismissal threats may be stronger in Washington D.C., I argue that the perceived credibility among teachers across the regions was similar. As discussed in Section 5, the credibility of the threats relies on two factors: the language

[53] Otherwise, the two environments are quite similar. After the first low rating, both sets of teachers construct a Corrective Action Plan with their principals. In addition, the NJDOE lists Washington D.C.'s IMPACT tool as an approved evaluation rubric.

of the law and the dismissal rates conditional on low summative ratings. While the actual policy language is similar in both regions, teachers would remain unaware of the dismissal rates conditional on confidential ratings. As a result, teachers use the language of the law and any observed performance-related dismissals to measure the perceived credibility of the dismissal threats.[54]

However, New Jersey teachers facing dismissal threats are much more experienced than Washington D.C. teachers. In the New Jersey RDD sample, the average experience was 15 years compared to 7 years in the Washington D.C. sample.[55] More experienced teachers may demonstrate fewer productivity effects for a variety of reasons. For example, effort becomes relatively less important as innate ability improves due to experience. While novice teachers must construct new lesson plans, experienced teachers may reuse lessons from previous years. Thus, the effects of weaker job security may diminish over time and generate the differences in results.

[54] I cannot completely rule out the actual credibility as a mechanism because Washington D.C. schools only retained about 50% of their minimally (partially) effective teachers, while New Jersey schools retained about 75% of their partially effective teachers.

[55] I calculated the average experience in the Washington D.C. sample by using a weighted average of the experience range midpoints from Dee and Wyckoff (2015).

A.7 Appendix Tables

Table A1: Summary Statistics

	Students	Teachers
Female	0.484	0.798
	(0.500)	(0.402)
Black	0.197	0.080
	(0.398)	(0.272)
Hispanic	0.271	0.075
	(0.445)	(0.263)
Urban	0.911	0.913
	(0.285)	(0.281)
FRPL	0.377	
	(0.485)	
ELL	0.045	
	(0.207)	
Special Ed.	0.194	
	(0.395)	
Math Proficient	0.528	
	(0.499)	
ELA Proficient	0.582	
	(0.493)	
Experience		13.276
		(8.904)
Years in District		11.644
		(8.264)
Summative Rating		3.386
		(0.322)
Obs	12,405,063	905,574
Unique Obs	2,164,750	231,815

Notes: This table provides summary statistics at the student-year and teacher-year levels. The row headers define the variable. The first column provides the student-year summary statistics, while the second column provides the teacher-year summary statistics. The standard deviations of each value are listed in parentheses below the means. The final two rows count the number of observations and the number of unique individuals in the sample.

Table A2: Summative Rating Weights By Year and Subject

	2014, 2017, 2018		2015, 2016	
	ELA 4-8 Math 4-7	Other	ELA 4-8 Math 4-7	Other
Teacher Practice	55%	85%	70%	80%
SGO - District	15%	15%	20%	20%
mSGP - State	30%		10%	

Notes: This table shows summative rating weights. The first two columns record the weights for the academic years ending in 2014, 2017, and 2018. The first column provides weights for high stakes subjects where standardized tests impact the summative ratings. The second column provides weights for all other teachers. The third and fourth columns are defined similarly for the academic years ending in 2015 and 2016. In this table, SGO and mSGP are acronyms for Student Growth Objectives and median Student Growth Percentiles, respectively.

Table A3: Sample Restrictions

	Math VA	ELA VA	Ratings
All Teachers	50,835	56,445	154,670
Has Non-Missing Student Data	38,715	43,082	51,814
Has Value-Added in Years 2 and 3	3,199	3,562	6,480
Has Summative Ratings in Years 2 and 3	NA	NA	3,871

Notes: This table shows the number of observations remaining after each sample restriction. The first column records the number of teachers used for the math value-added analysis. The second and third columns are defined similarly for ELA value-added and summative ratings, respectively. The first row includes all teachers with the performance measure listed in the column header. In the second row, I restrict the sample to math and ELA teachers in tested grades with non-missing student, class, and school characteristics. In the third row, I restrict the sample to teachers with value-added in years 2 and 3. This provides multiple observations of performance prior to tenure receipt. In the final row, I restrict the sample to teachers with summative ratings in years 2 and 3. I only use this restriction for the summative rating analysis.

