Identifying High-Quality Teachers

Kevin Ng

March 21, 2022

Abstract

This study evaluates techniques to identify high-quality teachers. Since tenure restricts dismissals of experienced teachers, schools must predict productivity and dismiss those expected to perform ineffectively prior to tenure receipt. Many states solely rely on evaluation scores to guide these personnel decisions without considering other dimensions of teacher performance. For example, New Jersey uses summative ratings, which primarily rely on supervisor evaluations. I use various predictive models to rank teachers based on predicted value-added and summative ratings. I then simulate revised personnel decisions and test for changes in average retained teacher performance. In this exercise, I adjust two factors that impact the quality of the predictions: the number of predictors and the length of the pretenure period. Both factors impact the precision of the predictions, though extended pretenure periods also may negatively impact selection into teaching. I find that revised algorithms using both value-added and summative ratings increase the average value-added of retained teachers by 0.01 standard deviations without decreasing summative ratings or diversity. This Pareto improvement equates to a present value gain of $2,240 per student. These returns are a product of using additional information rather than advanced algorithms, as I generate similar gains when using simple ordinary least squares regressions or advanced machine learning techniques. In comparison, algorithms that extend the pretenure period beyond one year do not provide enough additional information to significantly improve average retained teacher performance unless dismissal rates increase dramatically.

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

Public schools seek to retain high-quality teachers to improve student achievement. However, teacher tenure restricts dismissals of experienced educators. Prior to tenure receipt, schools must predict productivity and dismiss those expected to perform ineffectively. While these predictions could incorporate multiple dimensions of performance, many states solely rely on evaluations based on classroom observations. Evaluations measure characteristics that are distinct from value-added, such as lesson planning, classroom management, and professionalism. Without considering all measures of performance, schools may be ignoring important information when making these choices.

Even if schools use all available dimensions of performance, each annual metric only provides a noisy signal of quality. Identifying the optimal pretenure period length is critical to providing more reliable information about teacher ability while still attracting high-quality educators through compensating differentials associated with tenure. Previous research has been unable to evaluate the returns to additional performance metrics and longer pretenure periods because few datasets contain evaluations that are directly tied to personnel decisions, as well as test score data used to calculate value-added.

To evaluate the returns to utilizing additional performance metrics and longer pretenure periods, I use predictive models relying on either machine learning techniques or ordinary least squares (OLS) regressions. In analogous settings, several studies have used machine learning algorithms to predict performance (Kleinberg et al., 2017; Athey et al., 2007; Chandler et al., 2011; Abaluck et al., 2016). These studies leverage the strengths of machine learning (making predictions) rather than its weaknesses (estimating causal effects) (Mullainathan & Spiess, 2017; Kleinberg et al., 2015). For example, Kleinberg et al. (2017) evaluate whether these algorithms can improve bail decisions by simultaneously minimizing jailing and crime rates. I apply these techniques to evaluate whether these algorithms can generate a Pareto improvement in average retained teacher quality by simultaneously (weakly) increasing value-added and summative ratings. I also assess whether these improvements are a product of

using additional information or sophisticated algorithms.

In this paper, I use teacher-student linked administrative data from the New Jersey Department of Education (NJDOE) to explore the relationship between previous and subsequent value-added and summative ratings. I calculate value-added using a lagged test score model. The NJDOE provides annual summative ratings which are based on a combination of supervisor classroom observations and test scores. These ratings[1] serve as the sole determinant of performance-based personnel decisions in New Jersey. The study maintains external validity, as every state relies on some combination of supervisor evaluations and test scores to assess their teachers (Ross & Walsh, 2019). Relative to the New Jersey benchmark, I increase the weight on value-added relative to ratings and record the change in average retained teacher performance. I utilize various algorithms, such as random forests, to predict subsequent productivity based on previous performance. Using these predictions, I simulate revised personnel decisions that dismiss the bottom 10% of teachers, which approximately matches current pretenure dismissal rates. I measure the retained teachers' subsequent value-added and ratings to compare the algorithms. This study explores two factors that impact predictions: the number of performance measures and the length of the pretenure period.

Similar to Kleinberg et al. (2017), this analysis relies on imputed data. I compare the subsequent performance of teachers retained using the current system to that of teachers retained using rankings based on the revised algorithms. Since I do not observe the subsequent performance of teachers who leave the profession, I must impute their performance and assume that unobserved characteristics do not bias this prediction. While Kleinberg et al. (2017) rely on quasi-random assignment to strict and lenient judges to evaluate this selection-on-observables assumption, I leverage district strictness. Districts retain some discretion when dismissing low-performing teachers, so I compare strict districts with higher dismissal rates conditional on summative ratings to lenient districts with lower dismissal rates. If strict districts select on unobserved characteristics, imputations relying on these

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

teachers would overpredict performance in lenient districts. However, this test shows no evidence of prediction bias.

Using these assumptions, I then explore the returns to additional measures of teacher performance when using 10% dismissal rates. By supplementing summative ratings with value-added rather than following current policies that only consider ratings, districts can increase subsequent average value-added by 0.01 standard deviations without decreasing ratings or teacher diversity. This Pareto improvement equates to a present value gain of $2,240 per student (Chetty et al., 2014), which is over 9 times larger than the costs associated with the productivity decline at tenure receipt (Ng, 2021). These improvements stem from using additional information rather than just applying advanced machine learning techniques, as I generate similar gains when using OLS regressions. By using additional measures of performance, the proposed algorithms provide a virtually costless method to improve average retained teacher quality.

The Pareto improvement is feasible because value-added and summative ratings are not fixed over time. If they were constant, increasing retained teacher value-added would mechanically reduce retained teacher ratings, as the current system would already have maximized ratings. Since value-added and ratings vary over time, current retention decisions relying on previous ratings do not perfectly predict subsequent ratings. Incorporating value-added generates a Pareto improvement because the reweighted algorithm continues to predict ratings as effectively, while also improving the value-added predictions.

Next, I reestimate the predictive models when adjusting the pretenure period. At current dismissal rates of 10%, extending the pretenure period beyond one year does not significantly improve average teacher quality. Teacher performance approximately follows a normal distribution, so the bottom decile lies in the far left tail. In this region, quality is so widely dispersed that even noisy annual estimates often correctly classify these teachers. Thus, longer pretenure periods provide little additional information at the cost of reduced compensating differentials. In comparison, higher dismissal rates allow extended pretenure periods

to generate significant improvements in average teacher performance. Since more teachers are clustered near the middle of the distribution, even a little bit of noise could result in teachers being misclassified and dismissed. Consequently, additional pretenure years only provide valuable information when ranking teachers near the middle of the distribution.

Combining these results, I find that incorporating additional measures of teacher performance is a more effective technique to select high-quality teachers than extending the pretenure period. Unlike extending the pretenure period with a 10% dismissal rate, using additional measures generates statistically significant gains in average retained teacher quality. In addition, longer pretenure periods reduce compensating differentials, while additional measures of performance are unlikely to impact selection into teaching.

This paper contributes to the literature by estimating the returns to using additional performance measures and longer pretenure periods. I use value-added and summative ratings to make these predictions because previous work finds other characteristics, such as educational attainment and licensure test scores, are not correlated with subsequent teacher value-added (Hanushek, 1997; Buddin & Zamarro, 2009; Chingos & Peterson, 2011). Prior research predicts subsequent teacher performance using survey evaluation data with both novice and experienced teachers (Harris & Sass, 2014; Winters & Cowen, 2013; Chalfin et al., 2016). Since the returns to experience vary throughout a teacher's career (Kraft & Papay, 2015; Wiswall, 2013), predictions relying on experienced teachers may be inapplicable to novices. Using a rich dataset in a populous state, I restrict my analysis to teachers at the beginning of their careers. Previously, this restriction was infeasible because most datasets have too few novice teachers. In addition, prior work often relied on value-added or principal surveys as proxies for current retention decisions. In comparison, my study uses the actual rankings based on administrative summative rating data. Since this dataset provides a large sample of novice teachers, as well as the actual metrics used to inform personnel decisions, my paper provides the most comprehensive analysis of tenure receipt decisions to date.

2 Data

I use the NJDOE's teacher-student linked administrative test score data from 2012 to 2018. These math and English language arts (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.[2] These data include student gender, race, Free or Reduced-Price Lunch (FRPL) eligibility, English Language Learner (ELL) status, and special education status. The dataset also contains teacher gender, race, and experience.[3]

I use all New Jersey teachers to estimate the following model:

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

where A_{ijgst} is the test score of student i in teacher j's grade g class in school s and year t.[4] 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, ELL status, and special education status. The classroom controls (C_{it}) are class size and aggregated student controls. School covariates (S_{it}) include [5], enrollment, racial composition, and percentage of FRPL eligible.[6] Value-added is measured annually by Θ_{jt}.

To calculate career value-added, I use equation (1) but replace teacher-year fixed effects (Θ_{jt}) with teacher fixed effects (Θ_j). Thus, I estimate the following model:

$$A_{ijgst} = \alpha A_{it-1} + \beta X_{it} + \eta C_{it} + \lambda S_{it} + \Theta_j + \varepsilon_{ijgst}. \tag{2}$$

[2] Appendix Section A.1 addresses concerns about the transition to the PARCC exam in 2015.

[3] 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.

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

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

[6] The main results are robust to using school fixed effects rather than S_{it} to calculate value-added.

To avoid mechanical correlations, I estimate equation (2) by excluding any years that the machine learning algorithm uses as predictors. For example, if the algorithm predicts career value-added using annual value-added in years 1, 2, and 3, then I only use data after year 3 to estimate equation (2).

Summative ratings from 2014 to 2018 provide an additional measure of teacher quality.[7] These confidential ratings measure performance using a weighted average of Teacher Practice, Student Growth Objectives (SGO), and median Student Growth Percentiles (mSGP). In Teacher Practice, supervisors observe several classes using an NJDOE approved rubric. These rubrics evaluate various teaching competencies, such as lesson planning, classroom management, and professionalism. Administrators and teachers in each district collaborate to design their own SGO based on state standards. The SGO measure student growth based on the percentage of students improving their scores. Grades 4 to 8 ELA and grades 4 to 7 math teachers rely on mSGP, which measure score growth 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.[8]

Table A2 shows the weighting schemes for 2014 and 2017–2018 (first two columns), as well as 2015–2016 (last two columns).[9] Summative 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).

I limit the sample to teachers who have summative ratings and value-added estimates for their first three years of experience.[10] Using these restrictions, I focus the analysis on novice

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

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

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

[10] While all teachers receive summative ratings, this restriction limits the sample to only math and ELA teachers.

teachers with multiple measures of performance.[11]

2.1 Value-Added and Summative Ratings Correlation

Before evaluating the predictive models, I estimate the correlation between value-added and summative ratings among all teachers in Figure 1. As described in Jacob and Lefgren (2008), I adjust for measurement error in the value-added estimates. Panel A shows the correlation between math value-added and itself (solid black), ELA value-added (dashed red), and ratings (dashed and dotted blue). The x-axis records time between observations, so concurrent measures occur at $x = 0$, while measures separated by 5 years occur at $x = 5$. Panels B and C are designed similarly for ELA value-added and summative ratings.

In Figure 1, the correlations are stronger within performance measures than across them when holding time between observations constant. Correlations weaken as time increases. In fact, the solid black and dashed red lines in Panels A and B show that the correlation between concurrent math and ELA value-added exceeds the autocorrelation within either value-added measure over time. This suggests math and ELA value-added capture similar components of teacher effectiveness.

Figure 1 also shows that contemporaneous value-added and summative ratings only have correlation coefficients of about 0.12. Thus, ratings primarily capture elements of teacher effectiveness that are not measured by value-added. It is important to consider both metrics when making personnel decisions because improving teacher performance along one dimension does not necessarily increase performance along the other dimensions.

[11] 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 year 1 has the largest effect on sample size.

3 Empirical Analysis

3.1 Machine Learning Algorithm

In this section, I use random forests to predict subsequent performance. Random forests estimate a series of regression trees where each tree predicts subsequent performance by splitting the sample at nodes based on previous performance. While regression trees can perfectly fit in-sample data, this procedure would lead to overfitting for out-of-sample predictions. To overcome this source of bias, random forests create 500 bootstrapped datasets. In each dataset, I estimate a regression tree based on a randomly selected $\frac{1}{3}$ of the total regressors. To estimate these algorithms, I use 40% of the sample to impute missing performance data, another 40% to train the algorithm, and the remaining 20% to conduct the analysis.

I estimate several models using both value-added and ratings. First, I calculate mean summative ratings in the first three years. This method does not use machine learning techniques but closely reflects the current system in New Jersey. As discussed in Appendix Section A.3, performance-based personnel decisions solely rely on ratings.[12] In fact, Figure 2 shows mean pretenure ratings are positively correlated with retention rates.

Next, I consider machine learning algorithms. Using random forests, I rely on one performance measure to predict subsequent performance along the same metric. For example, I use previous math value-added to predict subsequent math value-added. I then repeat the process for ELA value-added and summative ratings.[13]

However, these algorithms are limited by a multidimensionality problem. Since random forests only permit one outcome variable and value-added is weakly correlated with summative ratings, the algorithms will struggle to generate improvements along all dimensions of performance simultaneously. Instead, the algorithms will maximize the outcome variable and have little effect on the other dimensions. I can reduce this multidimensional prob-

[12] In New Jersey, highly-rated pretenure teachers may still be dismissed without cause. However, I focus my analysis on performance-related dismissals.

[13] Using two metrics to predict subsequent performance produces similar results (not shown).

lem into a single dimension by using a composite measure that is a weighted average of value-added and summative ratings. Specifically, I standardize each performance measure to mean 0 and standard deviation 1 within a given experience year. I generate a single value-added measure by averaging non-missing standardized math and ELA value-added. I then construct a weighted average of the combined value-added measure and standardized summative rating.[14] Using a random forest, I predict subsequent composite values using previous composite measures.

After generating these algorithms, I use the predicted subsequent performance to rank teachers and simulate personnel decisions. I dismiss the lowest ranked teachers and compare the average subsequent performance of retained teachers across algorithms. However, prior to proceeding with the analysis, I address concerns about missing data.

3.2 Imputing Missing Data

To evaluate the prediction function, I compare the subsequent performance of teachers retained under the current system to that of teachers retained using the predictive models. However, I do not observe the performance of teachers who leave the profession. Thus, I encounter a one-sided problem for teachers who were dismissed and left the profession under the current system but would be retained using other algorithms.[15] To address this concern, I must impute the subsequent performance of teachers who left the profession.

Imputed performance may be biased if retention criteria rely on unobserved characteristics that impact subsequent performance (Kleinberg et al., 2017). In this context, there is limited scope for using unobserved traits because ratings capture many characteristics that would typically be unobserved, such as ineffective pedagogy and poor professionalism. Nonetheless, I must impute $E[y_j|x_j] = E[f(x_j)|x_j] = f(x_j)$ for teachers who leave the profession, where y_j is the subsequent performance of teacher j and $f(x_j)$ is a flexible function

[14] My main specification places equal weight on ratings and value-added, as recommended by the first component of a principal component analysis.

[15] I do not encounter this problem for teachers who were dismissed and switched districts.

of the teacher's previous summative rating, x_j. However, the data only allow me to estimate $E[y_j|x_j, r_j = 1]$, where r_j is an indicator function defining retention. If the conditional expectation is independent of retention, the imputation will be unbiased. Thus, I assume:

$$E[y_j|x_j] = E[y_j|x_j, r_j = 1]. \tag{3}$$

To evaluate this assumption, I leverage district strictness. Districts retain some discretion when dismissing low-performing teachers. As discussed in Appendix Section A.3, teachers who receive consecutive partially effective ratings may be dismissed for cause, though districts also may provide one more opportunity to improve. This allows districts to retain teachers using unobserved characteristics that are not captured by ratings. For instance, supervisors may recognize that one teacher earning low summative ratings has great potential, so they offer an additional opportunity for this teacher to improve.

