Identifying High-Quality Teachers

Kevin Ng*

December 5, 2023

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 rely on evaluation scores to guide these personnel decisions without considering other dimensions of teacher performance. I use predictive models to rank teachers based on expected 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 negatively impact selection into teaching. I estimate optimal weights on each performance measure to maximize measures of teacher quality using a range of utility parameters. These improvements are a product of using additional information (value-added) rather than advanced algorithms, as OLS regressions and advanced machine learning techniques produce similar gains. In comparison, prediction models 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.

Keywords: Teacher selection, predictive models, teacher evaluation

*Ng: CNA, 3003 Washington Boulevard, Arlington VA 22201 (email: ngk@cna.org). For helpful comments, I thank Michael Lovenheim, Evan Riehl, Zhuan Pei, Cornell Labor Seminar participants, and several anonymous referees. This work is based on my doctoral dissertation while attending Cornell University. I am grateful to the New Jersey Department of Education for assistance with the data. The views expressed in this paper are my own and do not reflect those of the CNA Corporation or any of its sponsors. They also 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.[1] While these predictions could incorporate multiple dimensions of performance, 17 states rely on evaluations based on classroom observations without including objective measures of student growth (Ross & Walsh, 2019). The remaining states place most weight on classroom observations to generate their overall ratings. Evaluations capture characteristics that are distinct from value-added, such as classroom management and professionalism. Without considering all metrics, 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.

In this paper, I use teacher-student linked administrative data from the New Jersey Department of Education (NJDOE) to evaluate three questions. First, can districts improve average retained teacher quality by supplementing evaluations with value-added? Second, are these improvements a product of additional information or sophisticated algorithms? Third, does extending the pretenure period improve average retained teacher performance?

To evaluate the returns to utilizing additional information and longer pretenure periods, I use predictive models relying on ordinary least squares (OLS) regressions. These models use early career teaching performance to identify which teachers will be most effective in the long-run. I then identify optimal weights on value-added and summative ratings that maximize measures of teacher quality using a range of plausible utility parameters.

To conduct this analysis, I calculate value-added using a lagged test score model. The NJDOE provides annual summative ratings, which are based on a combination of supervisor

[1] Schools may also improve performance through professional development opportunities and support.

classroom observations and student growth. These ratings[2] serve as the sole determinant of performance-based personnel decisions in New Jersey. Relative to the New Jersey benchmark, I increase the weight on value-added and record the change in average retained teacher performance. I 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 turnover rates.[3] I measure the retained teachers' subsequent value-added and ratings to compare the models. 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 prediction models. 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 dismissal residuals. Districts retain some discretion when dismissing low-performing teachers, so I compare districts with higher dismissal rates conditional on summative ratings to districts with lower dismissal rates. If high-dismissal districts select on unobserved characteristics, imputations relying on these teachers would overpredict performance in low-dismissal districts. However, this test shows no evidence of prediction bias.

Using these assumptions, I then explore the returns to supplementing summative ratings with value-added when using 10% dismissal rates. By incorporating value-added rather than following current policies that only consider ratings, districts can increase subsequent average value-added by 0.01 student test score standard deviations, as well as the diversity of the teacher labor force without causing a statistically significant decline in ratings. This

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

[3] The pretenure turnover rate is 13%, though I cannot separate voluntary from involuntary turnover.

improvement equates to a present value gain of $2,520 per student (Chetty et al., 2014), which is nearly 12 times larger than the costs associated with the productivity decline at tenure receipt (Ng, 2022). These improvements stem from using additional information (value-added), as I generate similar gains when using advanced machine learning techniques. In this analysis, I provide optimal weights on each performance measure to maximize teacher quality using a range of utility parameters. By using additional measures of performance, the proposed models provide a virtually costless method to improve average retained teacher quality.

Next, I reestimate the predictive models when adjusting the pretenure period. At current turnover rates of 10%, extending the pretenure period beyond one year does not generate statistically significant improvements in average teacher quality. Although the magnitude of the point estimates suggest similar gains to using additional data, the lack of statistical significance shows that these improvements are inconsistent. I find that 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 same time, extending the pretenure period reduces 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.

This paper contributes to the literature by estimating the returns to using additional information (value-added) 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 finds that performance measure weights play a critical role in the distribution of retained teacher proficiency rates (Steinberg & Kraft, 2017). Other work predicts subsequent teacher performance using evaluation data with both novice and experienced teachers (Harris & Sass, 2014; Winters & Cowen, 2013; Chalfin et al., 2016; Mihaly et al., 2013). Since the returns to experience vary throughout a teacher's career (Kraft & Papay, 2015; Wiswall, 2013; Hanushek & Rivkin, 2006), 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 (rather than low-stakes survey data), my paper directly addresses tenure receipt decisions.

This study also contributes to research regarding optimal teacher dismissal policies. Staiger and Rockoff (2010) argue that schools should screen candidates based on performance in their first few years. My paper extends the argument by demonstrating one technique to select high-quality teachers based on this early career performance. Other research has considered the impact of different factors on average retained teacher performance, such as introducing performance pay and increasing dismissal rates (Rothstein, 2015; Neal, 2011).

2 Data and Policy Context

Summative ratings from 2014 to 2018 measure performance using a weighted average of Teacher Practice, Student Growth Objectives (SGOs), and median Student Growth Percentiles (mSGPs).[4] In Teacher Practice, supervisors observe several classes using an NJDOE

[4] In Appendix Section A.1, I discuss the implementation of this evaluation system.

approved rubric.[5] 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 SGOs based on state standards. The SGOs 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 mSGPs, which measure score growth on state assessments. The mSGPs differ from value-added because they only account for previous test scores rather than a variety of student, classroom, and school characteristics.[6]

Table A1 shows the weighting schemes for 2014 and 2017–2018 (first two columns), as well as 2015–2016 (last two columns).[7] Summative ratings primarily rely on Teacher Practice with some weight placed on student growth. The odd columns record the weights for subjects that partially rely on state tests. The even columns show the weights for other subjects. Based on these weights, teachers receive a summative rating between 1.00 and 4.00. These ratings place teachers into one of four categories with minimum thresholds included in parentheses: ineffective (1.00), partially effective (1.85), effective (2.65), and highly effective (3.50).[8]

While Table A1 provides the actual weights, the "effective" weights depend on the distribution of scores in each component. Teacher Practice and mSGPs approximately follow normal distributions with a wide range of possible scores, though SGO scores are concentrated near perfect scores (State of New Jersey Department of Education, 2014, 2015). Given the limited dispersion of SGO scores and their low weight in Table A1, mSGPs and Teacher Practice scores have a greater impact on the variability of summative ratings across teachers.

In New Jersey, the Teacher Effectiveness and Accountability for the Children of New Jersey (TEACHNJ) Act defines teacher retention criteria. According to TEACHNJ, summative ratings dictate tenure receipt and job security. Teachers must earn two effective or

[5] "Teacher Practice Evaluation Instruments" (2019) contains the full list of approved evaluation instruments, including the widely-used Danielson Framework.

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

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

[8] Ideally, I would separate the ratings into Teacher Practice, SGOs, and mSGPs. However, the data only include combined summative ratings from 1.00 to 4.00.

highly effective ratings to earn tenure at the end of their fourth-year. 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.

To calculate value-added, 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.[9] 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.[10]

I use all teachers to estimate the following model separately for math and ELA:[11]

$$A_{ijgst} = \alpha A_{it-1} + \beta X_{it} + \eta_1 C_{it} + \eta_2 HS_i * C_{it} + \lambda S_{it} + \Theta_{jt} + \varepsilon_{ijgst} \tag{1}$$

where A_{ijgst} are the test scores of student i in teacher j's grade g class in school s and year t, which is standardized to have mean 0 and standard deviation 1 in each grade-year. I control for the student's previous year math and ELA test scores (A_{it-1}) because Walsh et al. (2018) show using both previous test scores improves the precision of the value-added estimates. I also include controls for student, classroom, and school characteristics. The

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

[10] Table A2 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.

[11] I use value-added because previous research has linked it to long-run student success (Chetty et al., 2014). In addition, Walsh and Isenberg (2015) find mSGP methods may generate biased estimates of teacher performance because they lack controls for student, classroom, and school characteristics. For example, they depress the scores of teachers who instruct many English language learners. Also, the mSGP measures are not available in the data.