Table A4: Tenure Effects on Class Composition

	(1) Size	(2) % Black	(3) % Hisp.	(4) % Female	(5) % FRPL	(6) % Spec. Ed.	(7) % ELL	(8) New Grade
Tenure	0.191	-0.017*	0.003	0.002	0.024*	-0.003	-0.005	-0.080***
	(0.390)	(0.010)	(0.012)	(0.006)	(0.014)	(0.014)	(0.005)	(0.015)
Num Classes	22,718	22,572	22,624	22,624	22,622	22,496	22,623	18,312

Notes: This table shows γ from equation (4). The column headers define the dependent variables.
Standard errors in parentheses and clustered at the school level.
* p<0.10, ** p<0.05, *** p<0.01

Table A5: Dismissal Threat Effects Bandwidth Sensitivity

Panel A: Math Value-Added

	(1) 0.1	(2) 0.2	(3) 0.3	(4) 0.4	(5) 0.5	(6) Full Sample	(7) MSE-Optimal
Dismissal Threat	-0.035	-0.005	0.015	0.007	-0.022	-0.016	0.003
	(0.055)	(0.034)	(0.029)	(0.025)	(0.024)	(0.016)	(0.035)
	[-0.112]	[-0.016]	[0.047]	[0.023]	[-0.072]	[-0.052]	[0.009]
Obs	753	1,878	3,259	7,354	11,823	60,008	1,742

Panel B: ELA Value-Added

	(1) 0.1	(2) 0.2	(3) 0.3	(4) 0.4	(5) 0.5	(6) Full Sample	(7) MSE-Optimal
Dismissal Threat	-0.013	-0.024	-0.009	-0.017	-0.053**	-0.016	-0.018
	(0.054)	(0.033)	(0.028)	(0.024)	(0.022)	(0.017)	(0.024)
	[-0.045]	[-0.083]	[-0.033]	[-0.058]	[-0.184]	[-0.057]	[-0.061]
Obs	713	1,854	3,301	7,733	12,478	66,436	8,134

Panel C: Summative Ratings

	(1) 0.1	(2) 0.2	(3) 0.3	(4) 0.4	(5) 0.5	(6) Full Sample	(7) MSE-Optimal
Dismissal Threat	-0.000	0.005	0.003	-0.010	0.033**	0.108***	0.008
	(0.029)	(0.021)	(0.017)	(0.016)	(0.015)	(0.015)	(0.021)
	[-0.000]	[0.016]	[0.009]	[-0.031]	[0.102]	[0.337]	[0.026]
Obs	2,828	7,081	12,907	33,361	55,747	324,132	6,039

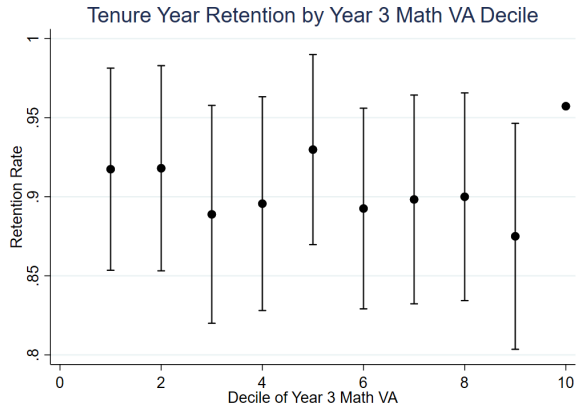
Notes: This table evaluates the bandwidth sensitivity of γ from equation (5). The column headers define the bandwidth in value-added standard deviations or summative rating points. Using the bandwidth selection method developed by Calonico et al. (2014), the MSE-Optimal bandwidths are 0.193, 0.411, and 0.185 for math value-added (Panel A), ELA value-added (Panel B), and summative ratings (Panel C), respectively. I limit the sample to tenured teachers within the given bandwidth.

Standard errors in parentheses and clustered at the school level. Performance units rescaled to standard deviation 1 in the dataset are included in brackets.