Suppose these unobserved characteristics, s_j, are independent of x_j and increase performance additively by $g(s_j)$ where $g(\cdot)$ is a flexible function and $E[g(s_j)] = 0$. Then, I can rewrite the conditional expectation as follows:

$$\begin{aligned} E[y_j|x_j, r_j = 1] &= E[f(x_j)|x_j, r_j = 1] + E[g(s_j)|x_j, r_j = 1] \\ &= f(x_j) + E[g(s_j)|r_j = 1]. \end{aligned}$$

For equation (3) to hold, I must show that $E[g(s_j)|r_j = 1] = 0$. First, I partition the sample into "lenient" districts that only rely on observed characteristics and "strict" districts that also consider unobserved characteristics. For example, lenient districts retain all teachers near the margin of ineffective teaching, while strict districts only keep marginal teachers if they have great potential that is not reflected in the ratings. Referring back to the theoretical framework, let $F(\cdot)$ and $G(\cdot)$ be flexible functions. In lenient districts, $r_j = 1$ if $F(x_j) > 0$ because they only rely on observed characteristics. In strict districts, $r_j = 1$ if $F(x_j) + G(s_j) > 0$ because they rely on both observed and unobserved characteristics.

I train a model on strict districts to estimate $E[y_j|x_j, r_j = 1] = f(x_j) + E[g(s_j)|r_j = 1]$. A prediction trained on strict districts would incorporate any positive selection generated by these districts that select on unobserved characteristics. In comparison, the actual performance in lenient districts provides an estimate of $E[y_j|x_j] = f(x_j)$ because these districts ignore unobserved characteristics.[16] By comparing predicted performance ($E[y_j|x_j, r_j = 1]$) to actual performance ($E[y_j|x_j]$) in the lenient districts, I test whether $E[g(s_j)|r_j = 1] = 0$.

In practice, I cannot determine which districts rely on unobserved characteristics. However, I can observe retention rates conditional on summative ratings. Consequently, I partition districts into strict and lenient halves using the following regression:

$$r_{jt} = \beta x_{jt} + \delta_t + \varepsilon_{jt}. \tag{4}$$

I regress the retention of teacher j after year t (r_{jt}) on summative ratings (x_{jt}) and year fixed effects (δ_t). To measure district strictness, I calculate the mean residual (ε_{jt}) for all teachers in the district other than teacher j. This leave-one-out mean avoids biasing a district's strictness by using the teacher's own retention decision. Positive residuals suggest teachers were retained more often than expected, while negative residuals suggest teachers were retained less often than expected. Figure 3 plots the positive relationship between leave-one-out mean district strictness and teacher retention. Thus, I partition the leave-one-out mean residuals into strict (below median) and lenient (above median) halves.[17] Since strict districts retain fewer teachers conditional on ratings, I assume that these districts rely on both summative ratings (observed) and unobserved characteristics, while lenient districts only rely on summative ratings. This partition also relies on a monotonicity assumption. I assume that any teacher retained in a strict district also would have been retained in a lenient district.

Ideally, I would focus on involuntary dismissals to identify strict and lenient districts.

[16] For lenient districts, $E[g(s_j)|r_j = 1] = E[g(s_j)|F(x_j) > 0] = E[g(s_j)] = 0$.

[17] The results are similar when using different deciles to partition the sample into strict and lenient districts (not shown).

Unfortunately, I cannot distinguish between voluntary and involuntary turnover.[18] Therefore, I may misclassify lenient districts as strict districts if they have high rates of voluntary teacher attrition. I would expect this problem to be especially prevalent in hard-to-staff districts that have difficulty filling vacancies and retaining teachers. These hard-to-staff districts tend to have high poverty rates and low proficiency rates. To evaluate this concern, I compare the characteristics of strict and lenient districts in Table 1. Relative to strict districts (third column), lenient districts (first column) have higher poverty (FRPL) rates, more ELL students, lower proficiency rates, and more minority students. All these differences are statistically significant at the 1% level as seen in the fifth column. This suggests that lenient districts actually have the characteristics of hard-to-staff districts. With high voluntary attrition rates but low turnover rates conditional on summative ratings, the lenient districts would have few opportunities to select on unobserved characteristics as they attempt to retain as many teachers as possible. As a result, voluntary attrition is unlikely to cause the misclassification of strict and lenient districts in equation (4).

After dividing the sample, I train a random forest on the strict districts. I use the first three years of math value-added to predict subsequent math value-added and repeat the process for ELA value-added and summative ratings.

Figure 4 plots the relationship between predicted and actual performance using algorithms trained on strict districts and applied to lenient districts. I include a 45-degree line and calculate the average difference between true and predicted outcomes. The average differences are statistically indistinguishable from 0 ranging from -0.023 to 0.019 standard deviations or summative rating points. Thus, I fail to reject the null hypothesis that $E[g(s_j)|r_j = 1] = 0$, so equation (3) holds. I do not find any evidence that districts use unobserved characteristics that impact subsequent performance to selectively retain teachers.

[18] In fact, it is very difficult to identify voluntary and involuntary turnover in any dataset. For example, some teachers may appear to voluntarily leave the district if they knew that they would soon be dismissed.

4 Additional Performance Measures

I use random forests to predict subsequent teacher value-added and summative ratings using previous performance. This exercise evaluates whether schools are optimally using all available data to inform personnel decisions. Specifically, I test for the existence of Pareto improvements where districts improve average performance along at least one dimension without decreasing average performance along the other dimensions. Pareto optimality is relevant in this environment because previous research has linked value-added to long-run student success (Chetty et al., 2014), though these tests fail to capture the development of non-cognitive skills that improve student outcomes (Jackson, 2018). Summative ratings may capture these critical components of student success. In addition, the education community often feels uncomfortable relying on these test-based metrics ("Value-added measures in teacher evaluation", 2019; "Taking action on the promise of the Every Student Succeeds Act", 2016), so they rely on summative ratings as a more comprehensive measure of teacher performance. In fact, every state uses evaluations to assess their teachers (Ross & Walsh, 2019). Since the trade-offs between value-added and summative ratings are not well understood, any revised ranking system must generate a Pareto improvement to demonstrate a definitive increase in average teacher quality.

To conduct the analysis, I split the sample into three parts. First, I use a random forest on 40% of the sample to impute missing data.[19] Second, I train the algorithm with a distinct 40% of the sample. Third, I use the remaining 20% of the sample to test the performance of the models. This holdout sample allows me to evaluate the efficacy of the algorithms.

To train the algorithms, I estimate several models using different combinations of previous performance to predict each outcome. First, I use previous performance to predict subsequent performance along the same metric. For example, I use previous math value-added to predict subsequent math value-added. I then repeat the process using ELA value-added

[19] To impute subsequent performance, I use previous performance along the same metric. For example, to impute math value-added after year 3, I use the first 3 years of math value-added as the predictors.

and summative ratings. Second, I use the composite measure described in Section 3.1 that is a weighted average of summative ratings and non-missing value-added. The composite measures are used as both the outcome and predictor in these models. Each of these models relies on three years of pretenure data to predict subsequent performance.[20]

Next, I compare the predicted performance from the algorithm to the actual performance in the holdout sample. Panels A, C, and E of Figure 5 plot actual performance against predicted performance. The algorithms accurately predict performance with mean squared errors of less than 0.03 standard deviations or summative rating points.

In addition, I compare retention rates to predicted performance in the holdout sample. Panels B, D, and F of Figure 5 plot retention rates by predicted performance. In Panel F, districts effectively remove teachers predicted to earn low summative ratings. However, Panels B and D show districts fail to remove teachers predicted to generate low value-added. These findings suggest revised rankings are more likely to improve value-added than ratings.

I then evaluate the algorithms by simulating personnel decisions. Specifically, I rank teachers based on their predicted performance generated from each algorithm. I simulate personnel decisions by removing the bottom $p \in \{1, 2, \ldots, 70\}$ percentile of teachers using each ranking system.[21] My main specification dismisses 10% of teachers because annual turnover rates for New Jersey teachers in their first three years of teaching are about 13%. Since some of these teachers voluntarily quit, 10% is a reasonable dismissal rate. As a comparison algorithm, I rank teachers based on their mean pretenure summative ratings.

Panel A of Figure 6 simulates average performance when using mean summative ratings to rank the teachers. The y-axis records the average performance of retained teachers, while the x-axis defines the percentile of teachers dismissed. For example, when $x = 10$, I dismiss the bottom 10% of teachers based on mean pretenure ratings. While this method accurately ranks teachers by subsequent summative ratings (dashed and dotted blue), it fails to generate any value-added gains (solid black and dashed red). The first row of Table 2

[20] I use a three-year pretenure period because 32 states use this length (Thomsen, 2020).
[21] I stop at the 70^{th} percentile because the samples become small.

shows summative ratings rise by 0.0313 points when dismissing the bottom 10% of teachers based on pretenure ratings relative to dismissing no teachers.[22] However, subsequent value-added remains unchanged by this dismissal policy. Ideally, schools would continue to generate these summative rating gains, while also improving subsequent value-added. Thus, I turn to machine learning algorithms.

First, I estimate random forest models using only one measure of performance. I use previous math value-added to predict subsequent math value-added and repeat the process for ELA value-added and summative ratings. Panels B–D of Figure 6 show the results.[23] I demonstrate improvements along the outcome measure but cannot increase value-added and ratings simultaneously. For example, Panel B relying on math value-added as the outcome generates strong gains in math value-added as dismissal rates increase with little change in ratings. Similarly, the Panel D shows ratings rise as dismissal rates increase when using an algorithm relying on ratings. However, value-added remains unchanged.

To compare the ranking systems, the remaining rows of Table 2 use a 10% dismissal rate and estimate the change in average retained teacher performance using the machine learning algorithms relative to the current mean summative rating dismissal policy. For example, dismissing the bottom 10% of teachers using algorithms relying on value-added increases subsequent value-added by 0.0238–0.0334 standard deviations, as seen in the second and third rows of Table 2. However, this policy causes summative ratings to decline by 0.0192–0.0301 points relative to the current system. Similarly, the third row using ratings generates no gains along any dimension relative to the current system.

While machine learning algorithms effectively predict the outcome variable, they poorly predict the other measures. As discussed in Section 3.1, I encounter a multidimensionality problem because random forests only permit one outcome variable and value-added is weakly

[22] 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.

[23] The graphs that rely on math (ELA) value-added as predictors or outcomes generate noisy estimates for ELA (math) value-added due to limited samples of elementary school teachers who teach both subjects.

correlated with ratings. Using a composite measure that is a weighted average of value-added and ratings, I reduce this multidimensional problem into a single dimension.

In the final five entries of Table 2, I use the composite measure and find that placing more weight on summative ratings (moving down the table) increases subsequent ratings but decreases subsequent value-added. Yet, these algorithms can simultaneously increase value-added and ratings in Figure 7. The second and third to last rows of Table 2 using a composite measure with 50%–70% weight on ratings increases subsequent value-added by 0.0099–0.0143 standard deviations without negatively impacting subsequent ratings.[24] Almost all these Pareto improvements are statistically significant at the 10% level. Although none of the algorithms improve summative ratings relative to the current system, I generate a Pareto improvement by increasing subsequent teacher value-added.[25]

These results show districts can generate Pareto improvements in average teaching performance by using a weighted average of value-added and summative ratings. As seen in brackets, the value-added gains are about 0.04 teacher standard deviations of math and ELA value-added. Using partial equilibrium estimates from Chetty et al. (2014), this equates to a present value gain of $2,240 per student.[26] This value is over 9 times larger than the productivity effects of tenure (Ng, 2021). In addition, this method is quite inexpensive to implement because all these data are already available to the school districts. Compared to previous research, these estimated gains lie in the upper tail of the confidence interval from Chalfin et al. (2016). I further contribute to the literature by demonstrating that other dimensions of performance need not decline to generate these gains.

[24] Pareto improvements are feasible because the current system fails to perfectly sort teachers by subsequent ratings and ignores value-added. As a result, machine learning algorithms can generate Pareto improvements by more effectively ranking similarly rated teachers by subsequent value-added.

[25] The point estimates are similar when using all teachers rather than just novices in Table A4. To maintain policy relevance, I continue to restrict the sample to novices because dismissing pretenured teachers is much more feasible than dismissing tenured teachers. Performance-related dismissal rates are about 19 times higher for non-tenured teachers than tenured teachers (National Center for Education Statistics, 2012).

[26] Chetty et al. (2014) estimates a 1 standard deviation increase in value-added for 1 grade generates a present value gain of $7,000 per student. I scale this estimate by the 0.04 teacher standard deviation gain and the 8 grades for which I can calculate value-added.

4.1 Information or Machine Learning Techniques?

Policymakers may be concerned about the "black box" nature of machine learning techniques. While random forests do not require functional form assumptions and flexibly account for non-linear relationships, the underlying coefficients are difficult to interpret. To provide a more transparent alternative, I impute and train the data using OLS. I continue to use the same imputing, training, and holdout samples.

To impute missing data, I regress subsequent math value-added, ELA value-added, and ratings on their analogous annual components in the imputing sample. Using these linear projections, I impute missing observations in the training and holdout samples. Then, I estimate OLS regressions in the training sample using the models described in Section 3.1.

Using OLS, Table 3 shows the baseline results (top row) and changes relative to the current system (remaining rows). The OLS estimates are nearly identical to the random forests results in Table 2.[27]

The estimates remain similar because the relationship between predictors and outcomes is linear. Figure A1 plots this linear relationship between year 3 and subsequent performance.[28] Both random forests and OLS account for linear relationships between predictors and outcomes, so they both perform equally well in this context. Random forests are more useful when incorporating additional data. For example, individual rating components capturing specific domains, such as classroom management or lesson planning, may have non-linear relationships. In this case, flexible machine learning algorithms could produce sizeable gains. Given the flexibility of machine learning techniques, I continue to use them for the remainder of the paper, though the analysis remains similar when using OLS.

[27] The largest difference occurs when comparing math value-added changes based on ELA value-added models. These analyses rely on small samples because few teachers have both math and ELA value-added.

[28] The relationship also is linear when comparing performance in other years.

4.2 Changing Demographics in Response to Reformed Algorithms

While I find that revised ranking systems can improve value-added without harming summative ratings, it also is critical to consider the impacts of these revised decisions on diversity. In New Jersey, male and non-white (Black or Hispanic) teachers are underrepresented in the profession relative to their corresponding student demographics. Table A1 shows that only 18.3% of teachers are male, while 51.6% of students are male. Similarly, non-white teachers comprise only 13.6% of the teacher labor force, while 40.2% of students are non-white.[29]

This underrepresentation may have negative impacts on in-group students. In fact, 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. With already limited access to in-group educators, male and non-white students may benefit from revised personnel decisions that reduce turnover among these teachers.

To evaluate impacts on diversity, Figure 8 plots changes in demographics associated with each algorithm. In Panels A and D using the current ranking system and the algorithm trained using summative ratings, the fraction of male (dashed and dotted blue) and non-white (solid black) teachers steadily declines as dismissal rates increase. This occurs because male and non-white teachers earn lower summative ratings. Figure 9 shows the cumulative distribution functions of the performance measures by teacher gender and race. In Panels E and F, the distribution of summative ratings for female (dashed red) and white (dashed red) teachers almost universally exceeds the distribution for male (dashed and dotted blue) and non-white (solid black) teachers, respectively. Similarly, Table 4 shows summary statistics of teacher performance by gender and race. In the final column, I find mean summative

[29] These gender and racial disparities are prevalent throughout the United States ("Characteristics of Public School Teachers", 2020; "Racial/Ethnic Enrollment in Public Schools", 2020).

ratings are 0.111 and 0.126 points higher for female and white teachers, respectively. Since male and non-white teachers consistently earn lower ratings, algorithms trained to maximize ratings simultaneously reduce these teachers' representation in the profession.