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. To separately identify classroom characteristics for elementary schools where many teachers have one class and secondary schools where most teachers instruct multiple classes, I interact these characteristics with an indicator for being in grade 7 or higher (HS_i). School covariates (S_{it}) include urbanicity[12], enrollment, racial composition, and percentage of FRPL eligible.[13] Value-added is measured annually by Θ_{jt} fixed effects.[14]

To calculate career value-added, I use equation (1) but replace teacher-year fixed effects (Θ_{jt}) with teacher fixed effects (Θ_j). I also include year fixed effects (γ_t) to account for shifts in the distribution of value-added over time. Thus, I estimate the following model:

$$A_{ijgst} = \alpha A_{it-1} + \beta X_{it} + \eta_1 C_{it} + \eta_2 HS_i * C_{it} + \lambda S_{it} + \Theta_j + \gamma_t + \varepsilon_{ijgst}. \quad (2)$$

To avoid mechanical correlations, I estimate equation (2) by excluding any years that the models use as predictors. For example, if the model 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).[15]

I limit the sample to teachers who have summative ratings and value-added estimates for their first three years of experience.[16] Using these restrictions, I focus the analysis on novice

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

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

[14] Test score floor and ceiling effects are unlikely to bias the estimates for tracked classes, as the scores follow a normal distribution centered at the average score. Only 0.44% of tests receive the lowest score, while 0.94% of tests receive the highest score. Based on this information, I would be most concerned about the slightly greater test truncation at the highest score. However, Resch and Isenberg (2018) found that ceiling effects only shrink value-added toward the middle of the distribution rather than send average value-added to the bottom of the distribution. Since my simulations only dismiss teachers at the bottom of the distribution, ceiling effects would not punish those teaching higher-level content.

[15] I also do not use Bayesian shrinkage estimators (Kane & Staiger, 2008) because they shrink annual value-added toward career averages. This introduces a mechanical correlation between each value-added estimate that would overstate the predictive power of the models.

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

teachers with multiple measures of performance.[17]

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.[18] 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 and weaken over time. In fact, the solid black and dashed red lines in Panels A and B show that the correlation between math and ELA value-added ranges between 0.17 and 0.6. 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 0.14. 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 dimension.

Ideally, I would estimate the correlation between value-added estimates and the individual summative rating components to demonstrate differences between the measures. However, I lack the data to estimate these effects. I would not expect a high correlation between SGOs and value-added because they rely on different types of tests written by different entities.

[17] 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. Ex-ante, it is critical to focus the analysis on year one teachers because early career returns to experience are quite rapid and non-linear (Kraft & Papay, 2015; Wiswall, 2013; Hanushek & Rivkin, 2006). Nonetheless, the results are similar when considering all teachers.

[18] Specifically, this adjustment rescales the correlation coefficient by the standard deviation of the observed value-added estimates divided by the standard deviation of the true value-added measure.

In addition, the state tests provide snapshots of performance on testing days, while the SGOs include long-term assessments, such as written pieces with multiple opportunities for revision. When considering the relationship between value-added and mSGPs, Guarino et al. (2014) identified a correlation of 0.87, however, they found large differences in the tails of the distributions. In fact, about 30 percent of bottom quartile mSGP teachers were not bottom quartile value-added teachers. This difference is particularly relevant in this setting where I consider policies to dismiss teachers at the bottom of the performance distribution.

3 Empirical Analysis

3.1 Policymaker's Problem

Prior to proceeding with the analysis, I introduce the policymaker's problem. The policymaker seeks to maximize teacher performance by selecting the highest quality teachers, as measured by value-added and summative ratings. In this two-dimensional space, the relative importance of value-added and summative ratings is not well understood. 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). While summative ratings may capture these critical components of student success, Kraft (2019) finds evidence that the classroom observation component of summative ratings also poorly captures the development of non-cognitive skills. Nonetheless, the education community often feels uncomfortable only relying on these test-based metrics based on single-day snapshots ("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 observations to assess their teachers (Ross & Walsh, 2019). Since the trade-offs between value-added and summative ratings are not well understood, I estimate optimal weights on each performance measure to maximize measures of teacher quality across a range of

utility functions. These utility functions allow the relative importance of value-added and summative ratings to vary.

4 Using Additional Performance Measures: Value-Added

In this section, I use OLS to predict later career performance based on early career performance. While more advanced machine learning techniques exist, I use OLS for greater policy relevance and transparency. Districts could conduct the OLS regressions and provide teachers with the weights (the coefficients from the regressions) used to generate the final score.[19] This exercise evaluates whether schools can incorporate additional performance measures (value-added) to better inform personnel decisions.[20]

To conduct the analysis, I split the sample into three parts. First, I use OLS on 40% of the sample to impute missing data.[21] Second, I estimate the prediction model 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 models.

To impute subsequent summative ratings and value-added, I use previous summative ratings and value-added.[22] This imputation is critical because I evaluate the model by comparing the subsequent performance of teachers retained under the current system to that of teachers retained using the predictive models. Since I do not observe the performance of teachers who leave the profession, I encounter a one-sided problem for teachers who were dismissed and left the profession under the current system but would be retained using

[19] In Section 4.1, I find that machine learning techniques generate similar results.

[20] District leaders may have additional information about teachers that are unavailable in the administrative data. However, summative ratings should incorporate this information, as they are comprehensive measures of teacher performance. For example, these rubrics include scores for professionalism that capture daily interactions between teachers and administrators. Administrators must carefully document these interactions in the summative ratings, as these ratings are subject to review in dismissal hearings. In addition, Section A.3 suggests that characteristics that are unobservable in the data but observable to school administrators do not bias predicted performance.

[21] I impute subsequent summative ratings using previous summative ratings and the average of nonmissing value-added. Similarly, I impute subsequent math (ELA) value-added using previous math (ELA) value-added and summative ratings. I assume dismissed teachers would continue to teach the same subject in the future when imputing value-added.

[22] I do not use demographics, so I avoid introducing gender and racial biases into the imputation.

other prediction models.[23] To address this concern, I must impute the subsequent performance of teachers who left the profession. Districts retain some discretion when dismissing low-performing teachers, so, in Appendix Section A.3, I compare districts with higher dismissal rates conditional on summative ratings to districts with lower dismissal rates. If high-dismissal districts select on unobserved characteristics, imputations relying on these teachers would overpredict performance in low-dismissal districts. However, this test shows no evidence of prediction bias.

Next, I estimate several prediction models using both value-added and ratings. First, I calculate mean summative ratings in the first three years, which closely reflects the current system in New Jersey. As discussed in Section 2, performance-based personnel decisions solely rely on ratings.[24] In fact, Figure 2 shows mean pretenure ratings in the holdout sample are positively correlated with retention rates. Next, 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 and summative ratings. However, these prediction models are limited by a multidimensionality problem to identify the optimal combination of value-added and ratings that maximizes the state's utility function. Practically, since regressions only permit one outcome variable and value-added is weakly correlated with summative ratings, the models will struggle to generate improvements along all dimensions of performance simultaneously. Instead, the models will maximize the outcome variable and have little effect on the other dimensions. I can reduce this multidimensional problem 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-

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

[24] In New Jersey, highly-rated pretenure teachers may still be dismissed without cause. However, I focus my analysis on simulated performance-related dismissals by imputing performance for all teachers and only removing those with the lowest average pretenure summative ratings. If I used observed turnover instead of this performance-based metric, the baseline estimate would include both performance- and non-performance-related turnover.

missing standardized math and ELA value-added. I then construct a weighted average of the combined value-added measure and standardized summative rating. The weights depend on the relative importance of value-added and ratings in the state's utility function. Using OLS regressions, I predict subsequent composite values using previous composite measures. Each of these models relies on three years of pretenure data to predict subsequent performance.[25]

Next, I compare the predicted performance from the model to the actual performance in the holdout sample. Panels A, C, and E of Figure 3 plot actual performance against predicted performance. The models accurately predict performance with mean squared errors of less than 0.028 student test score 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 3 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 prediction models by simulating personnel decisions. Specifically, I rank teachers based on their predicted performance generated from each model. I simulate personnel decisions by removing the bottom $p \in \{1, 2, \ldots, 70\}$ percentile of teachers using each ranking system.[26] 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%. I cannot distinguish between voluntary and involuntary turnover in the data, so I assume 10% is a reasonable dismissal rate.[27] As a comparison model, I rank teachers based on their mean pretenure summative ratings. Since I cannot distinguish between voluntary and involuntary turnover, simulating dismissals based on mean pretenure ratings focuses the analysis on involuntary turnover of low-performing teachers. Specifically, I generate the control group

[25] I use a three-year pretenure period because 32 states use this length (Thomsen, 2020).

[26] I stop at the 70^{th} percentile because the samples become small.

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

where I remove the lowest ranked teachers based on mean pretenure ratings (rather than using the actual observed turnover) and the treated group where I remove the lowest ranked teachers based on a given model. Since I limit turnover in both control and treated groups to involuntary dismissals, voluntary turnover will not bias my results.