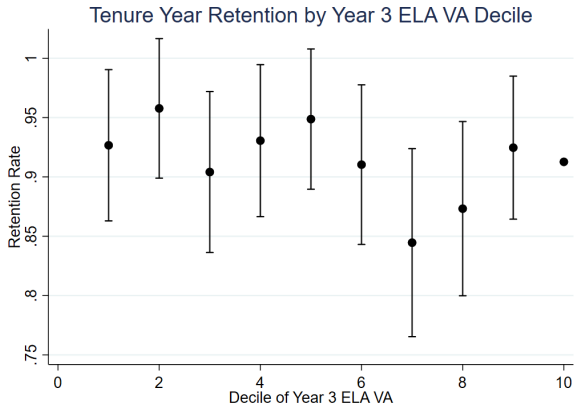
* p<0.10, ** p<0.05, *** p<0.01

A.8 Appendix Figures

Panel A: Math Value-Added



Panel B: ELA Value-Added



Panel C: Summative Ratings

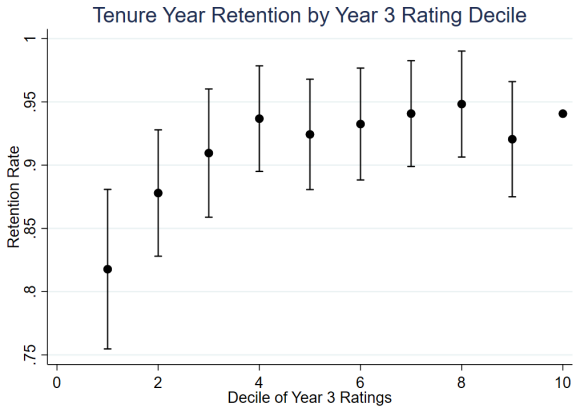
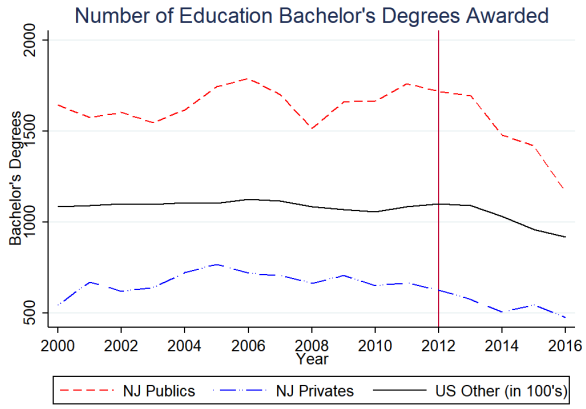


Figure A1: Tenure Year Retention by Decile

Notes: This figure plots retention rates and 95% confidence intervals by decile of year 3 performance. The x-axis records the decile of year 3 math value-added (Panel A), ELA value-added (Panel B), or summative rating (Panel C). The y-axis measures the retention rates. The omitted group is the top decile of year 3 performance.

Panel A: Education Bachelor's Degrees



Panel B: Education Master's Degrees

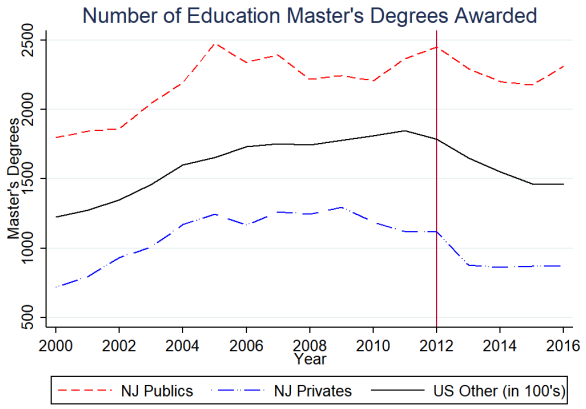


Figure A2: Trends in Education Degrees

Notes: This figure shows the number of bachelor's (Panel A) and master's (Panel B) degrees awarded by New Jersey public colleges, New Jersey private colleges, and all other United States colleges. The x-axis records the year, while the y-axis records the number of degrees awarded. The dashed red lines show New Jersey public colleges. The dashed and dotted blue lines show New Jersey private colleges. The solid black lines show all other United States colleges in hundreds of degrees. The red vertical lines illustrate the passage of TEACHNJ in 2012. This figure uses data from the Integrated Postsecondary Education Data System (IPEDS) of the National Center for Education Statistics.