Despite earning lower ratings, male and non-white teachers do not generate less value-added than their counterparts in Panels A–D of Figure 9. In fact, Panel B of Table 4 shows that non-white teachers' average value-added is 0.114-0.134 standard deviations higher. Thus, the fraction of male and non-white teachers stays constant or rises as dismissal rates increase for value-added algorithms in Panels B and C of Figure 8. Table 5 uses 10% dismissal rates to show baseline demographics (top row) and changes relative to the current system (remaining rows). In the top row, the current system reduces the male teacher composition by 2.1 percentage points relative to no dismissals. This difference is statistically significant at the 5% level. I also find a negative point estimate for non-white teachers but it is statistically indistinguishable from 0. The second and third rows of Table 5 show the algorithms that incorporate math value-added generate a statistically significant 4.21 percentage point increase in the fraction of male teachers relative to the current system. While the other male and non-white teacher estimates remain statistically indistinguishable from 0, the point estimates that use value-added are positive.

Focusing on the preferred specification using the composite measures with 50%–70% weight on ratings, I often find positive point estimates but no statistically significant differences relative to the current method in Table 5. In Panel E of Figure 8, the composite measure shows little change in diversity as dismissal rates increase. While male and non-white teachers earn lower ratings, their similar or higher value-added stabilizes diversity in the composite algorithms. Although these algorithms do not generate statistically significant increases in diversity, they certainly do not harm it relative to the current system.

5 Pretenure Period Length

In this section, I quantify the returns to longer pretenure periods by reestimating each model described in Section 3.1 using 1, 2, or 3 years of pretenure data.[30] While longer pretenure periods will improve the precision of estimated performance and almost always increase average retained teacher quality, I also must consider the corresponding reduced compensating differentials associated with weakened job security. In fact, Johnston (2018) finds teachers equate each additional pretenure year to a $415 reduction in salary. To overcome these costs, any improvement must be relatively large in magnitude and statistically significant.

Table 6 estimates the improvement in average teacher quality generated by extending the pretenure period from 1 to 3 years with a 10% dismissal rate. While I find positive point estimates, the gains are inconsistent with only one statistically significant value.[31] When using summative ratings in the first and fourth entries, value-added remains unchanged, while summative ratings increase by 0.0083–0.0115 points. I find analogous results for value-added in the second and third entries. Using the composite measure with 50-70% weight on summative ratings (the preferred models from Section 4), value-added increases by 0.0018–0.0132 standard deviations, while summative ratings increase by 0.0145–0.0148 points. However, these gains are not statistically significant. Thus, extended pretenure periods are negatively impacting selection into teaching without providing much additional information.[32]

Extending the pretenure period provides little additional information at low dismissal rates because principals can accurately identify low-performing teachers with limited data (Harris & Sass, 2014). In fact, extended pretenure periods only produce strong, statistically significant effects when dismissal rates are higher. Figure 10 shows the improvements in average teacher performance generated when extending the pretenure period from 1 to 3

[30] I do not include estimates using 4 or 5 pretenure years because I have too few observations.

[31] The gains are even weaker when extending the pretenure period from 1 to 2 years (Table A5) or from 2 to 3 years (Table A6).

[32] The estimates in Table 6 are similar in magnitude to those from using additional performance metrics in Section 4. However, unlike the extended pretenure period estimates, the revised algorithms from Section 4 generate consistent, statistically significant gains that are virtually costless to implement.

years. As dismissal rates rise for the composite measure in Panel E, the performance gains also increase for all three metrics.[33] Using a 50% dismissal rate, Table 7 shows large, statistically significant gains to extended pretenure periods. For example, the algorithm using a composite measure with 50% weight on ratings increases average retained teacher math value-added, ELA value-added, and ratings by 0.0539 standard deviations, 0.0265 standard deviations, and 0.0775 points, respectively. The math value-added and summative rating estimates are statistically significant at the 5% level. These returns are up to 0.07 standard deviations or points larger than the analogous point estimates from a 10% dismissal rate in Table 6.[34]

To depict a potential mechanism, I plot the kernel density of teacher performance in Figure 11. In Panels A and B, I graph the distribution of career math value-added as a proxy for true ability. The vertical line in Panel A shows the 10^{th} percentile, which illustrates a 10% dismissal rate. The red distributions represent noisy annual performance measures of a teacher whose true ability is 0.2 standard deviations below the 10^{th} percentile. This teacher should be dismissed but may be misclassified out of the bottom decile and retained due to this noise.[35] Using three years of data reduces the noise of the estimates and tightens the distribution. As a result, I shade the difference between the dashed and dotted lines, which represents the number of bottom decile teachers misclassified using one year of data who would be correctly classified using three years of data. The size of the distribution is scaled to the density of teachers at that point in the overall distribution of teacher quality. Since there are few teachers in this portion of the distribution, the gains from extending the

[33] Panels A–D of Figure 10 produce a similar pattern of results using the other algorithms.

[34] I find stronger gains in Table A7 when extending the pretenure period from 1 to 2 years than those generated when extending the pretenure period from 2 to 3 years in Table A8. For example, the composite measure specification with 50% weight on summative ratings shows that the extension from 1 to 2 years (Table A7) increases math value-added, ELA value-added, and summative ratings by 0.0434 standard deviations, 0.0244 standard deviations, and 0.0424 points, respectively. In comparison, the extension from year 2 to 3 (Table A8) shows math value-added, ELA value-added, and summative ratings changed by 0.0196 standard deviations, -0.0043 standard deviations, and 0.0298 points, respectively.

[35] To proxy for the variance of the distribution using 1 year of data (dashed lines) or 3 years of data (dotted lines), I calculate the mean squared errors relative to the career performance of teachers within 0.1 standard deviations of the mean.

pretenure period are very small. In comparison, Panel B conducts a similar analysis using 50% dismissal rates and generates much larger gains (shown in blue). The results are similar for ELA value-added and summative ratings.

In other words, more teachers are clustered near the middle of the distribution than in the tails. The bottom decile of teachers has math value-added that spans about 1 standard deviation,[36] while the 40^{th}–50^{th} percentiles span only 0.04 standard deviations. With similar annual performance noise throughout the distribution, it is much harder to classify teachers near the 50^{th} percentile than near the 10^{th} percentile. Thus, extended pretenure periods only produce meaningful gains when dismissal rates are closer to 50%.

These results align with Rothstein (2015) who used a structural model of teacher contracts. In several of his parameterizations, his graphs show that extending the pretenure period has little effect when dismissal rates are low. The gains accumulate and are largest when dismissal rates approach 40%.

From a policy perspective, these results would recommend shortening the pretenure period to only one year or increasing dismissal rates. The current three- to four-year pretenure period reduces compensating differentials relative to shorter pretenure periods without offering much additional useful information. Although schools must increase dismissal rates to use additional years of pretenure performance effectively, my simulation does not account 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), as well as changes in selection into teaching. Rothstein (2015) models selection into teaching and finds optimal dismissal rates vary between 10% and 71% based on the model's parameterization. Although I am unable to identify optimal dismissal rates, I account for selection by holding dismissal rates fixed when estimating improvements from extended pretenure periods. Consequently, this analysis informs optimal pretenure length conditional on dismissal rates. Specifically, a

[36] I truncated the tails of the graph to enlarge the image.

long pretenure period with low dismissal rates is a suboptimal combination of policies.[37]

6 Detecting Bias in the Data

In Section 4.2, I find that male and non-white teachers earn lower ratings despite having similar value-added. These rating disparities also appeared in previous research (Bailey et al., 2016; Drake et al., 2019; Sartain & Steinberg, 2020; Ng, 2021; Chi, 2021). In this section, I provide several tests to detect the presence of gender and racial biases in the results.

First, I evaluate biases by including teacher gender and race in the training algorithms. In Table A9, I estimate the improvements in average teacher performance when using a 10% dismissal rate, three years of performance data, and demographics.[38] With similar results to Table 2, Table A9 shows demographic data do not improve the algorithm's prediction accuracy. For example, the composite measure with 50% weight on summative ratings shows ratings do not change, while value-added increases by 0.0149–0.0162 standard deviations relative to the current system. This is reassuring because demographic data should be orthogonal to average performance and have no effect on the algorithms.

Next, I test for differences in residuals generated by the machine learning algorithms in Table A10. Specifically, I subtract the predicted performance from the actual performance for each group separately. Then, I calculate the difference between the residuals across groups. In the female-male (white-non-white) comparison, a negative value suggests that the algorithm underpredicts male (non-white) teacher performance relative to female (white) teacher performance. Although some of the point estimates are non-negligible, I do not identify any statistically significant differences or clear pattern of results across the performance measures. For example, the math value-added algorithm underpredicts male and non-white performance by 0.0089 and 0.0281 standard deviations, respectively. However,

[37] There also are practical limitations to a one-year pretenure period because teachers receive summative ratings at the end of the year and districts wish to use these ratings as opportunities for growth. However, a two-year pretenure period remains feasible.

[38] To maintain consistency with the results from Section 4, I continue to impute the data only using performance measures.

the ELA value-added algorithm overpredicts male and non-white performance by 0.0263 and 0.0270 standard deviations, respectively.[39]

Although I find summative rating disparities by gender and race, these tests do not provide any evidence that discrimination is biasing the estimated results. However, some of the tests are underpowered and rely on prior summative rating data, which may be inherently biased. Ideally, I would predict summative ratings using a more objective measure, such as value-added. Unfortunately, the correlation between value-added and summative ratings is too weak for one to predict the other. While I do not find any evidence that discrimination is biasing the machine learning algorithms, I cannot completely eliminate this possibility.

7 Conclusion

Schools are not optimally using all the data available to them. Utilizing predictive models, schools can increase subsequent average value-added by 0.01 standard deviations without decreasing ratings or increasing dismissal rates. These Pareto improvements are virtually costless to implement, as all the data are readily available. The gains are a product of using additional information rather than sophisticated methods, as the estimates are similar when using simple OLS or sophisticated machine learning techniques. The algorithms also do not harm diversity relative to the current system. Future research could supplement this paper by estimating the impact of summative ratings on long-run student outcomes, similar to Chetty et al. (2014) for value-added. Using these parameters, districts could identify optimal weights for value-added and ratings.

I also find that longer pretenure periods do not improve average teacher quality unless accompanied by higher dismissal rates. Schools can accurately classify bottom decile teachers after only one year of teaching. Thus, extra years of data provide little additional information, while also reducing compensating differentials. From a policy perspective, this

[39] The presence of heterogeneity also would suggest that discrimination may be impacting the algorithms. However, I find no heterogeneity by teacher, school, and student characteristics (not shown).

finding recommends either low dismissal rates with a one-year pretenure period or high dismissal rates with a longer pretenure period. Future research that estimates selection effects associated with longer pretenure periods and higher dismissal rates could supplement this analysis.

Overall, I find that incorporating additional measures of teacher performance is a more effective technique to select high-quality teachers than extending the pretenure period given current dismissal rates. Unlike extending the pretenure period with a 10% dismissal rate, using additional performance measures generates statistically significant improvements in average retained teacher performance without negatively impacting selection into teaching.

References

Abaluck, J., Agha, L., Kabrhel, C., Raja, A., & Venkatesh, A. (2016). The determinants of productivity in medical testing: Intensity and allocation of care. *American Economic Review*, *106*(12), 3730–64.

Athey, S., Katz, L., Krueger, A., Levitt, S., & Poterba, J. (2007). What does performance in graduate school predict? Graduate economics education and student outcomes. *American Economic Review*, *97*(2), 512–520.

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*.

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*.

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

Chalfin, A., Danieli, O., Hillis, A., Jelveh, Z., Luca, M., Ludwig, J., & Mullainathan, S. (2016). Productivity and selection of human capital with machine learning. *American Economic Review*, *106*(5), 124–27.

Chandler, D., Levitt, S., & List, J. (2011). Predicting and preventing shootings among at-risk youth. *American Economic Review*, *101*(3), 288–92.

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

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*.

Chingos, M., & Peterson, P. (2011). It's easier to pick a good teacher than to train one: Familiar and new results on the correlates of teacher effectiveness. *Economics of Education Review*, *30*(3), 449–465.

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.

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* (Tech. Rep.). National Bureau of Economic Research.

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.

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

Harris, D., & Sass, T. (2014). Skills, productivity and the evaluation of teacher performance.

Economics of Education Review, *40*, 183–204.

Jackson, C. (2018). What do test scores miss? The importance of teacher effects on non–test score outcomes. *Journal of Political Economy*, *126*(5), 2072–2107.

Jacob, B., & Lefgren, L. (2008). Can principals identify effective teachers? Evidence on subjective performance evaluation in education. *Journal of Labor Economics*, *26*(1), 101–136.

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

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2017). Human decisions and machine predictions. *The Quarterly Journal of Economics*, *133*(1), 237–293.

Kleinberg, J., Ludwig, J., Mullainathan, S., & Obermeyer, Z. (2015). Prediction policy problems. *American Economic Review*, *105*(5), 491–95.

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.

Mullainathan, S., & Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, *31*(2), 87–106.

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`.

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

Ng, K. (2021). *The effects of teacher tenure on productivity and selection* (Unpublished doctoral dissertation). Cornell University.

Racial/Ethnic Enrollment in Public Schools. (2020). *National Center for Education Statis-*

tics. Retrieved from https://nces.ed.gov/programs/coe/indicator_cge.asp.

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

Ross, E., & Walsh, K. (2019). State of the States: Teacher and Principal Evaluation Policy. *Washington DC: National Council on Teacher Quality*.

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

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.

Taking action on the promise of the Every Student Succeeds Act. (2016). *American Federation of Teachers*. Retrieved from https://www.aft.org/resolution/taking-action-promise-every-student-succeeds-act.

Thomsen, J. (2020). State legislation: Teaching quality - Tenure or continuing contract. *Education Commission of the States*.

Value-added measures in teacher evaluation. (2019). *National Association of Secondary School Principals*. Retrieved from https://www.nassp.org/top-issues-in-education/position-statements/value-added-measures-in-teacher-evaluation/.

Winters, M., & Cowen, J. (2013). Would a value-added system of retention improve the distribution of teacher quality? A simulation of alternative policies. *Journal of Policy*

Analysis and Management, *32*(3), 634–654.

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

Tables

Table 1: Summary Statistics by District Strictness

	Lenient		Strict		
	Mean	SD	Mean	SD	Diff
FRPL	0.453	0.296	0.296	0.255	0.157***
ELL	0.079	0.079	0.046	0.064	0.033***
Math Proficient	0.476	0.164	0.564	0.160	-0.088***
ELA Proficient	0.521	0.175	0.621	0.156	-0.101***
Black	0.197	0.189	0.161	0.208	0.036***
Hispanic	0.339	0.283	0.218	0.212	0.121***
Observations	3,874		3,874		

Notes: This table provides summary statistics for strict and lenient districts calculated at the district level. I define lenient (strict) districts as those with above (below) median leave-one-out average residuals from equation (4). The row headers define the variable. The first two columns provide statistics for lenient districts. The first column records the means of the variables, while the second column measures their standard deviations. The next two columns are defined similarly for strict districts. The final column calculates the difference in means and provides the significance level from a T-test of equality.