Panel A of Figure 4 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 makes few value-added gains (solid black and dashed red). The first row of Table 1 shows summative ratings rise by 0.0343 points when dismissing the bottom 10% of teachers[28] based on pretenure ratings relative to dismissing no teachers.[29] However, ELA value-added only increases by 0.0116 student test score standard deviations, while math value-added is unchanged. Ideally, schools would continue to generate these summative rating gains, while further improving subsequent value-added. Thus, I turn to OLS predictions.

First, I estimate OLS 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 4 show the results.[30] I demonstrate improvements along the outcome measure but rarely 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, Panel D shows ratings rise as dismissal rates increase when using a prediction

[28] I dismiss the lowest ranked teachers to approximately match statewide turnover rates and simplify the analysis. In practice, districts could adjust their own thresholds based on their teachers' performance.

[29] In all tables relying on performance as the dependent variable, the main effects are measured in student test score standard deviations. This can be interpreted as changes in student performance relative to the test. I also include a teacher-level standardized estimate of the effects in brackets by dividing the coefficient by the standard deviation of teacher performance in the sample. This measures value-added relative to all other teachers and allows me to estimate present value gains based on Chetty et al. (2014).

[30] 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.

model relying on ratings. However, value-added remains unchanged. I only generate mild rating gains in Panel C along with the increase in value-added when using a prediction model relying on ELA value-added.

Table 1 quantifies these changes using a 10% dismissal rate. The first row records the difference between no dismissals and 10% dismissal rates, while the remaining rows compare average retained teacher performance using the OLS models relative to the current mean summative rating dismissal policy. For example, dismissing the bottom 10% of teachers using models relying on value-added increases subsequent value-added by 0.0137–0.0265 student test score standard deviations, as seen in the second and third rows of Table 1. However, this policy causes summative ratings to decline by up to 0.0363 points relative to the current system. Similarly, the third row using ratings generates nearly no gains along any dimension relative to the current system.

Similar to Mihaly et al. (2013), I find that the models effectively predict the outcome variable but poorly predict the other measures. As discussed earlier, I encounter a multidimensionality problem that requires me to identify the optimal combination of value-added and ratings to maximize the state's utility function. 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 1, 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 models can simultaneously increase value-added and ratings in Figure 5. The third to last row of Table 1 using a composite measure with 50% weight on ratings increases subsequent value-added by 0.0129–0.0140 student test score standard deviations without resulting in a statistically significant decline in subsequent

ratings.[31] These improvements are all statistically significant at the 5% level.[32]

These results show districts can generate improvements in average teaching performance by using a weighted average of value-added and summative ratings.[33] These gains are feasible because the current system fails to perfectly sort teachers by subsequent ratings and ignores value-added. As a result, OLS models can generate improvements by more effectively ranking similarly rated teachers by subsequent value-added.

As seen in brackets, the average value-added gains are 0.045 teacher-level 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 $2,520 per student.[34] This value is nearly 12 times larger than the productivity effects of tenure (Ng, 2022). 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 significantly decline to generate these gains.

However, there is inherently a tradeoff between value-added and summative ratings. To depict this relationship, I estimated a series of composite models placing different weights on

[31] The results are not sensitive to sample variations across different specifications. Table A4 replicates Table 1 but only uses the 69 teachers in the holdout sample with math value-added, ELA value-added, and summative ratings. While the estimates are noisier and no longer statistically significant due to the smaller samples, the results remain similar in magnitude.

[32] The point estimates are similar when using all teachers rather than just novices in Table A5. 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).

[33] The composite models relied on fixed weights in the inputs and outputs. For example, the 30% ratings model used the 30% weight on both inputs and outputs. However, the results are similar when allowing the OLS model to flexibly assign weights to value-added and summative ratings. In fact, Table A6 estimates the change in performance using a flexible model where I regress each composite measure on three years of value-added and summative ratings. These results are quite similar to those found in Table 1. Since allowing the model to flexibly estimate the weights generates similar results and fixed weights are likely more palatable and easily explained to stakeholders, my main specification continues to use fixed weights.

[34] Chetty et al. (2014) estimates a 1 teacher-level 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 0.045 teacher-level standard deviation gain and the 8 grades for which I can calculate value-added.

value-added and summative ratings.[35] I then rely on the the following Cobb-Douglas utility function to identify optimal points:

$$Utility = MathVA^{\frac{1-x}{2}} * ELAVA^{\frac{1-x}{2}} * Ratings^{x} \tag{3}$$

In equation (3), utility is a function of subsequent math value-added, ELA value-added, and summative ratings. $X \in (0\%, 1\%, 2\%, ..., 100\%)$ records the relative weight on ratings. Figure 6 plots the optimal composite rating weight (y-axis) given the relative importance of ratings in the utility function. As expected, optimal composite measures place more weight on value-added when the utility function emphasizes value-added. Small sample sizes limit the variation in optimal points across adjacent weights resulting in many plateaus in the graph. For instance, the composite measure with 73% weight on ratings produces the greatest utility if there is at least 75% weight on ratings in the utility function. Figure 6 also shows that optimal points always rely on both ratings and value-added rather than just one or the other. For example, even if there is no utility weight on summative ratings, the optimal composite measure still puts 17% weight on summative ratings.

4.1 Information or Advanced Techniques?

In this section, I evaluate whether the improvements in subsequent teacher performance can be attributed to using additional performance measures (value-added) or more complicated algorithms. In Table 1, the linear regression using only ratings produces nearly identical results to the model using mean pretenure ratings. This suggests the gains may be a product of using more data rather than advanced techniques. To evaluate this hypothesis, I compare the OLS results to an analysis relying on machine learning algorithms. If techniques contribute to the gains, I expect the more advanced machine learning algorithms to outperform OLS. As described in Appendix Section A.4, I impute and train the data using random forests.

[35] I created 101 composite measure models ranging from 0% to 100% weight on summative ratings in increments of 1%.

Using random forests, Table 2 shows the baseline results comparing no dismissals to 10% dismissals using mean summative ratings (top row) and changes relative to the current system (remaining rows). The random forest estimates are very similar to, if not slightly worse than, the OLS results in Table 1. Thus, the gains are attributable to using more data.

In fact, 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.[36] 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 sizable gains. For policy relevance, I continue to use OLS for the remainder of the paper.

4.2 Changing Demographics in Response to Reformed Models

While I find that revised ranking systems can improve value-added without causing a statistically significant decline in summative ratings, it is also 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 A2 shows that only 18.2% of teachers are male, while 51.6% of students are male. Similarly, non-white teachers comprise only 13.5% of the teacher labor force, while 40.2% of students are non-white.[37]

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

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

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

(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 7 plots changes in demographics associated with each prediction model. In Panel A using the current ranking system, the fraction of male (dashed and dotted blue) and non-white (solid black) teachers steadily declines as dismissal rates increase.[38] This occurs because male and non-white teachers earn lower summative ratings. Figure 8 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 3 shows summary statistics of teacher performance by gender and race. In the final column, I find mean summative 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, prediction models estimated to maximize ratings simultaneously reduce these teachers' representation in the profession.[39]

Despite earning lower ratings, male and non-white teachers do not generate less value-added than their counterparts in Panels A–D of Figure 8. In fact, Panel B of Table 3 shows that non-white teachers' average value-added is 0.082–0.123 student test score standard deviations higher. Thus, the fraction of male teachers stays constant and the fraction of non-white teachers rises as dismissal rates increase for value-added models in Panels B and C of Figure 7. Table 4 uses 10% dismissal rates to show baseline demographics comparing no

[38] I find similar results using the model estimated using summative ratings. In Panel D, the fraction of male teachers also declines, though the fraction of non-white teachers remains steady.

[39] When conducting an analysis similar to Table 4 in Sartain and Steinberg (2020), I find that 23% of the racial gap and 81% of the gender gap in ratings persist when controlling for past performance, classroom characteristics, grade, and school (not shown). In addition, using the same dataset, Ng (2022) finds these rating disparities are larger when black and male teachers are evaluated by white and female principals, respectively. The persistent racial and gender gap along with increased out-of-group rating disparities suggest evaluation biases contribute to the lower ratings for male and Black teachers.

dismissals to 10% dismissals using mean summative ratings (top row) and changes relative to the current system (remaining rows). In the top row, the current system reduces the male teacher composition by 3.56 percentage points relative to no dismissals. This difference is statistically significant at the 1% level. I find a negative point estimate for non-white teachers but it is not statistically distinguishable from 0. The second and third rows of Table 4 show the models that incorporate value-added generate statistically significant 0.49 to 3.4 percentage point increases in the fraction of male and non-white teachers relative to the current system.