Panel A: Turnover Rates by Experience *Panel B: Turnover Rates by Calendar Year*

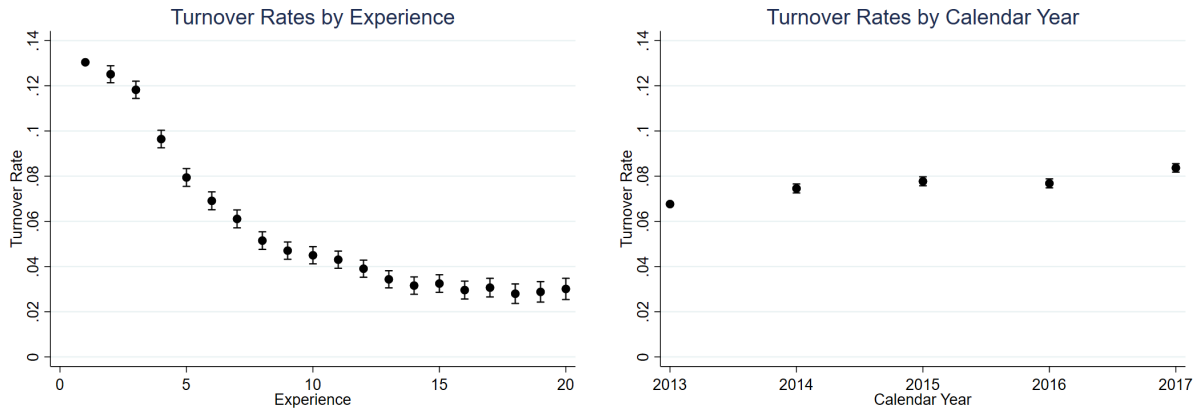
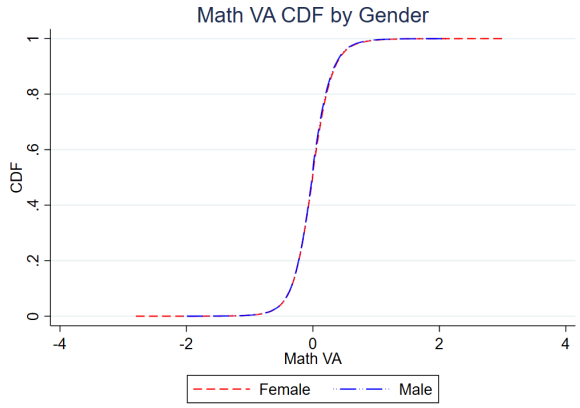


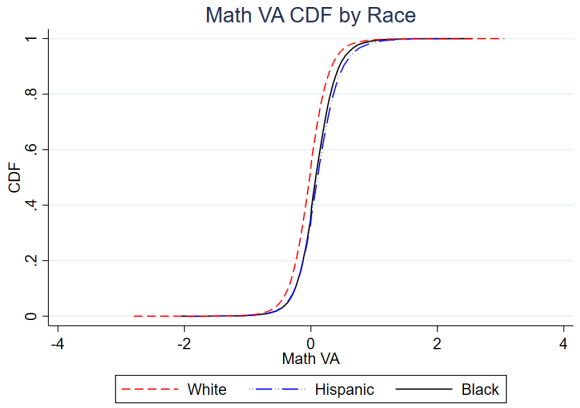
Figure A3: Turnover Rates

Notes: This figure shows the turnover rates and 95% confidence intervals by experience and calendar year. The omitted groups are teachers with 1 year of experience (Panel A) and teachers in 2013 (Panel B). In Panel A, I also omit teachers with more than 20 years of experience.