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

Table 2: Difference in Performance when Dismissing 10% of Teachers

	Math VA	N	ELA VA	N	Ratings	N
Mean Ratings (Baseline)	0.0008	212	0.0005	232	0.0313***	390
	(0.0035)		(0.0043)		(0.0042)	
	[0.0026]		[0.0016]		[0.0971]	
Math using Math	0.0238***	212	0.0334***	61	-0.0301***	212
	(0.0068)		(0.0127)		(0.0097)	
	[0.0770]		[0.1166]		[-0.0934]	
ELA using ELA	0.0252***	61	0.0246*	232	-0.0192*	232
	(0.0067)		(0.0126)		(0.0099)	
	[0.0818]		[0.0858]		[-0.0595]	
Ratings using Ratings	-0.0011	212	-0.0001	232	-0.0017	390
	(0.0155)		(0.0061)		(0.0073)	
	[-0.0035]		[-0.0004]		[-0.0052]	
Composite using						
10% Ratings	0.0270*	212	0.0219***	232	-0.0201***	383
	(0.0162)		(0.0062)		(0.0072)	
	[0.0876]		[0.0764]		[-0.0624]	
30% Ratings	0.0240***	212	0.0201***	232	-0.0181***	383
	(0.0051)		(0.0042)		(0.0065)	
	[0.0779]		[0.0701]		[-0.0563]	
50% Ratings	0.0130*	212	0.0143**	232	-0.0047	383
	(0.0071)		(0.0064)		(0.0064)	
	[0.0420]		[0.0499]		[-0.0146]	
70% Ratings	0.0099	212	0.0115*	232	-0.0018	383
	(0.0067)		(0.0064)		(0.0059)	
	[0.0322]		[0.0402]		[-0.0055]	
90% Ratings	-0.0007	212	0.0003	232	0.0009	383
	(0.0061)		(0.0055)		(0.0054)	
	[-0.0023]		[0.0011]		[0.0030]	

Notes: This table estimates the change in performance generated when dismissing the bottom 10% of teachers using three years of data. These models use random forest algorithms defined in Section 3.1. The row headers define the algorithm's outcome and predictors. The first row shows the change in performance generated when dismissing the bottom 10% of teachers using mean summative ratings relative to no dismissals. The remaining rows record changes relative to the first row using the algorithms defined in the row header. The first two columns show the change in math value-added and number of holdout observations. The remaining columns are defined similarly for ELA value-added and summative ratings.

Standard errors generated using 1,000 bootstrapped samples in parentheses. 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: Difference in Performance using OLS

	Math VA	N	ELA VA	N	Ratings	N
Mean Ratings (Baseline)	-0.0010	212	0.0019	232	0.0337***	390
	(0.0034)		(0.0041)		(0.0041)	
	[-0.0033]		[0.0068]		[0.1046]	
Math using Math	0.0289***	212	0.0296**	61	-0.0263***	212
	(0.0059)		(0.0123)		(0.0092)	
	[0.0935]		[0.1033]		[-0.0816]	
ELA using ELA	0.0454***	61	0.0231*	232	-0.0178*	232
	(0.0059)		(0.0126)		(0.0093)	
	[0.1473]		[0.0807]		[-0.0552]	
Ratings using Ratings	0.0027	212	-0.0012	232	0.0037	390
	(0.0160)		(0.0060)		(0.0067)	
	[0.0086]		[-0.0041]		[0.0116]	
Composite using						
10% Ratings	0.0261	212	0.0247***	232	-0.0227***	383
	(0.0165)		(0.0060)		(0.0067)	
	[0.0847]		[0.0861]		[-0.0706]	
30% Ratings	0.0299***	212	0.0232***	232	-0.0129*	383
	(0.0051)		(0.0087)		(0.0070)	
	[0.0970]		[0.0809]		[-0.0401]	
50% Ratings	0.0198***	212	0.0163***	232	-0.0074	383
	(0.0071)		(0.0063)		(0.0060)	
	[0.0641]		[0.0569]		[-0.0229]	
70% Ratings	0.0150**	212	0.0101*	232	0.0008	383
	(0.0068)		(0.0060)		(0.0055)	
	[0.0488]		[0.0352]		[0.0026]	
90% Ratings	0.0045	212	-0.0005	232	0.0024	383
	(0.0054)		(0.0052)		(0.0049)	
	[0.0147]		[-0.0016]		[0.0074]	

Notes: This table estimates the change in performance generated when dismissing the bottom 10% of teachers using three years of data. These models use linear regressions. The row headers define the algorithm's outcome and predictors. The first row shows the change in performance generated when dismissing the bottom 10% of teachers using mean summative ratings relative to no dismissals. The remaining rows record changes relative to the first row using the algorithms defined in the row header. The first two columns show the change in math value-added and number of holdout observations. The remaining columns are defined similarly for ELA value-added and summative ratings.

Standard errors generated using 1,000 bootstrapped samples in parentheses. 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: Summary Statistics by Gender and Race

Panel A: Gender

	Male		Female		
	Mean	SD	Mean	SD	Diff
Math VA	-0.009	0.302	-0.024	0.281	0.015
ELA VA	-0.012	0.270	-0.019	0.284	0.007
Ratings	3.128	0.357	3.239	0.304	-0.111***
Observations	1,770		7,949		

Panel B: Race

	Non-white		White		
	Mean	SD	Mean	SD	Diff
Math VA	0.096	0.286	-0.038	0.281	0.134***
ELA VA	0.080	0.301	-0.034	0.275	0.114***
Ratings	3.110	0.400	3.236	0.298	-0.126***
Observations	1,334		8,385		

Notes: This table records mean performance by gender (Panel A) and race (Panel B). The row headers define the performance variable. The first two columns provide statistics for male and non-white teachers. The first column provides the mean of the variable, while the second column measures its standard deviation. The third and fourth columns are defined similarly for female and white teachers. The final column calculates the difference in means and provides the significance level from a T-test of equality for the coefficients.

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

Table 5: Demographic Changes when Dismissing 10% of Teachers

	Male	N	Non-white	N
Mean Ratings (Baseline)	-0.0210**	390	-0.0052	390
	(0.0082)		(0.0062)	
Math using Math	0.0421***	212	0.0158	212
	(0.0149)		(0.0105)	
ELA using ELA	0.0192	232	0.0000	232
	(0.0146)		(0.0105)	
Ratings using Ratings	0.0124	390	-0.0013	390
	(0.0149)		(0.0111)	
Composite using				
10% Ratings	0.0248*	383	0.0106	383
	(0.0143)		(0.0113)	
30% Ratings	0.0218**	383	0.0135	383
	(0.0110)		(0.0092)	
50% Ratings	0.0073	383	0.0019	383
	(0.0111)		(0.0085)	
70% Ratings	0.0073	383	-0.0010	383
	(0.0104)		(0.0083)	
90% Ratings	0.0073	383	0.0019	383
	(0.0102)		(0.0081)	

Notes: This table shows the change in demographics generated when dismissing the bottom 10% of teachers using three years of data. These models use the random forest algorithms defined in Section 3.1. The first row shows the change generated when dismissing the bottom 10% of teachers using mean summative ratings relative to no dismissals. The remaining rows record changes relative to the first row using the algorithms defined in the row header. The first column shows the change in the fraction of male teachers, while the second column records the number of holdout observations. The remaining columns are defined similarly for the fraction of non-white teachers.

Standard errors generated using 1,000 bootstrapped samples in parentheses.

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

Table 6: Gains from Extending Pretenure from 1 to 3 Years: 10% Dismissal Rate

	Math VA	N	ELA VA	N	Ratings	N
Mean Ratings	-0.0022	212	-0.0048	232	0.0083	390
	(0.0042)		(0.0048)		(0.0053)	
	[-0.0070]		[-0.0167]		[0.0258]	
Math using Math	0.0123	212	0.0176	61	-0.0040	212
	(0.0117)		(0.0392)		(0.0178)	
	[0.0400]		[0.0614]		[-0.0124]	
ELA using ELA	0.0222*	61	0.0203	232	0.0102	232
	(0.0123)		(0.0391)		(0.0173)	
	[0.0721]		[0.0709]		[0.0317]	
Ratings using Ratings	-0.0031	212	0.0014	232	0.0115	390
	(0.0321)		(0.0139)		(0.0156)	
	[-0.0101]		[0.0048]		[0.0357]	
Composite using						
10% Ratings	0.0247	212	0.0100	232	0.0105	383
	(0.0311)		(0.0132)		(0.0154)	
	[0.0801]		[0.0349]		[0.0328]	
30% Ratings	0.0187	212	0.0091	232	0.0068	383
	(0.0114)		(0.0366)		(0.0164)	
	[0.0608]		[0.0318]		[0.0213]	
50% Ratings	0.0018	212	0.0083	232	0.0148	383
	(0.0125)		(0.0141)		(0.0101)	
	[0.0059]		[0.0290]		[0.0459]	
70% Ratings	0.0132	212	0.0062	232	0.0145	383
	(0.0119)		(0.0142)		(0.0098)	
	[0.0429]		[0.0217]		[0.0449]	
90% Ratings	-0.0026	212	-0.0001	232	0.0192**	383
	(0.0119)		(0.0134)		(0.0097)	
	[-0.0083]		[-0.0003]		[0.0598]	

Notes: This table shows the change in performance generated when extending the pretenure period from 1 to 3 years and dismissing the bottom 10% of teachers. I use the random forest algorithms defined in Section 3.1. The row headers define the outcome and predictors. The first two columns show the change in math value-added and number of holdout observations. The remaining columns are defined similarly for ELA value-added and summative ratings.

Standard errors generated using 1,000 bootstrapped samples in parentheses. 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: Gains from Extending Pretenure from 1 to 3 Years: 50% Dismissal Rate

	Math VA	N	ELA VA	N	Ratings	N
Mean Ratings	0.0129	212	-0.0068	232	0.0512***	390
	(0.0139)		(0.0159)		(0.0137)	
	[0.0417]		[-0.0236]		[0.1590]	
Math using Math	0.0547**	212	-0.0010	61	-0.0233	212
	(0.0221)		(0.0516)		(0.0287)	
	[0.1775]		[-0.0036]		[-0.0723]	
ELA using ELA	0.0596***	61	0.0820	232	0.0453	232
	(0.0222)		(0.0530)		(0.0299)	
	[0.1932]		[0.2862]		[0.1407]	
Ratings using Ratings	0.0154	212	0.0108	232	0.0775***	390
	(0.0450)		(0.0241)		(0.0264)	
	[0.0498]		[0.0377]		[0.2407]	
Composite using						
10% Ratings	0.0386	212	0.0256	232	0.0427	383
	(0.0427)		(0.0250)		(0.0264)	
	[0.1250]		[0.0895]		[0.1327]	
30% Ratings	0.0709***	212	0.0227	232	0.0762***	383
	(0.0173)		(0.0418)		(0.0231)	
	[0.2299]		[0.0794]		[0.2368]	
50% Ratings	0.0539**	212	0.0265	232	0.0775***	383
	(0.0231)		(0.0251)		(0.0217)	
	[0.1747]		[0.0924]		[0.2406]	
70% Ratings	0.0323	212	0.0131	232	0.0794***	383
	(0.0218)		(0.0254)		(0.0202)	
	[0.1048]		[0.0459]		[0.2465]	
90% Ratings	0.0093	212	-0.0185	232	0.0834***	383
	(0.0207)		(0.0229)		(0.0186)	
	[0.0301]		[-0.0644]		[0.2591]	

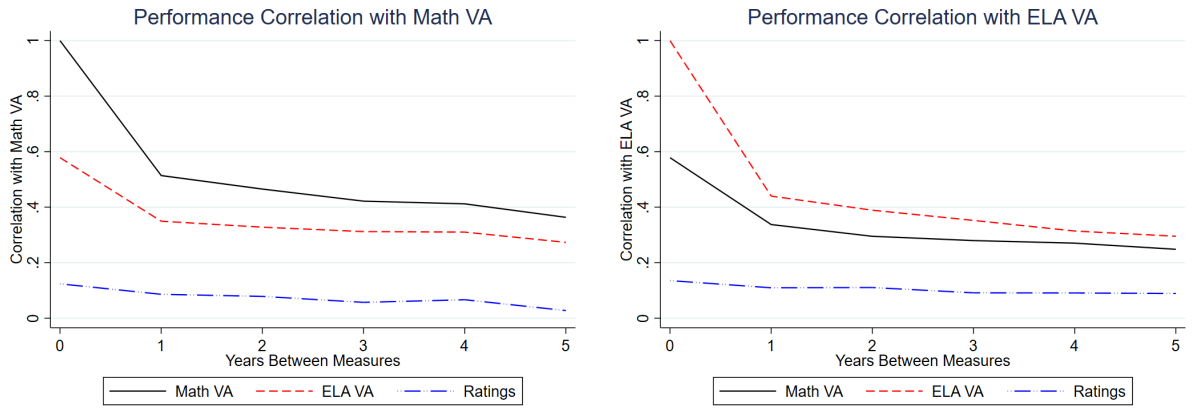
Notes: This table shows the change in performance generated when extending the pretenure period from 1 to 3 years and dismissing the bottom 50% of teachers. I use the random forest algorithms defined in Section 3.1. The row headers define the outcome and predictors. The first two columns show the change in math value-added and number of holdout observations. The remaining columns are defined similarly for ELA value-added and summative ratings.

Standard errors generated using 1,000 bootstrapped samples in parentheses. Performance units rescaled to standard deviation 1 in the dataset are included in brackets.

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

Figures

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



Panel C: Summative Ratings

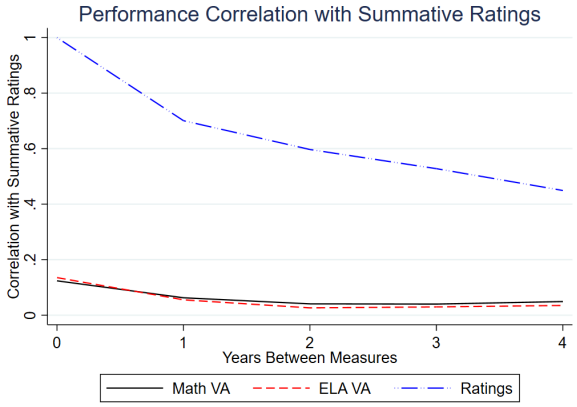


Figure 1: Performance Correlations

Notes: This figure plots the within-teacher correlation between each of the performance measures. Following Jacob and Lefgren (2008), I correct for measurement error. The x-axis measures the time between performance measures, while the y-axis measures the correlation with the metric labeled in each graph. Solid black lines depict the correlations between the y-axis variable and math value-added. Dashed red lines depict this relationship with ELA value-added, while dashed and dotted blue lines depict this relationship with ratings.

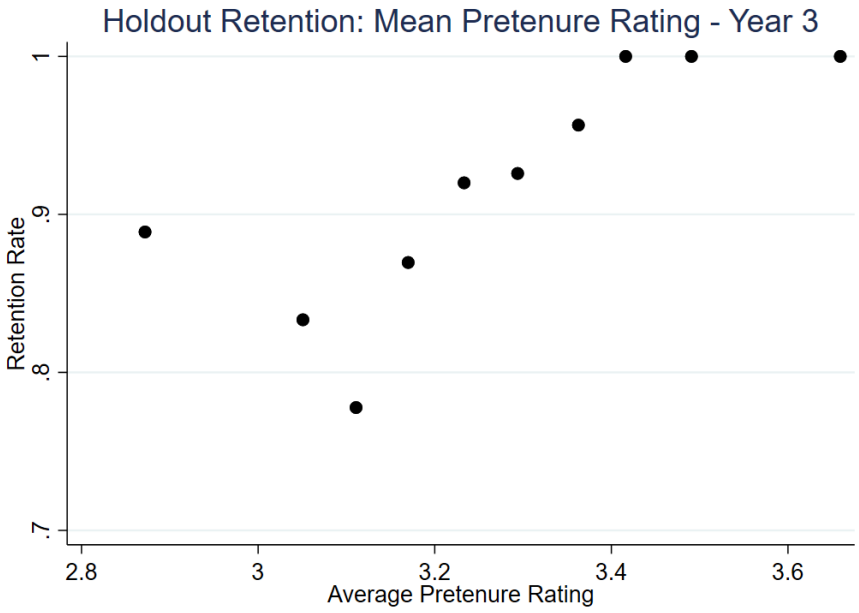


Figure 2: Retention and Mean Summative Ratings

Notes: This figure shows the relationship between mean summative ratings in years 1–3 and retention rates. The x-axis records the mean pretenure summative rating in 10 equal-sized bins, while the y-axis records the average retention rate within that bin. The sample is restricted to holdout observations.