Focusing on the composite measures, models that place less weight on summative ratings increase male and non-white teacher composition. As a result, placing at least 50% weight on value-added generates statistically significant increases in male and non-white teacher composition ranging from 0.88 to 3.3 percentage points. All of these estimates are statistically significant at the 5% or 1% level. Panel E of Figure 7 using the composite measure with 50% weight on ratings corroborates this finding, as male teacher composition declines more slowly than in Panel A, while non-white teacher composition increases. While male and non-white teachers earn lower ratings, their similar or higher value-added stabilizes diversity in the composite models. Thus, the revised models improve diversity relative to the current system.[40]

5 Pretenure Period Length

In this section, I quantify the returns to longer pretenure periods by reestimating each model described in Section 4 using 1, 2, or 3 years of pretenure data.[41] 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

[40] Although male and non-white teachers earn lower summative ratings, Appendix Section A.5 finds no evidence that discrimination is biasing the OLS models.

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

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 5 estimates the improvement in average teacher quality generated by extending the pretenure period from 1 to 3 years with a 10% dismissal rate. Using the additional data produces positive point estimates, which is consistent with arguments to use multiple years of data when making decisions based on value-added estimates (Harris & Sass, 2014; Staiger & Rockoff, 2010; Rothstein, 2015). However, the gains are inconsistent with only a few statistically significant values.[42] When using summative ratings in the first and fourth entries, value-added remains unchanged, while summative ratings increase by 0.0168–0.0196 points. I find analogous results for value-added and the composite measures but the value-added gains are never statistically significant. While the estimates in Table 5 are similar in magnitude to those from using additional information in Table 1, the value-added gains are inconsistent when extending the pretenure period resulting in the lack of statistical significance. Also, unlike the extended pretenure period estimates, the revised prediction models from Section 4 do not reduce compensating differential through weakened job security.[43]

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 9 shows the improvements in average teacher performance generated when extending the pretenure period from 1 to 3 years. As dismissal rates rise for the composite measure in Panel E, the performance gains also increase for all three metrics.[44] Using a 50% dismissal rate, Table 6 shows large, statistically significant gains to extended pretenure periods. For example, the model using a composite measure with 50% weight on ratings increases average retained teacher math value-added,

[42] The gains are even weaker when extending the pretenure period from 1 to 2 years (Table A7) or from 2 to 3 years (Table A8).

[43] The revised models are not completely costless, as some teachers have left the teaching profession in response to an increased reliance on quantitative measures of performance (Perryman & Calvert, 2020). However, both extended pretenure periods and revised models rely on quantitative measures.

[44] Panels A–D of Figure 9 produce a similar pattern of results using the other models.

ELA value-added, and ratings by 0.0715 student test score standard deviations, 0.0403 student test score standard deviations, and 0.0666 points, respectively. These estimates are all statistically significant at the 5% or 1% levels. These gains are up to 0.06 student test score standard deviations or points larger than the analogous point estimates from a 10% dismissal rate in Table 5.

To depict a potential mechanism, I plot the kernel density of teacher performance in Figure 10. 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 student test score 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.[45] 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 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 student test score standard deviation,[46] while the 40^{th}–50^{th} percentiles span only 0.04 student test score standard deviation. 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.

[45] 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 student test score standard deviations of the mean.

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

Thus, extended pretenure periods only produce meaningful gains when dismissal rates are closer to 50%.[47]

These results align with several papers using structural models of teacher contracts (Rothstein, 2015; Staiger & Rockoff, 2010). In Rothstein (2015), 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%.[48]

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. However, I refrain from making a definitive policy prescription because administrators may use first-year performance results for professional development opportunities that may dramatically improve performance for some teachers. In addition, 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 long pretenure period with low dismissal rates is a suboptimal combination of policies.[49]

[47] While it is easy to identify low-performing teachers, the improvements across all measures of performance are much more consistent when using the composite measure. Therefore, it is still difficult to improve summative ratings or value-added without incorporating the other in the prediction model.

[48] In addition, Staiger and Rockoff (2010) find these improvements are limited if principals must retain low-performing teachers until the tenure receipt decision. In practice, teachers can be dismissed throughout the pretenure period. My analysis focuses on comparing the performance of retained teachers following tenure receipt, which avoids this comparison. However, policies that retained relatively low-performing teachers on the margin of tenure receipt would only increase the costs of extended pretenure periods.

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

6 Conclusion

Schools can simultaneously incorporate value-added and summative ratings to better inform personnel decisions. Utilizing predictive models, districts can increase subsequent average value-added by 0.01 student test score standard deviations, as well as the diversity of the teacher labor force without causing a statistically significant decline in ratings. The gains are a product of using additional information (value-added) rather than sophisticated methods, as the estimates are similar when using simple OLS or advanced machine learning techniques. These improvements are virtually costless to implement, as all the data are readily available.

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.[50] This finding suggests a more efficient policy would have either low dismissal rates with a short 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.

Based on these results, schools should flexibly incorporate value-added and summative ratings to inform personnel decisions. The models placing additional weight on value-added generate meaningful, statistically significant increases in subsequent value-added. Although this often simultaneously reduces subsequent ratings, several cases result in, at most, trivial declines. Similarly, the value-added returns to extending the pretenure period are smaller when using models that only rely on ratings. As discussed earlier, value-added has been linked to long-run student success (Chetty et al., 2014), whereas the returns to summative ratings remain unclear. While the tradeoffs between value-added and summative ratings are not known, these reforms would only worsen long-run student outcomes if, compared to

[50] However, longer pre-tenure periods allow districts to use early performance data to inform professional development. This may dramatically improve performance for some teachers.

value-added, summative ratings had a much *greater* positive impact on student outcomes. Future research could supplement this paper by estimating the precise impact of summative ratings on long-run student outcomes, similar to Chetty et al. (2014) for value-added.

At the same time, the current system relying solely on ratings reduces diversity due to increased male and non-white teacher dismissal rates. The reduced gender and racial representation may worsen male and non-white student outcomes. Adding value-added to these models reduces these disparities. Thus, flexibly incorporating value-added and summative ratings to inform personnel decisions would improve measures of teacher quality and teacher diversity.

This proposed reform is also practical to implement. Since OLS and machine learning generate similar gains, policymakers could just implement simple regressions to predict future performance based on past performance.[51] Machine learning techniques may prove useful if individual components of the summative ratings become available. The flexibility of machine learning techniques to account for non-linear relationships could yield even greater gains.

This study's findings maintain external validity because the Teacher Practice component of New Jersey's summative ratings rely on the same evaluation instruments as other states. For example, one of the approved rubrics, the Danielson Framework, has been used in 31 states ("Our Story", 2022).

Despite its external validity, there are four major limitations of this research. First, it only focuses on math and ELA teachers in grades with standardized tests. However, other subject-grades could increase the weight on alternative local test-based metrics of performances, such as SGOs. As discussed in Section 2, SGOs are designed by administrators and teachers to measure student growth and rely on the same standards as the state tests. If future research can show SGOs are informative, they may serve as an alternative option in a similar analysis for subject-grades without standardized tests. In this case, districts

[51] Individual districts may lack the expertise to calculate value-added and the predictive models but each state's Department of Education already has personnel that calculate mSGPs. These individuals would be familiar with value-added and OLS, so these calculations may be conducted at the state-level before being disseminated to individual districts.

could increase the weight on SGOs for grades without standardized tests to better identify high-quality teachers in these subjects. Otherwise, these policies would only be relevant for math and ELA teachers in tested grades.

Second, this research only has limited measures of non-cognitive skill development, which are critical to student outcomes (Jackson, 2018). Districts may measure non-cognitive skills using absences, suspensions, and grade repetition, as well as individual components of Teacher Practice, such as the classroom management category. These measures may be incorporated as both inputs and outputs, though further research would need to confirm the validity of these particular performance metrics. Future research incorporating additional measures of non-cognitive skills could further supplement this analysis.

Third, revised evaluation policies may generate distortions in performance and selection. Following a multitask principal-agent model (Holmstrom & Milgrom, 1991; Baker, 2002), teacher performance may be sensitive to the metrics used to evaluate the employee. Due to limited sample sizes, my analysis relies on some test scores that do not impact the mSGP component of the summative rating. In fact, 23% of the sample does not teach in subject-grades that produce mSGPs. Prior work has found that revised personnel decisions generate a behavioral response in performance (Dinerstein & Opper, 2022) and distortions in hiring practices (Ng, 2022). This analysis cannot plausibly estimate the impacts of revised evaluation policies on the reallocation of teacher effort across multiple dimensions of performance. However, policymakers could implement this policy and fine-tune the weights in response to observed distortions in behavior.

Fourth, I do not have disaggregated individual components of summative ratings. Since summative ratings include mSGPs and each state places different weights on test scores and evaluations, the precise recommended weights will vary across states. For instance, if another state's composite did not consider any test scores, the model for that state would likely find an optimal weight that places greater emphasis value-added compared to New Jersey's. In this case, the increased weight on value-added may be less appealing to state

leaders, especially since they had previously placed little weight on test scores. Nonetheless, the methodology from this paper could be easily applied to other states to identify unique optimal weights and provide a menu of improved performance-related dismissal policies.