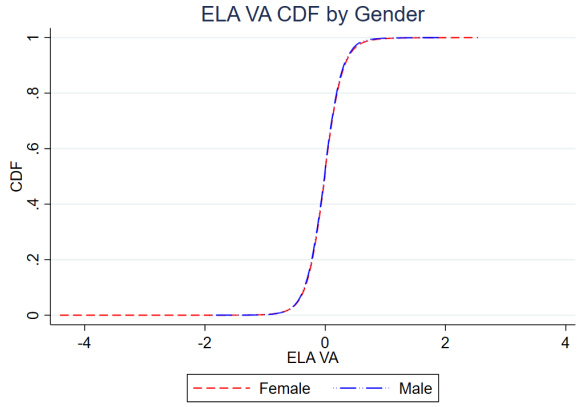
Panel A: Math Value-Added by Gender



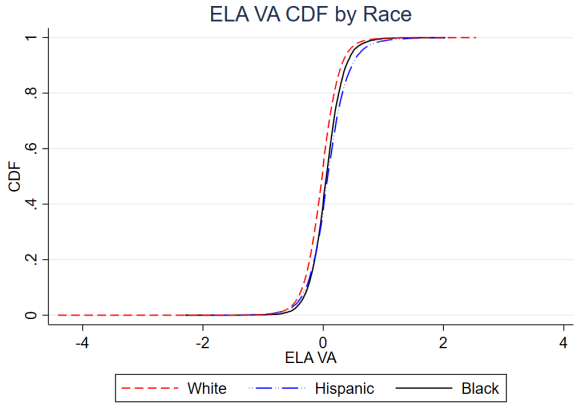
Panel B: Math Value-Added by Race



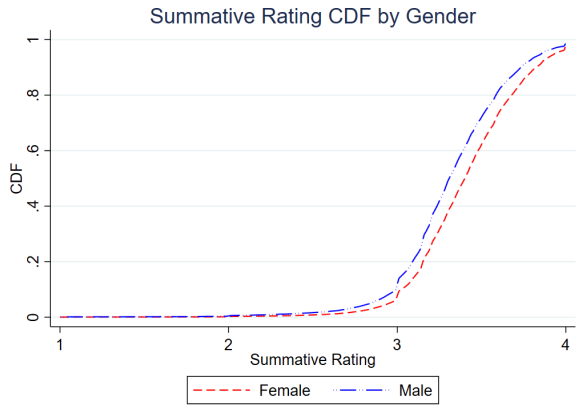
Panel C: ELA Value-Added by Gender



Panel D: ELA Value-Added by Race



Panel E: Ratings by Gender



Panel F: Ratings by Race

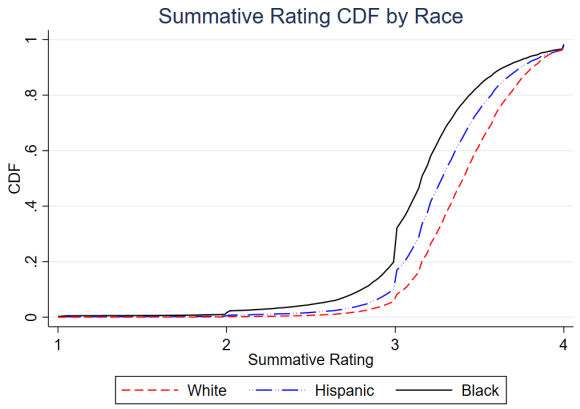
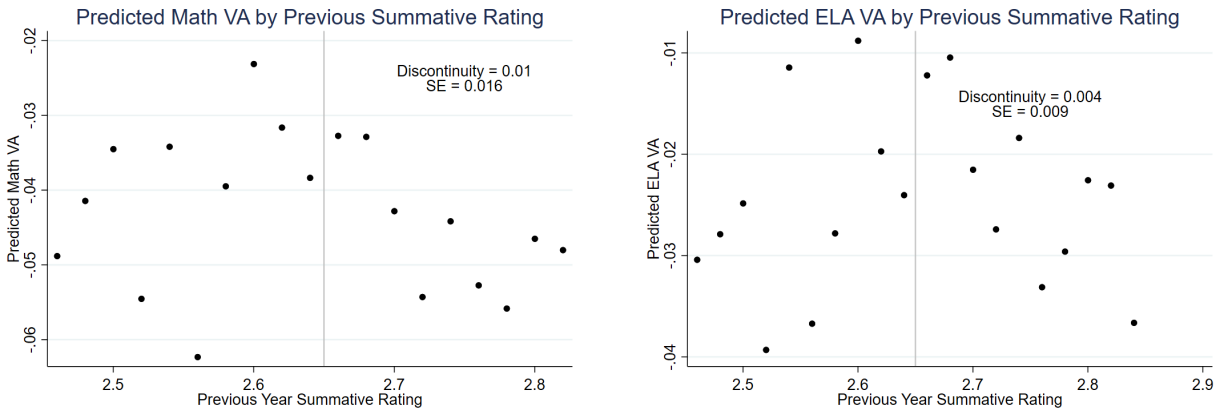


Figure A4: Performance CDF by Gender and Race

Notes: This figure shows the cumulative density of performance by gender (Panels A, C, and E) and race (Panels B, D, and F). The x-axis records performance, while the y-axis records the density. For gender (race), dashed red lines show female (White) teachers, while dashed and dotted blue lines depict male (Hispanic) teachers. Solid black lines show Black teachers.