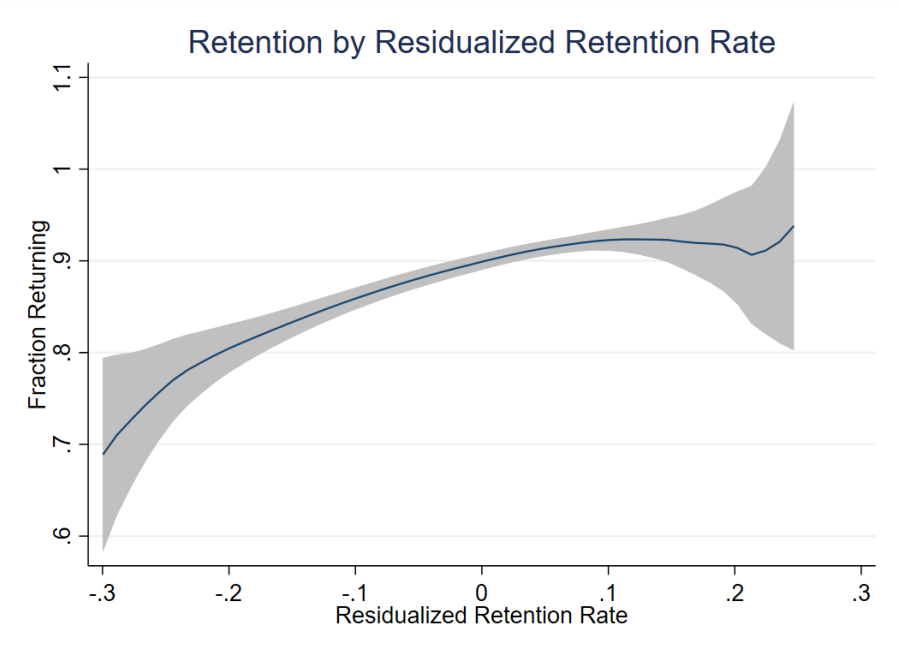
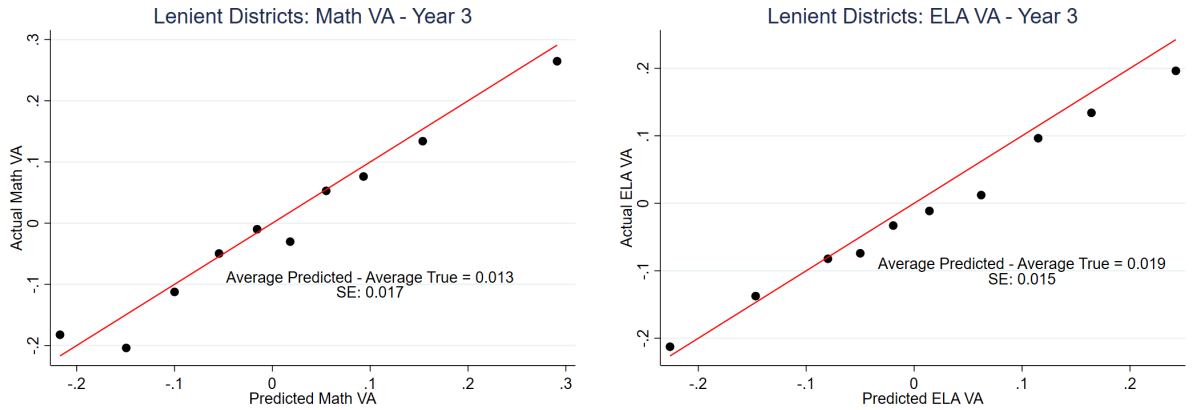


Figure 3: Retention by Residualized Retention Rate

Notes: This figure plots a local quadratic regression of the retention rate against the leave-one-out mean residual from equation (4). The x-axis records the leave-one-out mean residual, while the y-axis shows the retention rate. The plotted line uses a local quadratic regression with the Epanechnikov kernel and a bandwidth of 0.118. The shaded area shows the 95% confidence interval. The graph is truncated at residuals of -0.3 and 0.3. This truncation includes over 98% of the observations. Observations outside of this range are sparce and generate noisy estimates.

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



Panel C: Summative Ratings

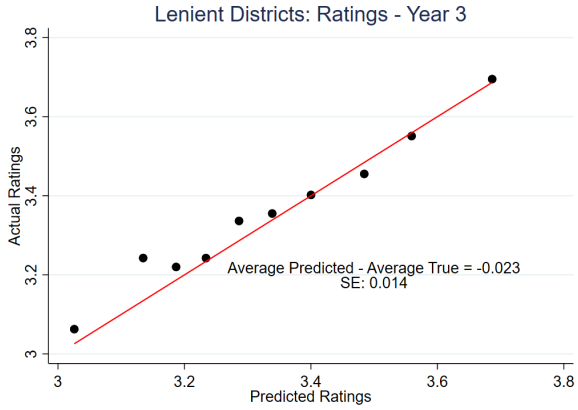
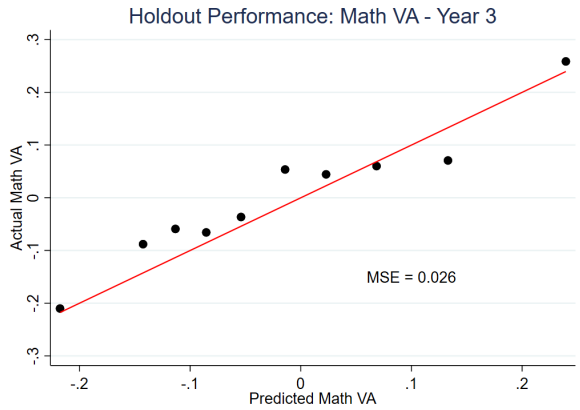


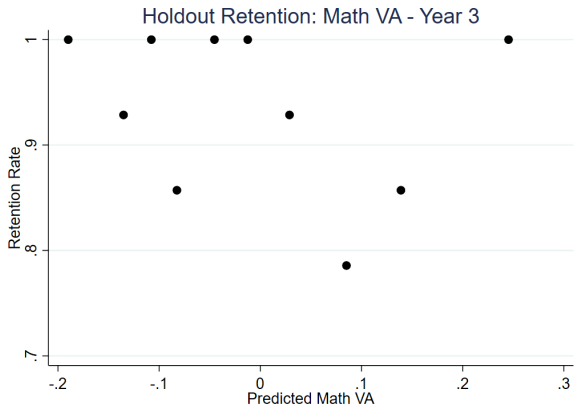
Figure 4: Actual and Predicted Performance in Lenient Districts Trained on Strict Districts

Notes: This figure shows the relationship between predicted and actual subsequent performance in lenient districts based on algorithms trained in strict districts. I use the random forest algorithms defined in Section 3.1 based on three years of data. The performance measure of interest is labeled in each graph. The x-axis records the mean predicted performance in 10 equal-sized bins, while the y-axis records the average actual performance within that bin. I define lenient (strict) districts as those with above (below) median leave-one-out average residuals from equation (4). In each graph, I include 45° lines, the mean deviation, and the standard error of the deviation.

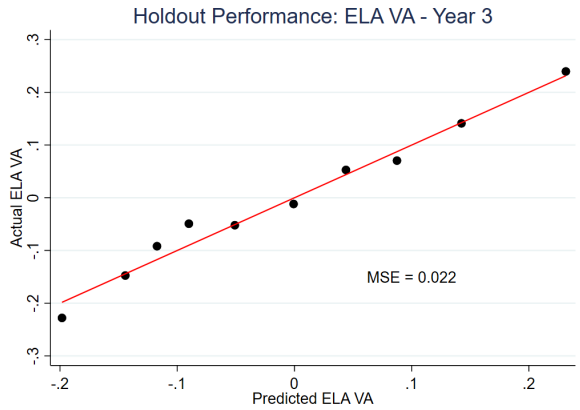
Panel A: Math Value-Added Performance



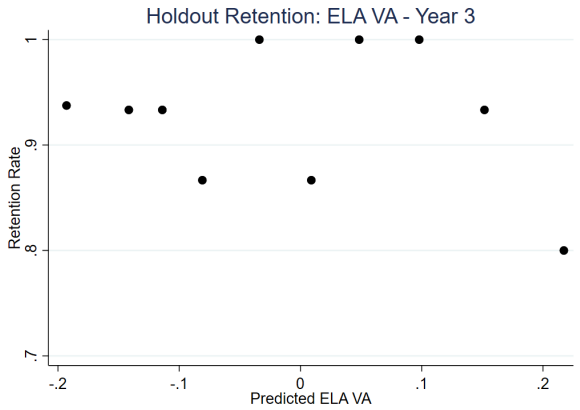
Panel B: Math Value-Added Retention



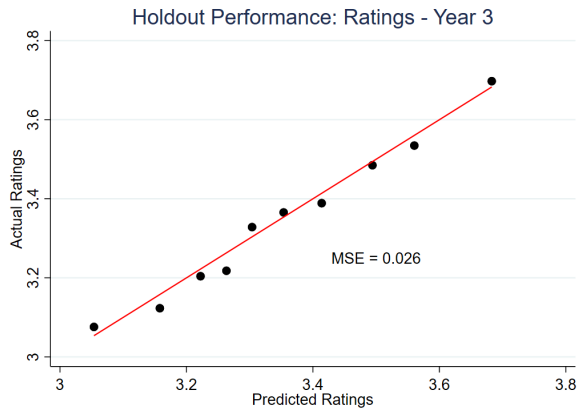
Panel C: ELA Value-Added Performance



Panel D: ELA Value-Added Retention



Panel E: Summative Ratings Performance



Panel F: Summative Ratings Retention

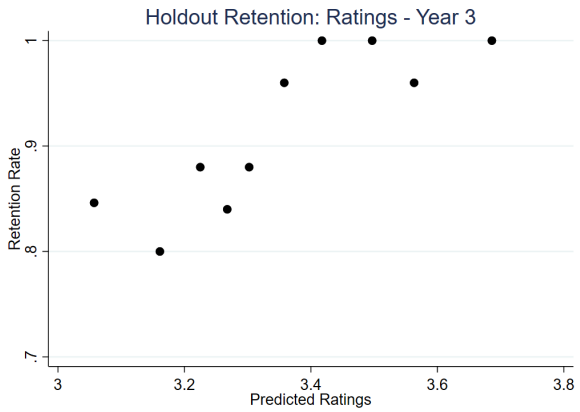
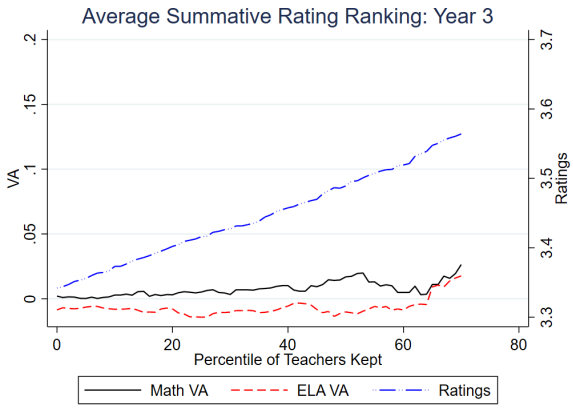


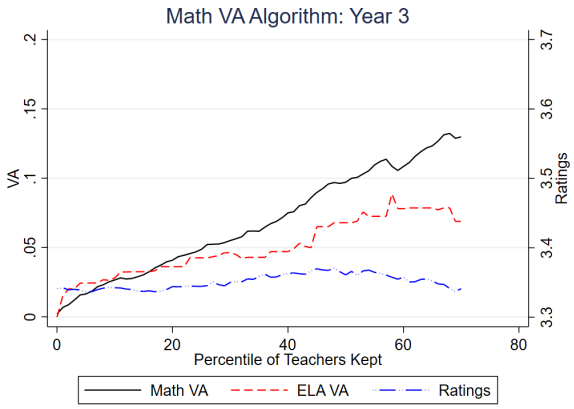
Figure 5: Actual Outcomes and Predicted Performance

Notes: This figure shows the relationship between actual outcomes and predicted subsequent performance in the holdout sample. I use the random forest algorithms defined in Section 3.1 based on three years of data. Panels A, C, and E show the relationship between predicted and actual performance, while Panels B, D, and F show the relationship between predicted performance and retention rates. Panels A and B use math value-added, Panels C and D use ELA value-added, and Panels E and F use summative ratings. The x-axis records the mean predicted performance in 10 equal-sized bins, while the y-axis records the average actual performance or retention rate within that bin. In the left graphs, I include 45° lines and the mean squared error (MSE) of the predictions.

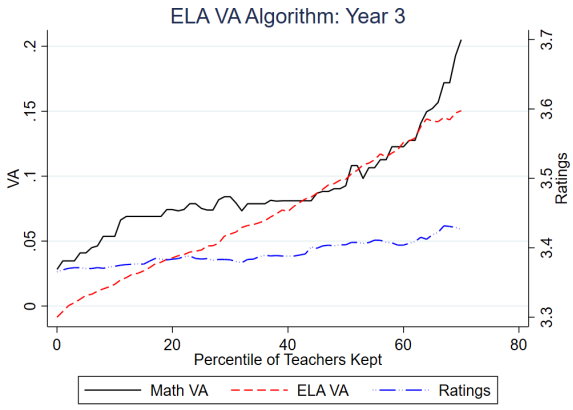
Panel A: Average Summative Ratings



Panel B: Math Value-Added Algorithm



Panel C: ELA Value-Added Algorithm



Panel D: Summative Ratings Algorithm

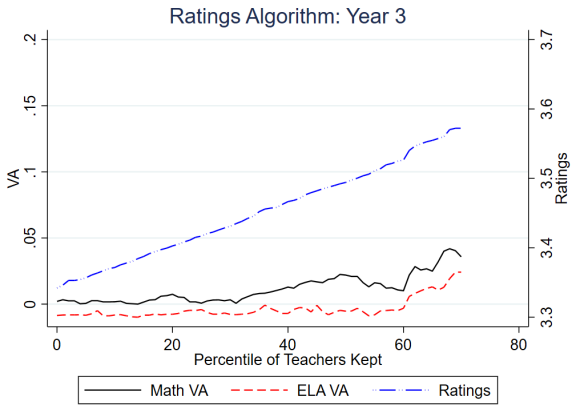
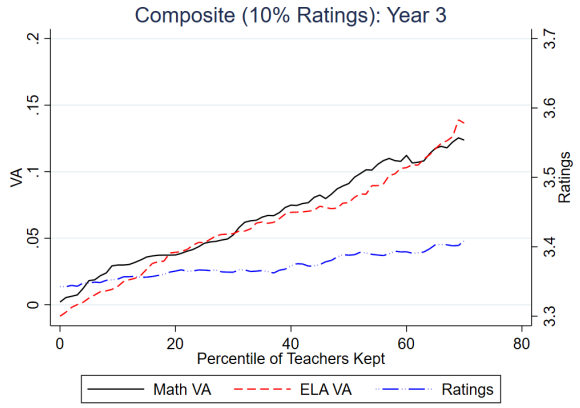


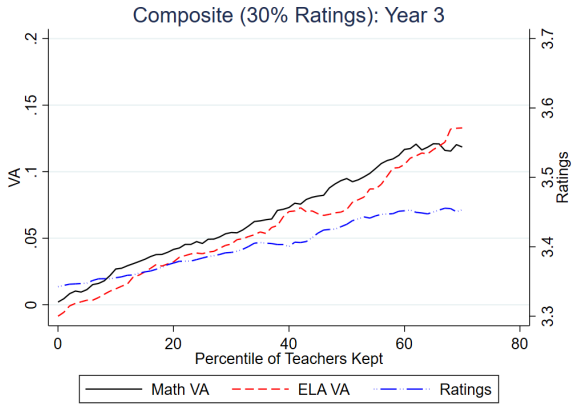
Figure 6: Mean Performance by Percentile

Notes: This figure plots the mean subsequent performance when changing minimum performance standards. Panel A uses mean summative ratings in the teacher's first three years. Panels B–D use the random forest algorithms defined in Section 3.1 based on three years of data. The x-axis shows the minimum percentile retained and the y-axis shows the performance of retained teachers. The left y-axis measures value-added standard deviations, while the right y-axis measures summative rating points. The solid black line shows math value-added, while the dashed red line shows ELA value-added. The dashed and dotted blue line shows summative ratings.