Overall, I find that incorporating additional measures of teacher performance (value-added) 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 consistent improvements in average retained teacher performance.

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*.

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

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

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.

Dinerstein, M., & Opper, I. (2022). *Screening with multitasking* (Tech. Rep.). National Bureau of Economic Research.

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)* (Tech. Rep.). 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.

Grissom, J., & Bartanen, B. (2022). Potential race and gender biases in high-stakes teacher observations. *Journal of Policy Analysis and Management*, *41*(1), 131–161.

Guarino, C., Reckase, M., Stacy, B., & Wooldridge, J. (2014). A comparison of growth percentile and value-added models of teacher performance. Working paper# 39. *Education Policy Center at Michigan State University*.

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. (2006). Teacher quality. *Handbook of the Economics of Education*, *2*, 1051–1078.

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.

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

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 Conference*.

Kane, T., & Staiger, D. (2008). *Estimating teacher impacts on student achievement: An experimental evaluation* (Tech. Rep.). National Bureau of Economic Research.

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. (2019). Teacher effects on complex cognitive skills and social-emotional competencies. *Journal of Human Resources*, *54*(1), 1–36.

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.

Mihaly, K., McCaffrey, D., Staiger, D., & Lockwood, J. (2013). A composite estimator of effective teaching. *Seattle, WA: Bill & Melinda Gates Foundation*.

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`.

Neal, D. (2011). The design of performance pay in education. *Handbook of the Economics of Education*, *4*, 495–550.

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

Ng, K. (2022). *The effects of teacher tenure on productivity and selection* (Unpublished doctoral dissertation). Cornell University.

Our Story. (2022). *The Danielson Group*. Retrieved from `https://danielsongroup.org/our-story/`.

Perryman, J., & Calvert, G. (2020). What motivates people to teach, and why do they leave? Accountability, performativity and teacher retention. *British Journal of Educational Studies*, *68*(1), 3–23.

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

Resch, A., & Isenberg, E. (2018). How do test scores at the ceiling affect value-added estimates? *Statistics and Public Policy*, *5*(1), 1–6.

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).

Staiger, D., & Rockoff, J. (2010). Searching for effective teachers with imperfect information. *Journal of Economic perspectives*, *24*(3), 97–118.

State of New Jersey Department of Education. (2014). 2013-14 Educator evaluation implementation report.

State of New Jersey Department of Education. (2015). 2014-15 Educator evaluation implementation report.

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

Steinberg, M., & Kraft, M. (2017). The sensitivity of teacher performance ratings to the design of teacher evaluation systems. *Educational Researcher*, *46*(7), 378–396.

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.

Teacher Practice Evaluation Instruments. (2019). *New Jersey Department of Education*. Retrieved from https://www.state.nj.us/education/AchieveNJ/resources/rfq.shtml.

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/.

Walsh, E., Dotter, D., & Liu, A. (2018). *Can more teachers be covered? The accuracy, credibility, and precision of value-added estimates with proxy pre-tests* (Tech. Rep.). Mathematica Policy Research.

Walsh, E., & Isenberg, E. (2015). How does value added compare to student growth percentiles? *Statistics and Public Policy*, *2*(1), 1–13.

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: Difference in Performance using OLS

	Math VA	N	ELA VA	N	Ratings	N
Mean Ratings (Baseline)	0.0031	206	0.0116**	229	0.0343***	375
	(0.0065)		(0.0052)		(0.0052)	
	[0.0101]		[0.0404]		[0.1066]	
Math using Math	0.0265***	206	-0.0015	69	-0.0098	206
	(0.0064)		(0.0119)		(0.0073)	
	[0.0858]		[-0.0051]		[-0.0305]	
ELA using ELA	-0.0097	69	0.0137**	229	-0.0363***	229
	(0.0230)		(0.0064)		(0.0101)	
	[-0.0316]		[0.0477]		[-0.1128]	
Ratings using Ratings	-0.0074	206	-0.0048	229	0.0027	375
	(0.0058)		(0.0042)		(0.0035)	
	[-0.0241]		[-0.0168]		[0.0085]	
Composite using						
10% Ratings	0.0246***	206	0.0120**	229	-0.0165***	366
	(0.0071)		(0.0059)		(0.0061)	
	[0.0797]		[0.0420]		[-0.0513]	
30% Ratings	0.0200***	206	0.0140**	229	-0.0087*	366
	(0.0069)		(0.0057)		(0.0052)	
	[0.0647]		[0.0488]		[-0.0270]	
50% Ratings	0.0129**	206	0.0140***	229	-0.0017	366
	(0.0060)		(0.0053)		(0.0050)	
	[0.0420]		[0.0488]		[-0.0054]	
70% Ratings	0.0085*	206	0.0039	229	0.0068**	366
	(0.0046)		(0.0042)		(0.0033)	
	[0.0274]		[0.0137]		[0.0211]	
90% Ratings	0.0010	206	0.0022	229	0.0068**	366
	(0.0048)		(0.0039)		(0.0034)	
	[0.0031]		[0.0078]		[0.0210]	

Notes: This table estimates the change in performance generated when dismissing the bottom 10% of teachers using three years of data measured in student test score standard deviations or summative rating points. These models use OLS regressions defined in Section 4. The row headers define the model'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 comparison group changes in the remaining rows. These rows record changes relative to the *first row* using the models 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 (teacher-level standard deviations) in the dataset are included in brackets.

* 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)	0.0047	206	0.0102**	229	0.0307***	375
	(0.0064)		(0.0051)		(0.0051)	
	[0.0151]		[0.0357]		[0.0955]	
Math using Math	0.0220***	206	0.0037	69	-0.0132	206
	(0.0082)		(0.0153)		(0.0096)	
	[0.0712]		[0.0129]		[-0.0409]	
ELA using ELA	-0.0160	69	0.0081	229	-0.0287***	229
	(0.0195)		(0.0064)		(0.0097)	
	[-0.0518]		[0.0284]		[-0.0893]	
Ratings using Ratings	-0.0051	206	-0.0050	229	-0.0028	375
	(0.0056)		(0.0048)		(0.0050)	
	[-0.0165]		[-0.0176]		[-0.0086]	
Composite using						
10% Ratings	0.0144*	206	0.0127*	229	-0.0178**	366
	(0.0086)		(0.0066)		(0.0073)	
	[0.0467]		[0.0444]		[-0.0554]	
30% Ratings	0.0165**	206	0.0093*	229	-0.0088	366
	(0.0065)		(0.0055)		(0.0060)	
	[0.0536]		[0.0326]		[-0.0273]	
50% Ratings	0.0039	206	0.0061	229	-0.0084	366
	(0.0069)		(0.0058)		(0.0058)	
	[0.0125]		[0.0213]		[-0.0259]	
70% Ratings	0.0059	206	0.0013	229	0.0031	366
	(0.0052)		(0.0047)		(0.0047)	
	[0.0193]		[0.0045]		[0.0096]	
90% Ratings	0.0020	206	-0.0025	229	-0.0029	366
	(0.0054)		(0.0049)		(0.0052)	
	[0.0064]		[-0.0086]		[-0.0091]	

Notes: This table estimates the change in performance generated when dismissing the bottom 10% of teachers using three years of data measured in student test score standard deviations or summative rating points. These models use random forest algorithms defined in Section A.4. The row headers define the model'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 comparison group changes in the remaining rows. These rows record changes relative to the *first row* using the models 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 (teacher-level standard deviations) in the dataset are included in brackets.