Panel A: Math Value-Added *Panel B: ELA Value-Added*



Panel C: Summative Ratings

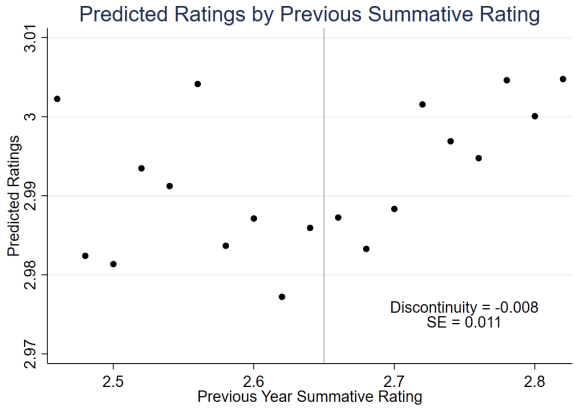


Figure A5: Predicted Performance by Previous Summative Rating

Notes: This figure shows RDD scatterplots when predicting each performance measure using only predetermined characteristics. I limit the sample to tenured teachers. The x-axis records the teacher's previous year summative rating, while the y-axis measures the predicted performance measure. The scatterplot shows average performance in 0.02 summative rating point bins. Each graph includes the estimated discontinuity and standard error generated by using a linear spline above and below the threshold with the optimal bandwidth developed by Calonico et al. (2014). Specifically, I use bandwidths of 0.193, 0.411, and 0.185 for math value-added (Panel A), ELA value-added (Panel B), and summative ratings (Panel C), respectively.

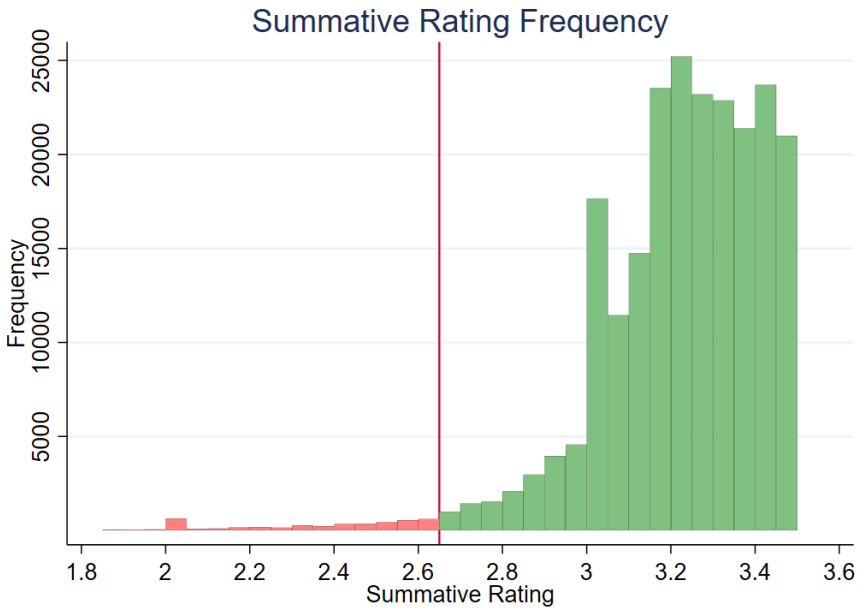


Figure A6: Frequency of Previous Summative Ratings

Notes: This figure shows the McCrary (2008) running variable density test. I limit the sample to tenured teachers. The x-axis records the teacher's previous year summative rating. The red vertical line separates partially effective teachers from effective teachers. The red bars count partially effective ratings, while the green bars count effective ratings.