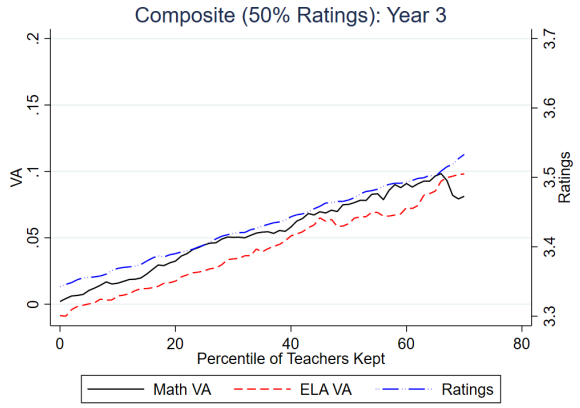
Panel A: 10% Summative Rating Weight



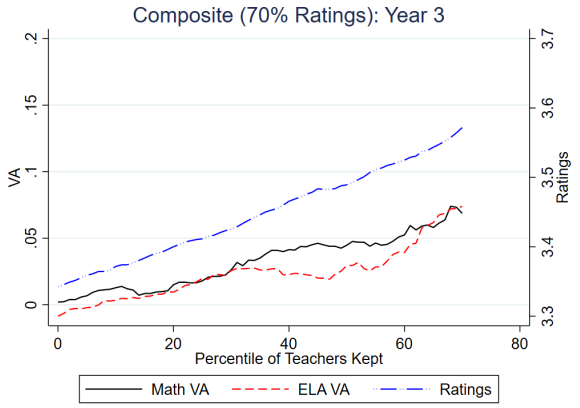
Panel B: 30% Summative Rating Weight



Panel C: 50% Summative Rating Weight



Panel D: 70% Summative Rating Weight



Panel E: 90% Summative Rating Weight

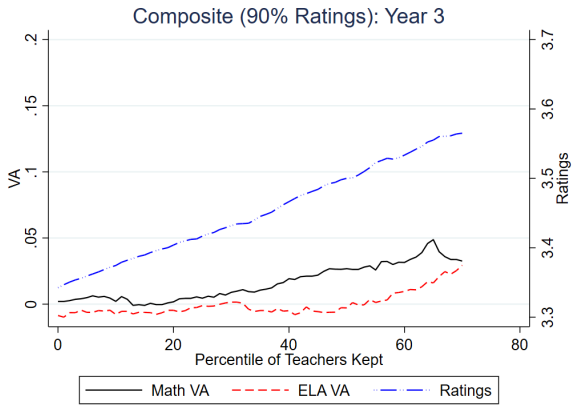
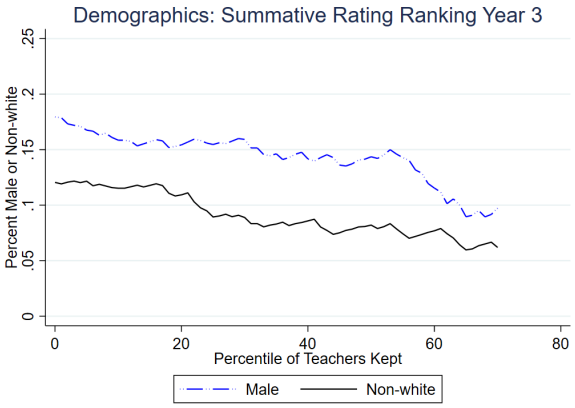


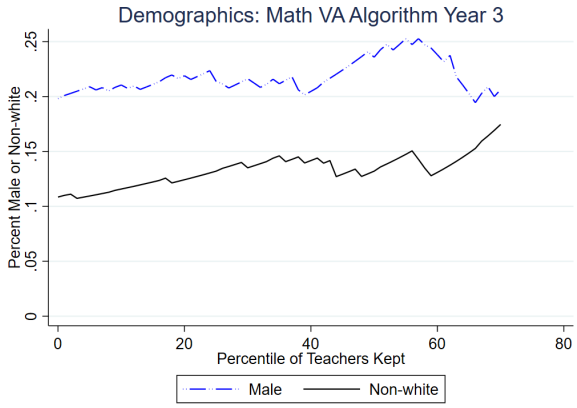
Figure 7: Mean Performance by Percentile using Composite Measure

Notes: This figure plots the mean subsequent performance when changing minimum performance standards. I use the random forest algorithms defined in Section 3.1 based on three years of data. To train the algorithm, I use the composite measure defined in Section 3.1 based on the weights defined in each graph's title. The x-axis shows the minimum percentile retained and the y-axis shows the performance of retained teachers. The left y-axis measures value-added standard deviations, while the right y-axis measures summative rating points. The solid black line shows math value-added, while the dashed red line shows ELA value-added. The dashed and dotted blue line shows summative ratings.

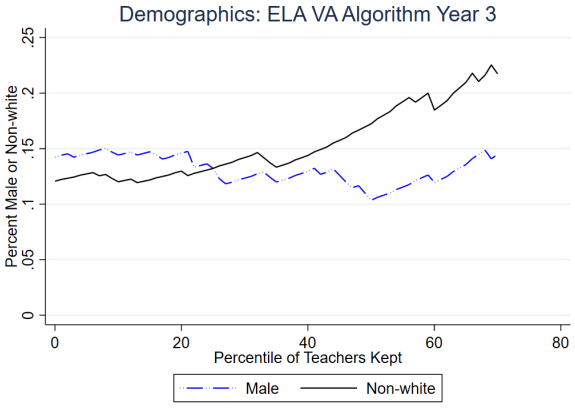
Panel A: Average Summative Ratings



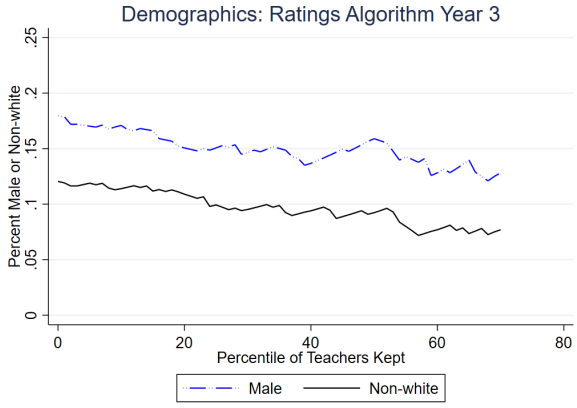
Panel B: Math Value-Added Algorithm



Panel C: ELA Value-Added Algorithm



Panel D: Summative Ratings Algorithm



Panel E: 50% Summative Rating Weight Composite

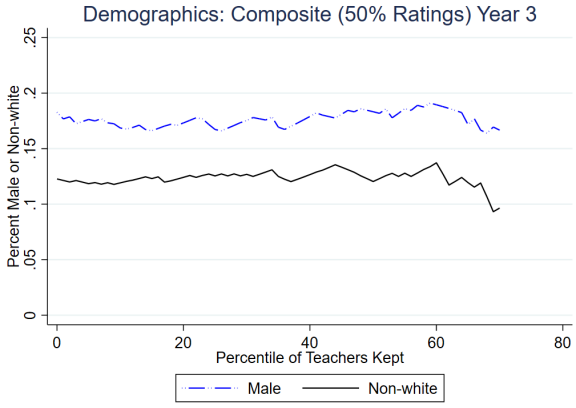
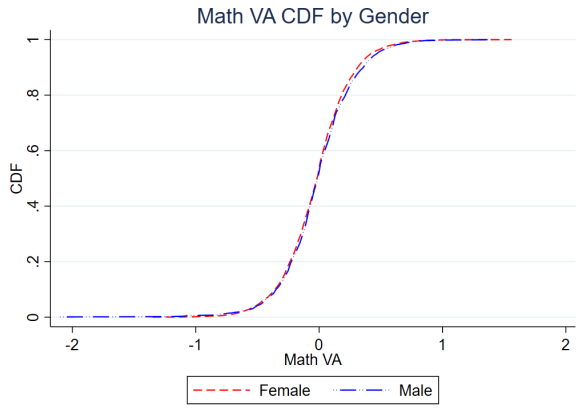


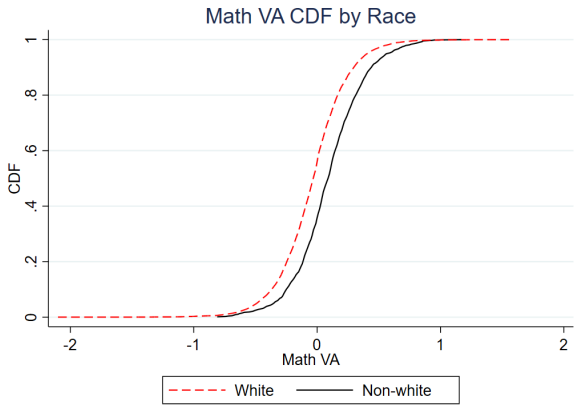
Figure 8: Mean Demographics by Percentile using ML

Notes: This figure plots the mean gender and race of retained teachers when changing minimum performance standards. Panel A uses mean summative ratings in the teacher's first three years. Panels B–E use the random forest algorithms defined in Section 3.1 based on three years of data. Panel E uses the composite measure defined in Section 3.1 with 50% weight on value-added and 50% weight on summative ratings. The x-axis shows the minimum percentile retained and the y-axis shows the demographics of retained teachers. The solid black line shows the results for race, while the dashed and dotted blue line shows the results for gender. The non-white category includes Black and Hispanic teachers.

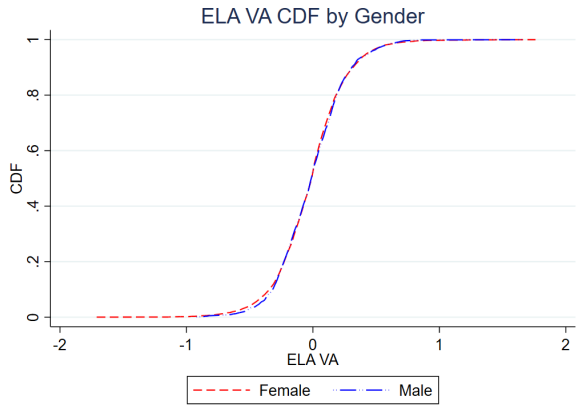
Panel A: Math Value-Added by Gender



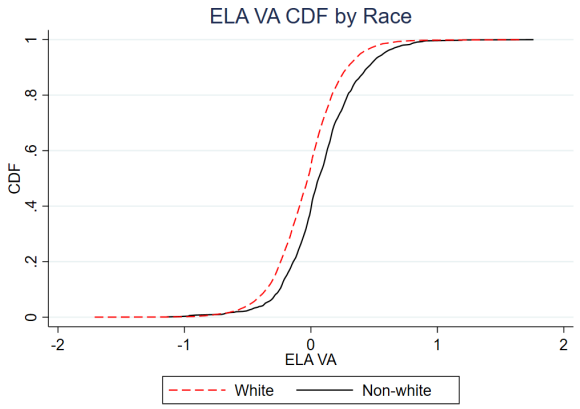
Panel B: Math Value-Added by Race



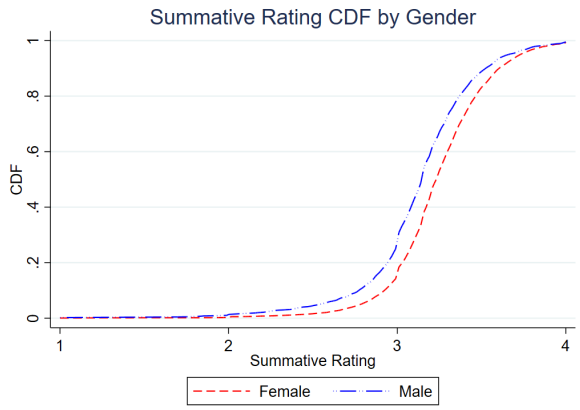
Panel C: ELA Value-Added by Gender



Panel D: ELA Value-Added by Race



Panel E: Summative Ratings by Gender



Panel F: Summative Ratings by Race

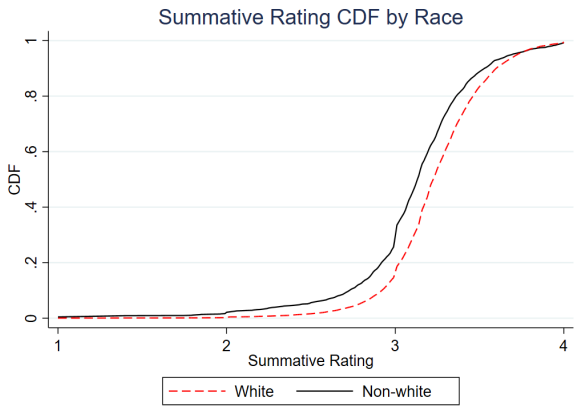
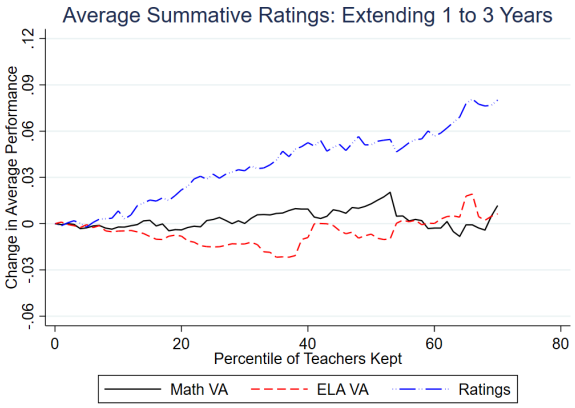


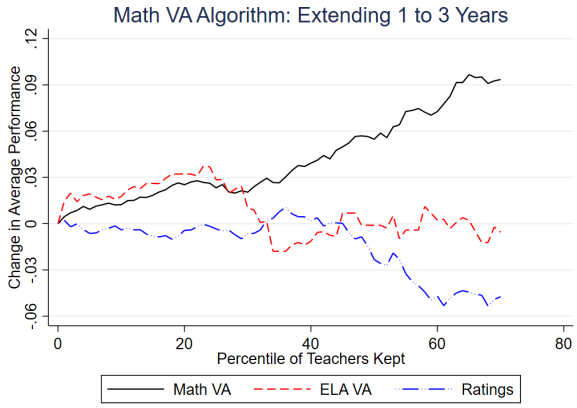
Figure 9: Performance CDF by Gender and Race

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, dashed red lines show female teachers, while dashed and dotted blue lines depict male teachers. For race, dashed red lines show white teachers, while solid black lines show non-white teachers.