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

Table 3: Summary Statistics by Gender and Race

Panel A: Gender

	Male	Female	Difference
Math VA	-0.009	-0.017	0.008
	(0.295)	(0.273)	(0.010)
ELA VA	-0.017	-0.016	-0.001
	(0.253)	(0.267)	(0.009)
Ratings	3.128	3.239	-0.111***
	(0.358)	(0.304)	(0.009)
Observations	1,747	7,856	

Panel B: Race

	Non-white	White	Difference
Math VA	0.091	-0.031	0.123***
	(0.276)	(0.275)	(0.011)
ELA VA	0.054	-0.028	0.082***
	(0.285)	(0.259)	(0.010)
Ratings	3.110	3.236	-0.126***
	(0.401)	(0.298)	(0.012)
Observations	1,312	8,291	

Notes: This table records mean performance by gender (Panel A) and race (Panel B) measured in student test score standard deviations or summative rating points. The row headers define the performance variable. The first column provides statistics for male and non-white teachers, while the second column provides statistics for female and white teachers. The standard deviations of each value are listed in parentheses below the means. 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 4: Difference in Demographics when Dismissing 10% of Teachers

	Male	N	Non-white	N
Mean Ratings (Baseline)	-0.0356***	375	-0.0019	375
	(0.0092)		(0.0060)	
Math using Math	0.0206*	206	0.0049**	206
	(0.0117)		(0.0023)	
ELA using ELA	0.0340***	229	0.0194***	229
	(0.0103)		(0.0050)	
Ratings using Ratings	0.0148*	375	0.0000	375
	(0.0086)		(0.0061)	
Composite using				
10% Ratings	0.0330***	366	0.0148***	366
	(0.0081)		(0.0030)	
30% Ratings	0.0239***	366	0.0118***	366
	(0.0085)		(0.0031)	
50% Ratings	0.0269***	366	0.0088**	366
	(0.0095)		(0.0041)	
70% Ratings	0.0117	366	0.0027	366
	(0.0090)		(0.0059)	
90% Ratings	0.0147	366	-0.0004	366
	(0.0090)		(0.0062)	

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 OLS models defined in Section 4. 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 comparison group changes in the remaining rows. These rows record changes relative to the *first row* using the models 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 5: Gains from Extending Pretenure from 1 to 3 Years: 10% Dismissal Rate

	Math VA	N	ELA VA	N	Ratings	N
Mean Ratings	0.0021	206	0.0070	229	0.0168***	375
	(0.0057)		(0.0057)		(0.0057)	
	[0.0068]		[0.0244]		[0.0522]	
Math using Math	0.0183	206	0.0185	69	0.0168	206
	(0.0136)		(0.0315)		(0.0183)	
	[0.0594]		[0.0645]		[0.0523]	
ELA using ELA	-0.0053	69	0.0084	229	-0.0040	229
	(0.0325)		(0.0130)		(0.0150)	
	[-0.0171]		[0.0294]		[-0.0125]	
Ratings using Ratings	-0.0053	206	0.0022	229	0.0196**	375
	(0.0132)		(0.0107)		(0.0079)	
	[-0.0172]		[0.0076]		[0.0608]	
Composite using						
10% Ratings	0.0133	206	0.0119	229	0.0135	366
	(0.0142)		(0.0127)		(0.0106)	
	[0.0433]		[0.0417]		[0.0420]	
30% Ratings	0.0198	206	0.0147	229	0.0153	366
	(0.0138)		(0.0125)		(0.0099)	
	[0.0641]		[0.0513]		[0.0475]	
50% Ratings	0.0123	206	0.0159	229	0.0165*	366
	(0.0138)		(0.0121)		(0.0095)	
	[0.0400]		[0.0555]		[0.0514]	
70% Ratings	0.0052	206	0.0088	229	0.0234**	366
	(0.0135)		(0.0117)		(0.0093)	
	[0.0168]		[0.0307]		[0.0727]	
90% Ratings	0.0020	206	0.0058	229	0.0241***	366
	(0.0132)		(0.0111)		(0.0088)	
	[0.0065]		[0.0202]		[0.0749]	

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 measured in student test score standard deviations or summative rating points. I use the OLS models defined in Section 4. 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 (teacher-level standard deviations) in the dataset are included in brackets.

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

Table 6: Gains from Extending Pretenure from 1 to 3 Years: 50% Dismissal Rate

	Math VA	N	ELA VA	N	Ratings	N
Mean Ratings	0.0036	206	0.0188	229	0.0373***	375
	(0.0164)		(0.0147)		(0.0135)	
	[0.0115]		[0.0658]		[0.1158]	
Math using Math	0.0551**	206	0.0339	69	0.0525**	206
	(0.0232)		(0.0397)		(0.0258)	
	[0.1787]		[0.1184]		[0.1630]	
ELA using ELA	0.0386	69	0.0360	229	0.0417	229
	(0.0430)		(0.0232)		(0.0255)	
	[0.1252]		[0.1257]		[0.1296]	
Ratings using Ratings	0.0280**	206	0.0238**	229	0.0542***	375
	(0.0132)		(0.0107)		(0.0079)	
	[0.0908]		[0.0830]		[0.1685]	
Composite using						
10% Ratings	0.0396*	206	0.0322	229	0.0453**	366
	(0.0233)		(0.0228)		(0.0188)	
	[0.1285]		[0.1125]		[0.1408]	
30% Ratings	0.0546**	206	0.0429**	229	0.0519***	366
	(0.0217)		(0.0218)		(0.0177)	
	[0.1770]		[0.1499]		[0.1612]	
50% Ratings	0.0715***	206	0.0403**	229	0.0666***	366
	(0.0201)		(0.0201)		(0.0152)	
	[0.2319]		[0.1408]		[0.2068]	
70% Ratings	0.0424**	206	0.0304	229	0.0689***	366
	(0.0172)		(0.0190)		(0.0136)	
	[0.1374]		[0.1060]		[0.2139]	
90% Ratings	0.0310**	206	0.0173	229	0.0539***	366
	(0.0150)		(0.0176)		(0.0116)	
	[0.1006]		[0.0604]		[0.1673]	

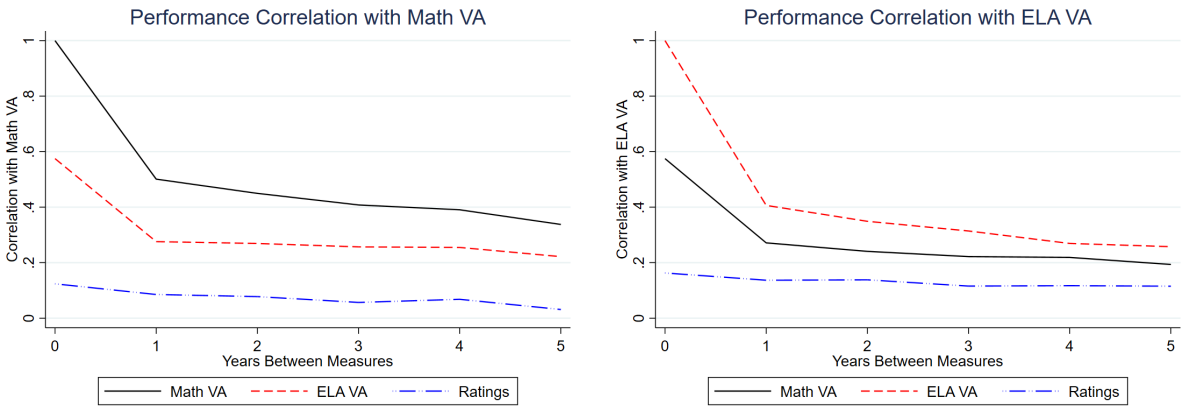
Notes: This table shows the change in performance generated when extending the pretenure period from 1 to 3 years and dismissing the bottom 50% of teachers measured in student test score standard deviations or summative rating points. I use the OLS models defined in Section 4. The row headers define 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 (teacher-level standard deviations) 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

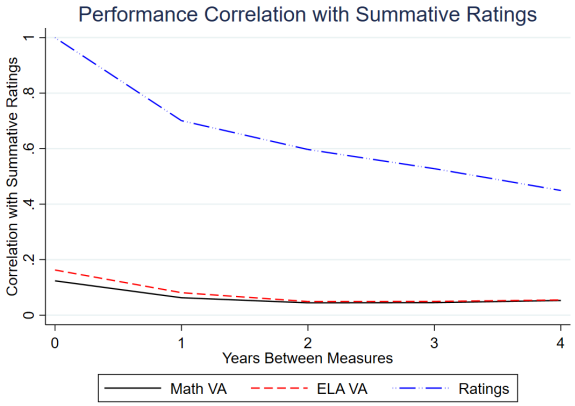


Figure 1: Performance Correlations

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.

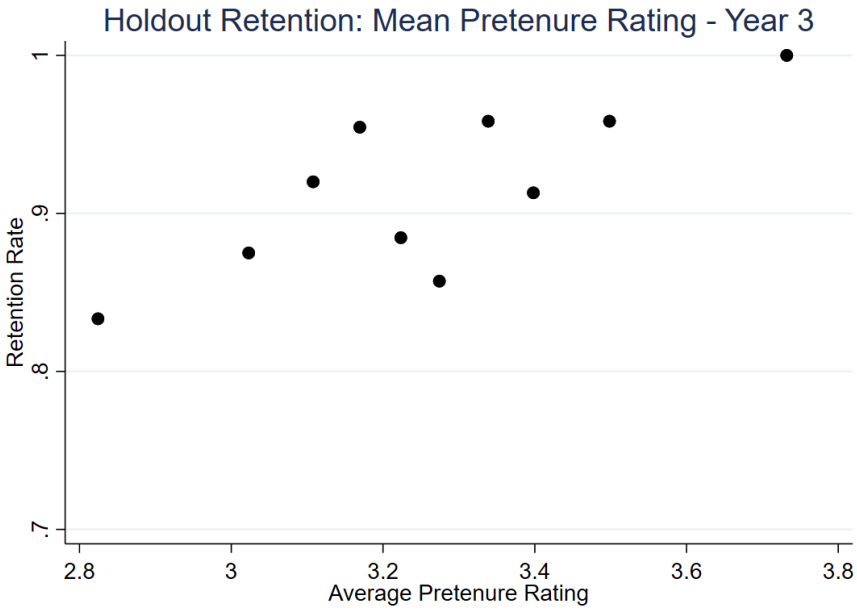
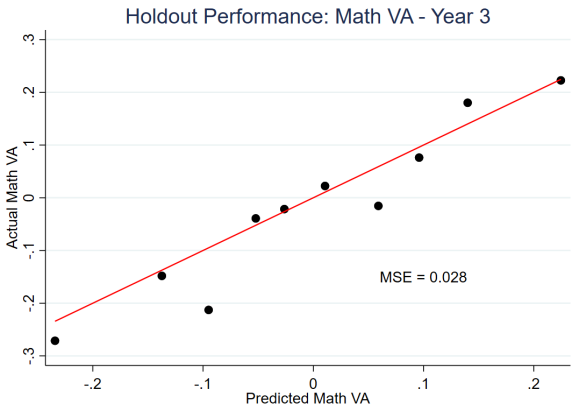


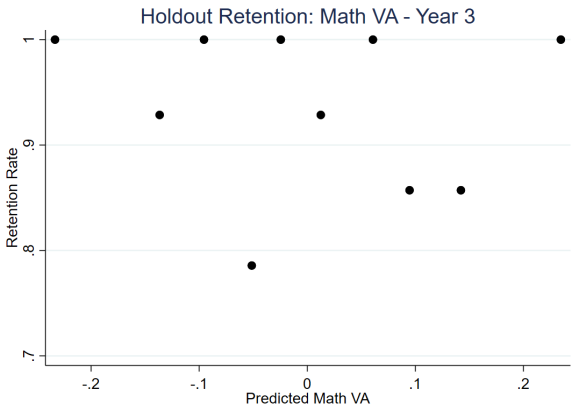
Figure 2: Retention and Mean Summative Ratings

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.

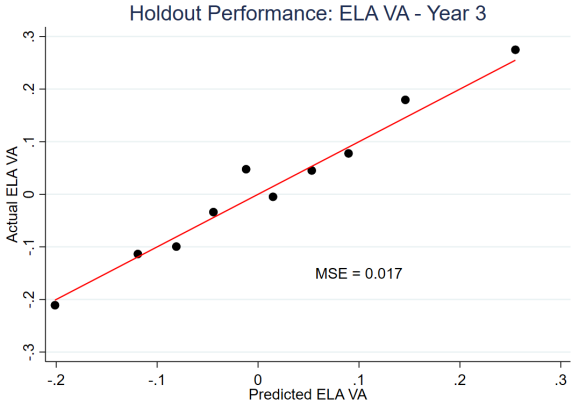
Panel A: Math Value-Added Performance



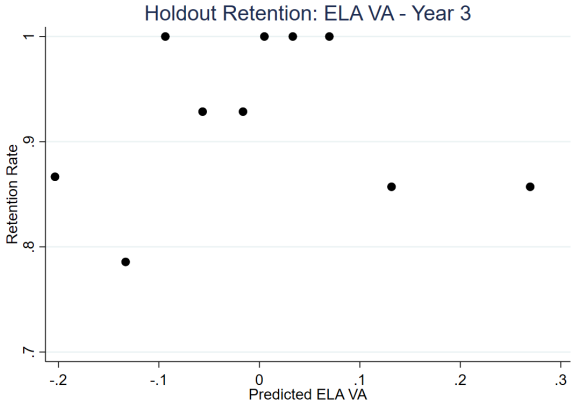
Panel B: Math Value-Added Retention



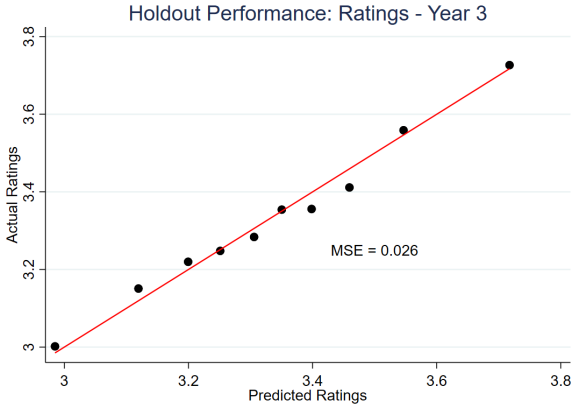
Panel C: ELA Value-Added Performance



Panel D: ELA Value-Added Retention



Panel E: Summative Ratings Performance



Panel F: Summative Ratings Retention

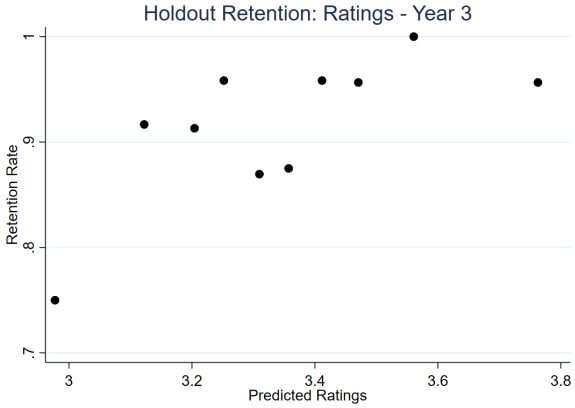
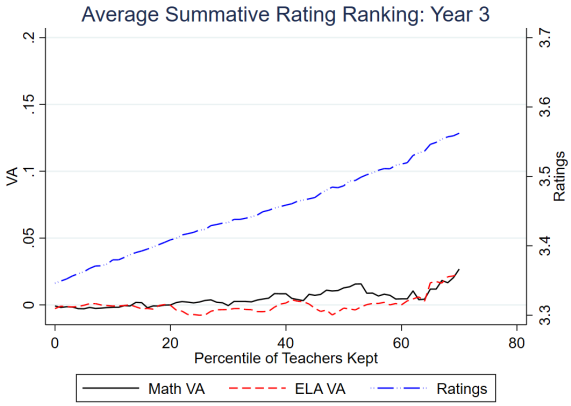


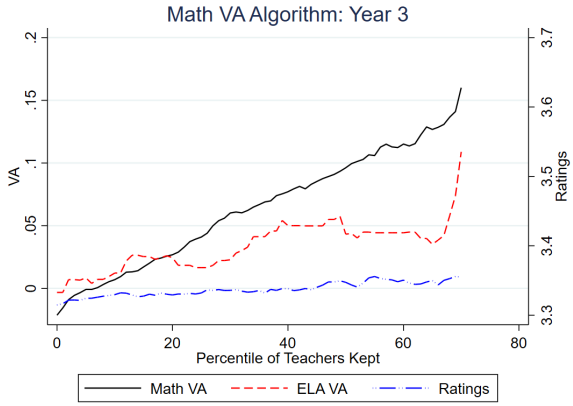
Figure 3: Actual Outcomes and Predicted Performance

Notes: This figure shows the relationship between actual outcomes and predicted subsequent performance in the holdout sample measured in student test score standard deviations or summative rating points. I use the OLS models defined in Section 4 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 mea 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.