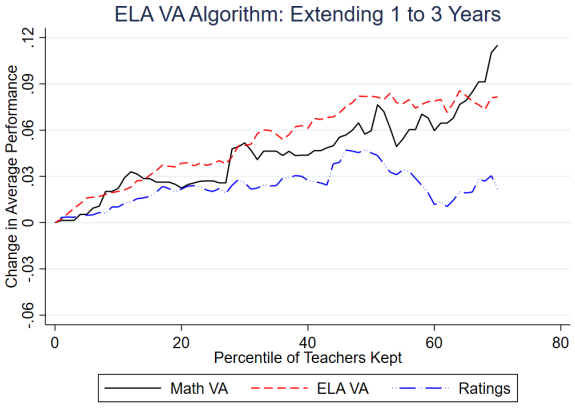
Panel A: Average Summative Ratings



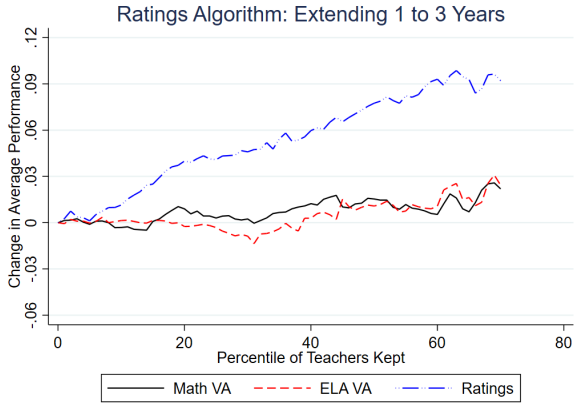
Panel B: Math Value-Added Algorithm



Panel C: ELA Value-Added Algorithm



Panel D: Summative Ratings Algorithm



Panel E: 50% Summative Rating Weight Composite

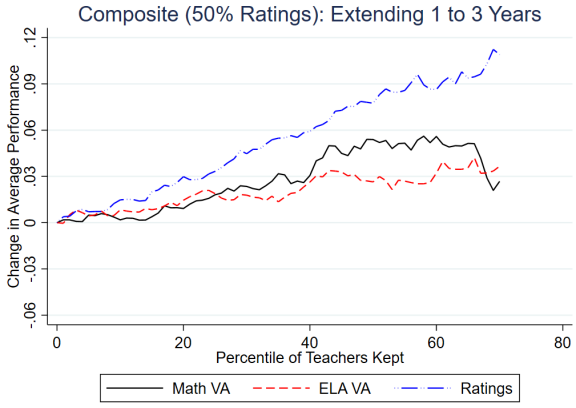
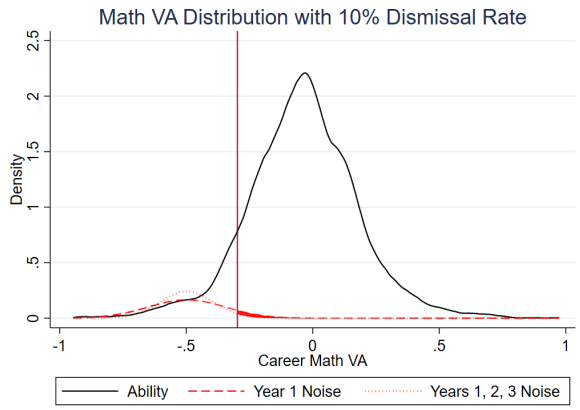


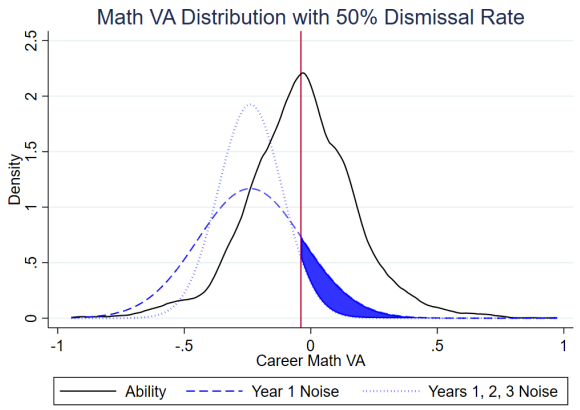
Figure 10: Extending Pretenure Period from 1 to 3 Years using Composite Ranking

Notes: This figure plots the change in average subsequent performance when using 3 years of data rather than 1 year of data. Panel A uses mean summative ratings. Panels B–E use the random forest algorithms defined in Section 3.1. Panel E uses the composite measure defined in Section 3.1 with 50% weight on value-added and 50% weight on summative ratings. The x-axis shows the minimum percentile retained and the y-axis shows the change in performance of retained teachers when extending the pretenure period from 1 to 3 years. The solid black line shows math value-added, while the dashed red line shows ELA value-added. The dashed and dotted blue line shows summative ratings.

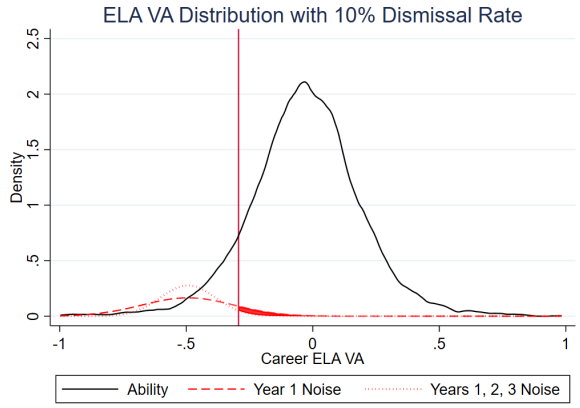
Panel A: Math Value-Added 10% Dismissal



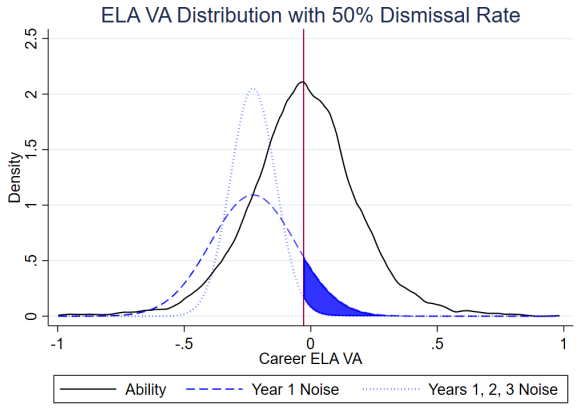
Panel B: Math Value-Added 50% Dismissal



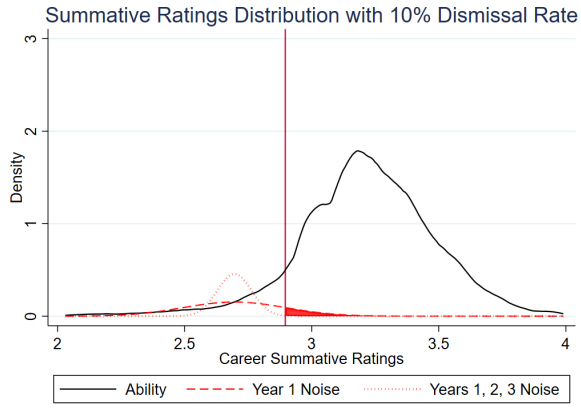
Panel C: ELA Value-Added 10% Dismissal



Panel D: ELA Value-Added 50% Dismissal



Panel E: Summative Ratings 10% Dismissal



Panel F: Summative Ratings 50% Dismissal

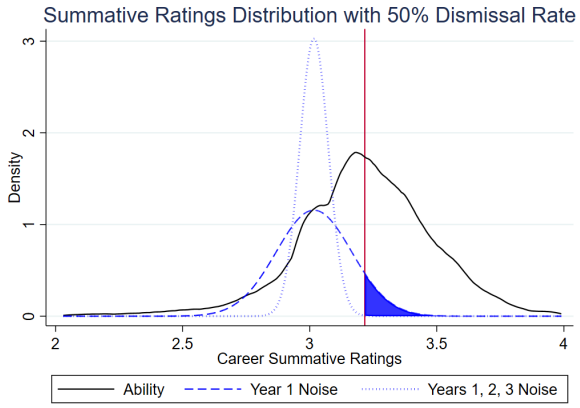


Figure 11: Performance Distribution and Noise

Notes: This figure plots performance kernel densities. Panels A and B show the results for math value-added. In Panel A, the vertical line shows the 10^{th} percentile performance. The red dashed line shows annual within-teacher variation at 0.2 standard deviations below the 10^{th} percentile, while the dotted line shows within-teacher variation in three-year pretenure performance. The standard deviations are calculated using the mean squared errors of annual performance relative to career performance for observations within 0.1 standard deviations. The red area represents the additional density of top 90 percentile teachers correctly classified using three years rather than one year of data. The distributions are scaled to the density of career value-added at that point. Panel B is defined similarly for 50^{th} percentile teachers. Panels C–F are defined similarly for ELA value-added and summative ratings.

A Appendix

A.1 Transition from NJASK and HSPA to PARCC

The transition from the NJASK and HSPA to the PARCC in 2014 could confound the results if the estimated value-added differed between the tests. To evaluate the reliability of the value-added estimate across tests, I measure the within-teacher correlation in value-added across years. If the correlation in teacher-year value-added within one test (NJASK/HSPA or PARCC) matches the correlation in teacher-year value-added across tests (between the NJASK/HSPA and PARCC), the assessments likely estimate a similar value-added.

Table A11 show the value-added correlations within teachers over time. In Panel A, the math value-added correlations across tests are similar to the correlations within tests. For example, the correlation between 2015 and 2016 PARCC math value-added is 0.44, while the correlation between 2014 NJASK and 2015 PARCC math value-added is 0.43. In Panel B, the ELA value-added correlations are higher within tests than across tests. However, since I find no evidence of math value-added bias and all the value-added results are similar across subjects, the test transition appears to generate little bias.

A.2 New Jersey Summative Rating Implementation

Teacher summative ratings were carefully implemented in New Jersey following the passage of the 2012 Teacher Effectiveness and Accountability for the Children of New Jersey (TEACHNJ) Act (State of New Jersey Department of Education, 2017). This law provided districts with the autonomy to implement their own evaluation systems. 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.

A.3 Teacher Retention Criteria

In New Jersey, teacher retention criteria are defined by the TEACHNJ Act. According to TEACHNJ, summative ratings dictate tenure receipt and job security. Teachers must earn two effective or highly effective ratings to earn tenure. In addition, tenured teachers rated ineffective or partially effective for consecutive years may receive a charge of inefficiency. However, districts retain some discretion, as they may offer a third opportunity to teachers whose second rating was partially effective. After receiving a charge of inefficiency, the teacher's tenure status is subject to an arbitration process of no more than 48 days. If the arbitrator rules in favor of the district, the teacher's employment is terminated.

A.4 Appendix Tables

Table A1: Summary Statistics

	Students	Teachers
Female	0.484	0.817
	(0.500)	(0.387)
Black	0.197	0.063
	(0.398)	(0.242)
Hispanic	0.271	0.078
	(0.445)	(0.269)
Non-white	0.402	0.136
	(0.490)	(0.342)
Urban	0.911	0.910
	(0.285)	(0.286)
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)	
Graduate Degree		0.334
		(0.472)
Experience		2.900
		(1.435)
Years in District		2.733
		(1.451)
Summative Rating		3.267
		(0.313)
Obs	12,405,063	24,251
Unique Obs	2,164,750	10,396

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. The non-white category includes Black and Hispanic, which are not mutually exclusive.

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
Teachers with Non-Missing Data	40,703	46,132	154,670
Has Both VA and Ratings	33,600	37,854	52,029
Has Year 1 Performance	2,885	3,196	4,760
Has Performance up to Year 3	1,063	1,181	1,937
Training Sample	434	466	776
Imputing Sample	417	483	771
Holdout Sample	212	232	390

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 with both value-added and summative ratings. In the third row, I restrict the sample to novice teachers with performance data in year 1. In the fourth row, I restrict the sample to teachers with performance data in years 1, 2, and 3. The final three rows records the number of observations in the training, imputing, and holdout samples. These sample represent approximately 40%, 40%, and 20% of the remaining samples, respectively.

Table A4: Difference in Performance Using All Teachers

	Math VA	N	ELA VA	N	Ratings	N
Mean Ratings (Baseline)	0.0037***	3,331	0.0009	3,690	0.0384***	5,902
	(0.0013)		(0.0011)		(0.0012)	
	[0.0120]		[0.0032]		[0.1192]	
Math using Math	0.0219***	3,331	0.0185***	1,232	-0.0290***	3,331
	(0.0017)		(0.0034)		(0.0022)	
	[0.0710]		[0.0645]		[-0.0902]	
ELA using ELA	0.0134***	1,232	0.0212***	3,690	-0.0281***	3,690
	(0.0016)		(0.0034)		(0.0024)	
	[0.0436]		[0.0739]		[-0.0872]	
Ratings using Ratings	0.0026	3,331	0.0006	3,690	0.0025	5,902
	(0.0034)		(0.0017)		(0.0023)	
	[0.0085]		[0.0022]		[0.0078]	
Composite using						
10% Ratings	0.0208***	3,331	0.0205***	3,690	-0.0251***	5,789
	(0.0036)		(0.0017)		(0.0024)	
	[0.0676]		[0.0715]		[-0.0779]	
30% Ratings	0.0179***	3,331	0.0176***	3,690	-0.0146***	5,789
	(0.0011)		(0.0016)		(0.0011)	
	[0.0579]		[0.0613]		[-0.0454]	
50% Ratings	0.0139***	3,331	0.0123***	3,690	-0.0045***	5,789
	(0.0019)		(0.0018)		(0.0017)	
	[0.0452]		[0.0428]		[-0.0141]	
70% Ratings	0.0106***	3,331	0.0075***	3,690	-0.0007	5,789
	(0.0017)		(0.0016)		(0.0015)	
	[0.0343]		[0.0261]		[-0.0020]	
90% Ratings	0.0037**	3,331	0.0023	3,690	0.0013	5,789
	(0.0016)		(0.0014)		(0.0013)	
	[0.0120]		[0.0080]		[0.0039]	

Notes: This table shows the change in performance generated when dismissing the bottom 10% of teachers using random forest algorithms defined in Section 3.1 and three years of data. In this table, I use all teachers in my dataset rather than just novice teachers. The first row shows the change in performance generated when dismissing the bottom 10% of teachers using mean summative ratings relative to no dismissals. The remaining rows record changes relative to the first row using the algorithms defined in the row header. The first two columns show the change in math value-added and number of holdout observations. The remaining columns are defined similarly for ELA value-added and summative ratings.

Standard errors generated using 1,000 bootstrapped samples in parentheses. Performance units rescaled to standard deviation 1 in the dataset are included in brackets.

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

Table A5: Gains from Extending Pretenure from 1 to 2 Years: 10% Dismissal Rate

	Math VA	N	ELA VA	N	Ratings	N
Mean Ratings	0.0011	345	0.0052	352	0.0050	592
	(0.0035)		(0.0043)		(0.0041)	
	[0.0035]		[0.0182]		[0.0154]	
Math using Math	-0.0004	345	-0.0022	114	-0.0049	345
	(0.0115)		(0.0318)		(0.0146)	
	[-0.0014]		[-0.0077]		[-0.0153]	
ELA using ELA	-0.0049	114	0.0038	352	0.0019	352
	(0.0118)		(0.0322)		(0.0145)	
	[-0.0160]		[0.0133]		[0.0059]	
Ratings using Ratings	0.0019	345	0.0082	352	0.0052	592
	(0.0274)		(0.0124)		(0.0140)	
	[0.0061]		[0.0285]		[0.0163]	
Composite using						
10% Ratings	0.0051	345	0.0033	352	0.0028	583
	(0.0271)		(0.0123)		(0.0141)	
	[0.0165]		[0.0117]		[0.0087]	
30% Ratings	-0.0035	345	0.0087	352	0.0042	583
	(0.0114)		(0.0300)		(0.0126)	
	[-0.0115]		[0.0305]		[0.0131]	
50% Ratings	0.0012	345	0.0064	352	0.0025	583
	(0.0112)		(0.0122)		(0.0088)	
	[0.0040]		[0.0224]		[0.0078]	
70% Ratings	0.0064	345	0.0049	352	0.0119	583
	(0.0110)		(0.0124)		(0.0086)	
	[0.0206]		[0.0172]		[0.0369]	
90% Ratings	-0.0024	345	0.0013	352	0.0095	583
	(0.0109)		(0.0119)		(0.0085)	
	[-0.0077]		[0.0046]		[0.0296]	

Notes: This table shows the change in performance generated when extending the pretenure period from 1 to 2 years and dismissing the bottom 10% of teachers. I use the random forest algorithms defined in Section 3.1. The row headers define the outcome and predictors. The first two columns show the change in math value-added and number of holdout observations. The remaining columns are defined similarly for ELA value-added and summative ratings.

Standard errors generated using 1,000 bootstrapped samples in parentheses. Performance units rescaled to standard deviation 1 in the dataset are included in brackets.

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

Table A6: Gains from Extending Pretenure from 2 to 3 Years: 10% Dismissal Rate

	Math VA	N	ELA VA	N	Ratings	N
Mean Ratings	0.0007	212	-0.0060	232	0.0036	390
	(0.0040)		(0.0040)		(0.0048)	
	[0.0023]		[-0.0210]		[0.0111]	
Math using Math	0.0117	212	0.0111	61	0.0031	212
	(0.0141)		(0.0435)		(0.0210)	
	[0.0380]		[0.0387]		[0.0095]	
ELA using ELA	0.0192	61	0.0151	232	0.0050	232
	(0.0139)		(0.0431)		(0.0211)	
	[0.0621]		[0.0528]		[0.0154]	
Ratings using Ratings	-0.0008	212	-0.0057	232	-0.0007	390
	(0.0432)		(0.0140)		(0.0171)	
	[-0.0027]		[-0.0198]		[-0.0022]	
Composite using						
10% Ratings	0.0148	212	0.0077	232	0.0087	383
	(0.0433)		(0.0139)		(0.0171)	
	[0.0480]		[0.0269]		[0.0270]	
30% Ratings	0.0126	212	0.0027	232	0.0023	383
	(0.0130)		(0.0416)		(0.0197)	
	[0.0410]		[0.0094]		[0.0072]	
50% Ratings	0.0040	212	0.0059	232	0.0131	383
	(0.0133)		(0.0141)		(0.0102)	
	[0.0129]		[0.0205]		[0.0405]	
70% Ratings	0.0063	212	0.0019	232	0.0014	383
	(0.0130)		(0.0137)		(0.0101)	
	[0.0203]		[0.0065]		[0.0042]	
90% Ratings	0.0023	212	-0.0057	232	0.0056	383
	(0.0130)		(0.0137)		(0.0097)	
	[0.0075]		[-0.0200]		[0.0175]	

Notes: This table shows the change in performance generated when extending the pretenure period from 2 to 3 years and dismissing the bottom 10% of teachers. I use the random forest algorithms defined in Section 3.1. The row headers define the outcome and predictors. The first two columns show the change in math value-added and number of holdout observations. The remaining columns are defined similarly for ELA value-added and summative ratings.

Standard errors generated using 1,000 bootstrapped samples in parentheses. Performance units rescaled to standard deviation 1 in the dataset are included in brackets.

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

Table A7: Gains from Extending Pretenure from 1 to 2 Years: 50% Dismissal Rate

	Math VA	N	ELA VA	N	Ratings	N
Mean Ratings	-0.0016	345	0.0024	352	0.0427***	592
	(0.0119)		(0.0130)		(0.0120)	
	[-0.0051]		[0.0083]		[0.1326]	
Math using Math	0.0286	345	0.0179	114	-0.0539**	345
	(0.0220)		(0.0440)		(0.0232)	
	[0.0928]		[0.0625]		[-0.1675]	
ELA using ELA	0.0609***	114	0.0557	352	0.0305	352
	(0.0210)		(0.0420)		(0.0241)	
	[0.1976]		[0.1944]		[0.0946]	
Ratings using Ratings	-0.0062	345	0.0110	352	0.0516**	592
	(0.0388)		(0.0213)		(0.0228)	
	[-0.0200]		[0.0386]		[0.1604]	
Composite using						
10% Ratings	0.0273	345	-0.0021	352	0.0199	583
	(0.0348)		(0.0223)		(0.0228)	
	[0.0886]		[-0.0073]		[0.0619]	
30% Ratings	0.0278	345	0.0182	352	0.0358**	583
	(0.0176)		(0.0335)		(0.0170)	
	[0.0902]		[0.0636]		[0.1112]	
50% Ratings	0.0434**	345	0.0244	352	0.0424**	583
	(0.0219)		(0.0218)		(0.0183)	
	[0.1406]		[0.0852]		[0.1317]	
70% Ratings	0.0277	345	0.0096	352	0.0463***	583
	(0.0212)		(0.0212)		(0.0174)	
	[0.0899]		[0.0334]		[0.1439]	
90% Ratings	0.0226	345	-0.0033	352	0.0565***	583
	(0.0206)		(0.0205)		(0.0168)	
	[0.0731]		[-0.0114]		[0.1754]	

Notes: This table shows the change in performance generated when extending the pretenure period from 1 to 2 years and dismissing the bottom 50% of teachers. I use the random forest algorithms defined in Section 3.1. The row headers define the outcome and predictors. The first two columns show the change in math value-added and number of holdout observations. The remaining columns are defined similarly for ELA value-added and summative ratings.

Standard errors generated using 1,000 bootstrapped samples in parentheses. Performance units rescaled to standard deviation 1 in the dataset are included in brackets.

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

Table A8: Gains from Extending Pretenure from 2 to 3 Years: 50% Dismissal Rate

	Math VA	N	ELA VA	N	Ratings	N
Mean Ratings	0.0111	212	-0.0113	232	0.0161	390
	(0.0131)		(0.0123)		(0.0117)	
	[0.0360]		[-0.0395]		[0.0500]	
Math using Math	0.0289	212	-0.0058	61	0.0263	212
	(0.0241)		(0.0553)		(0.0347)	
	[0.0937]		[-0.0203]		[0.0816]	
ELA using ELA	-0.0382	61	0.0294	232	-0.0050	232
	(0.0242)		(0.0535)		(0.0363)	
	[-0.1239]		[0.1028]		[-0.0156]	
Ratings using Ratings	0.0192	212	-0.0015	232	0.0309	390
	(0.0604)		(0.0242)		(0.0246)	
	[0.0623]		[-0.0052]		[0.0961]	
Composite using						
10% Ratings	0.0168	212	0.0243	232	0.0170	383
	(0.0536)		(0.0243)		(0.0251)	
	[0.0545]		[0.0849]		[0.0527]	
30% Ratings	0.0281	212	0.0025	232	0.0310	383
	(0.0185)		(0.0488)		(0.0230)	
	[0.0912]		[0.0087]		[0.0962]	
50% Ratings	0.0196	212	-0.0043	232	0.0298	383
	(0.0245)		(0.0236)		(0.0220)	
	[0.0635]		[-0.0150]		[0.0927]	
70% Ratings	0.0053	212	-0.0181	232	0.0193	383
	(0.0241)		(0.0233)		(0.0206)	
	[0.0173]		[-0.0630]		[0.0600]	
90% Ratings	-0.0161	212	-0.0188	232	0.0267	383
	(0.0211)		(0.0212)		(0.0174)	
	[-0.0522]		[-0.0657]		[0.0830]	

Notes: This table shows the change in performance generated when extending the pretenure period from 2 to 3 years and dismissing the bottom 50% of teachers. I use the random forest algorithms defined in Section 3.1. The row headers define the outcome and predictors. The first two columns show the change in math value-added and number of holdout observations. The remaining columns are defined similarly for ELA value-added and summative ratings.

Standard errors generated using 1,000 bootstrapped samples in parentheses. Performance units rescaled to standard deviation 1 in the dataset are included in brackets.

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

Table A9: Difference in Performance with Demographics: 10% Dismissal Rate

	Math VA	N	ELA VA	N	Ratings	N
Mean Ratings (Baseline)	0.0008	212	0.0005	232	0.0313***	390
	(0.0035)		(0.0043)		(0.0042)	
	[0.0026]		[0.0016]		[0.0971]	
Math using Math	0.0250***	212	0.0315**	61	-0.0320***	212
	(0.0061)		(0.0125)		(0.0094)	
	[0.0810]		[0.1099]		[-0.0996]	
ELA using ELA	0.0341***	61	0.0276**	232	-0.0108	232
	(0.0065)		(0.0127)		(0.0100)	
	[0.1105]		[0.0965]		[-0.0336]	
Ratings using Ratings	-0.0028	212	-0.0022	232	-0.0013	390
	(0.0164)		(0.0059)		(0.0069)	
	[-0.0091]		[-0.0076]		[-0.0042]	
Composite using						
10% Ratings	0.0255	212	0.0246***	232	-0.0182***	383
	(0.0164)		(0.0062)		(0.0070)	
	[0.0828]		[0.0859]		[-0.0565]	
30% Ratings	0.0246***	212	0.0245***	232	-0.0152**	383
	(0.0051)		(0.0040)		(0.0066)	
	[0.0798]		[0.0856]		[-0.0473]	
50% Ratings	0.0162**	212	0.0149**	232	-0.0095	383
	(0.0071)		(0.0059)		(0.0063)	
	[0.0524]		[0.0521]		[-0.0294]	
70% Ratings	0.0073	212	0.0099	232	-0.0024	383
	(0.0067)		(0.0061)		(0.0057)	
	[0.0237]		[0.0346]		[-0.0074]	
90% Ratings	0.0009	212	0.0023	232	0.0023	383
	(0.0060)		(0.0057)		(0.0054)	
	[0.0030]		[0.0079]		[0.0071]	

Notes: This table shows the change in performance generated when dismissing the bottom 10% of teachers using three years of data. These models add race (white or non-white) and gender as predictors to the random forest algorithms defined in Section 3.1. The first row shows the change generated when dismissing the bottom 10% of teachers using mean summative ratings relative to no dismissals. The remaining rows record changes relative to the first row using the algorithms defined in the row header. The first two columns show the change in math value-added and number of holdout observations. The remaining columns are defined similarly for ELA value-added and summative ratings.

Standard errors generated using 1,000 bootstrapped samples in parentheses. Performance units rescaled to standard deviation 1 in the dataset are included in brackets.

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

Table A10: Residual Difference by Group Using 3 Years of Data

	Female - Male	N	White - Non-white	N
Math using Math	-0.0089	212	-0.0281	212
	(0.0339)		(0.0472)	
	[-0.0288]		[-0.0911]	
ELA using ELA	0.0263	232	0.0270	232
	(0.0334)		(0.0465)	
	[0.0918]		[0.0943]	
Ratings using Ratings	-0.0173	390	-0.0192	390
	(0.0329)		(0.0314)	
	[-0.0538]		[-0.0596]	

Notes: This table compares group-wide average residuals generated by the machine learning algorithms described in Section 3.1. The row headers define the algorithm's outcome and predictors. The first column shows the difference between male and female teachers, while the second column shows the difference between white and non-white teachers. In the female-male (white-non-white) comparison, a negative value suggests that the algorithm underpredicts male (non-white) teacher performance relative to female (white) teacher performance.

Standard errors generated using 1,000 bootstrapped samples in parentheses. Performance units rescaled to standard deviation 1 in the dataset are included in brackets.

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

Table A11: Annual VA Correlation by Year

Panel A: Math Value-Added

	2013	2014	2015	2016	2017	2018
NJASK:2013	1.00					
NJASK:2014	0.48	1.00				
PARCC:2015	0.38	0.43	1.00			
PARCC:2016	0.35	0.43	0.44	1.00		
PARCC:2017	0.36	0.40	0.41	0.49	1.00	
PARCC:2018	0.33	0.39	0.40	0.48	0.51	1.00

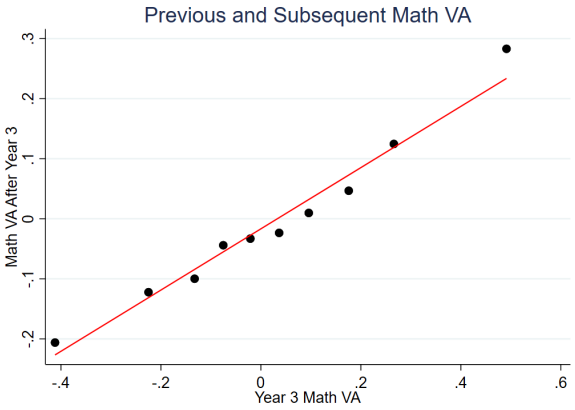
Panel B: ELA Value-Added

	2013	2014	2015	2016	2017	2018
NJASK:2013	1.00					
NJASK:2014	0.40	1.00				
PARCC:2015	0.27	0.29	1.00			
PARCC:2016	0.30	0.33	0.39	1.00		
PARCC:2017	0.27	0.29	0.35	0.45	1.00	
PARCC:2018	0.27	0.30	0.36	0.44	0.46	1.00

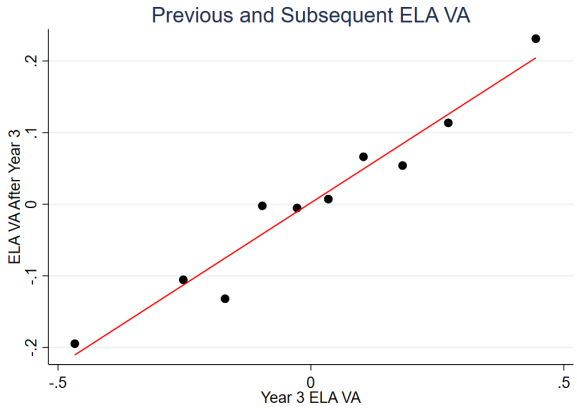
Notes: This table shows within-teacher math (Panel A) and ELA (Panel B) value-added correlations over time. The rows and columns define the test year used to generate the value-added estimate. NJASK exams were administered in 2013 and 2014, while PARCC exams were administered from 2015 to 2018.

A.5 Appendix Figures

Panel A: Math Value-Added



Panel B: ELA Value-Added



Panel C: Summative Ratings

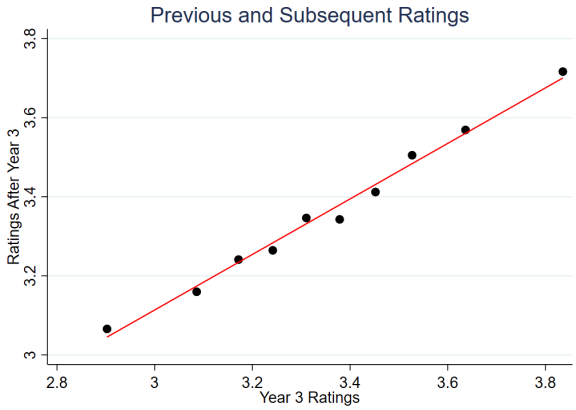


Figure A1: Relationship Between Previous and Subsequent Performance

Notes: This figure shows the relationship between year 3 performance and the actual subsequent performance. The performance measure of interest is labeled in each graph. The x-axis records the average year 3 performance in 10 equal-sized bins, while the y-axis records the average subsequent performance within that bin. The graphs include a line of best fit generated using a linear regression.