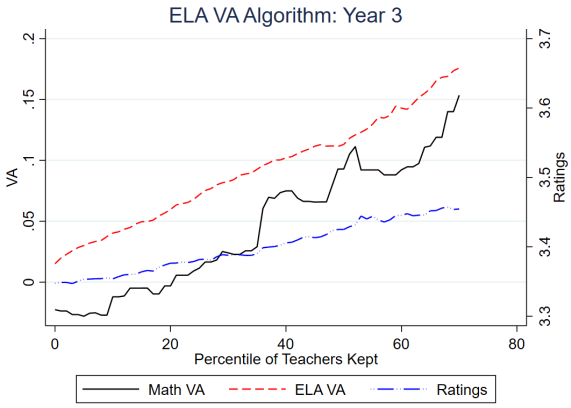
Panel A: Average Summative Ratings



Panel B: Math Value-Added Prediction Model



Panel C: ELA Value-Added Prediction Model Model



Panel D: Summative Ratings Prediction

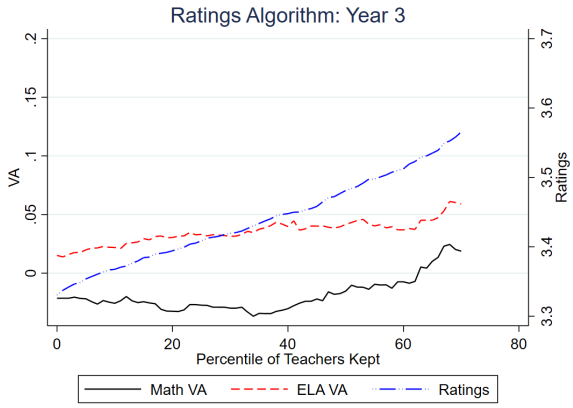
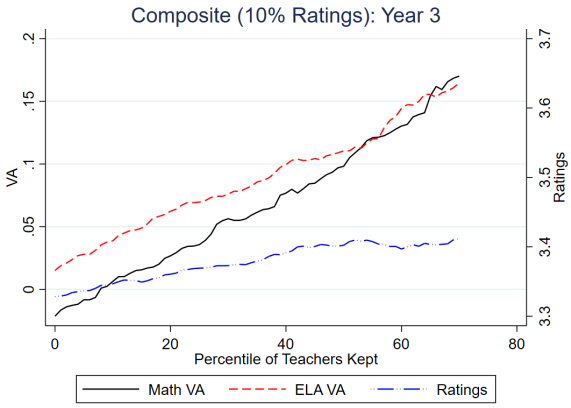


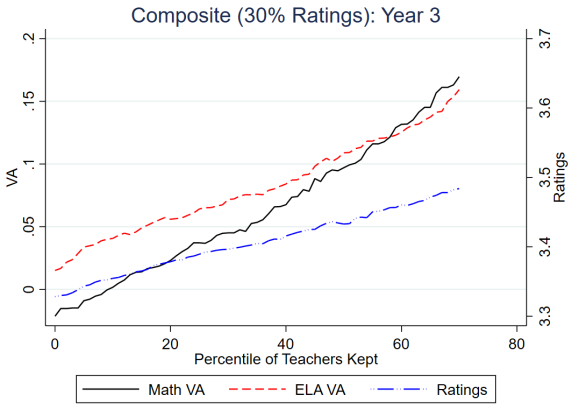
Figure 4: Mean Performance by Percentile

Notes: This figure plots the mean subsequent performance when changing minimum performance standards measured in student test score standard deviations or summative rating points. Panel A uses mean summative ratings in the teacher's first three years. Panels B–D use the OLS models defined in Section 4 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 student test score 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.

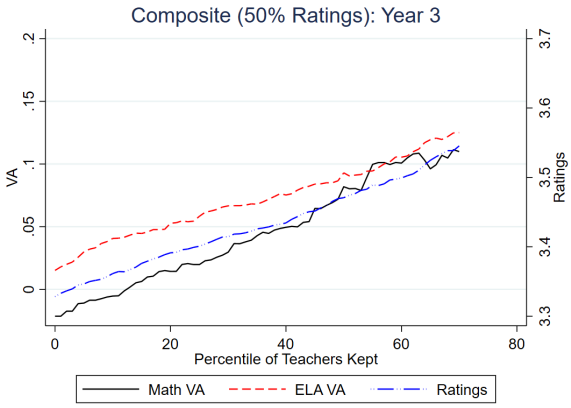
Panel A: 10% Summative Rating Weight



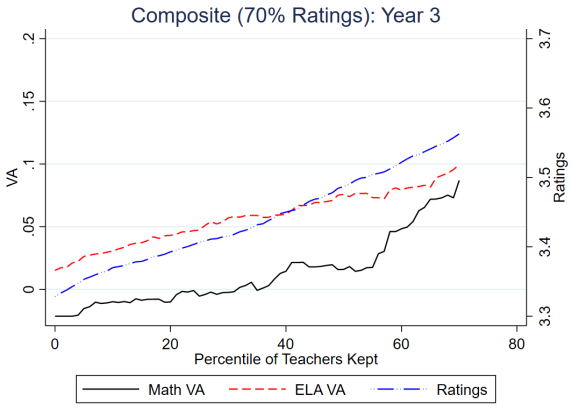
Panel B: 30% Summative Rating Weight



Panel C: 50% Summative Rating Weight



Panel D: 70% Summative Rating Weight



Panel E: 90% Summative Rating Weight

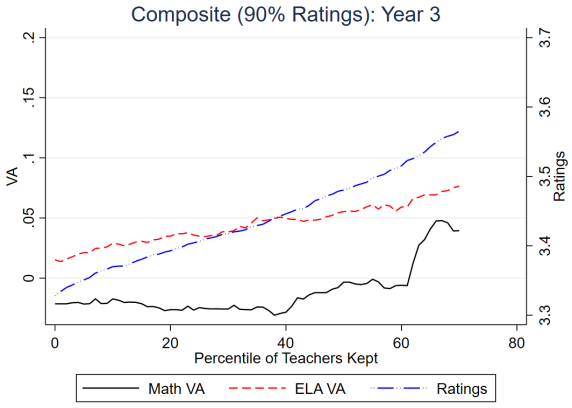


Figure 5: Mean Performance by Percentile using Composite Measure

Notes: This figure plots the mean subsequent performance when changing minimum performance standards measured in student test score standard deviations or summative rating points. I use the OLS models defined in Section 4 based on three years of data. To estimate the model, I use the composite measure defined in Section 4 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 student test score 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.

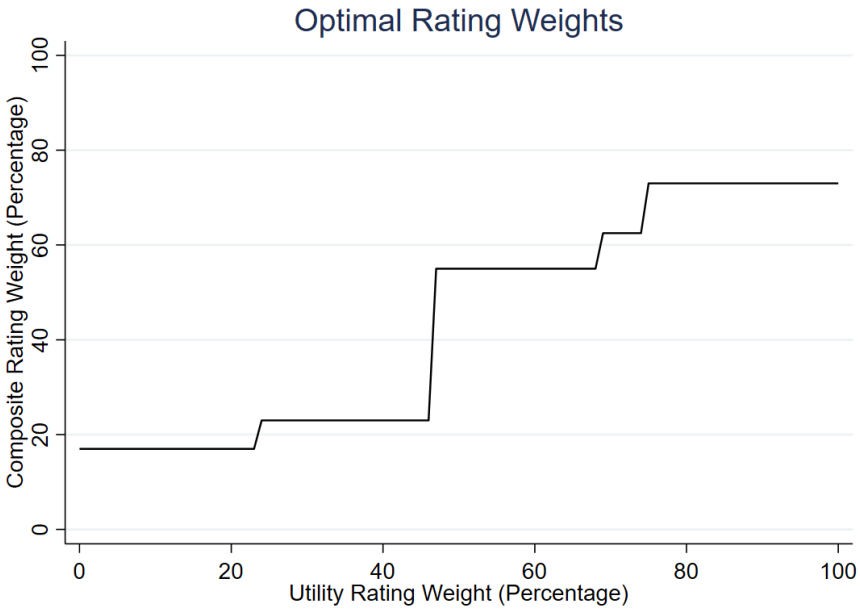
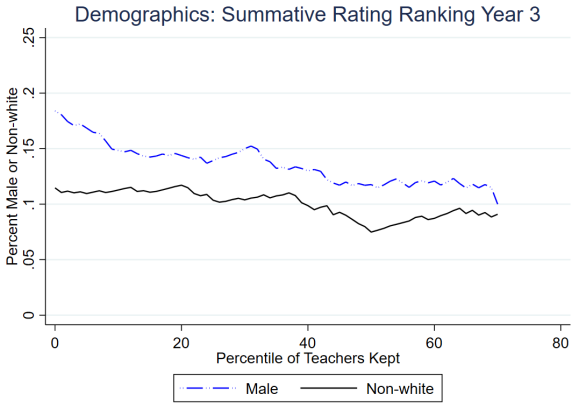


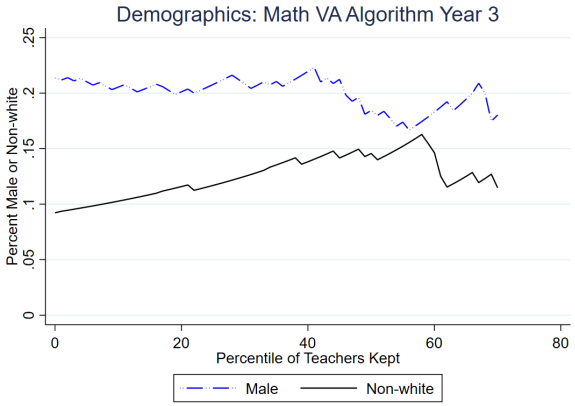
Figure 6: Optimal Summative Rating Weight Given Utility Weight

Notes: This figure plots the optimal summative rating weights on the composite measure given the summative rating weight in the utility function described in equation (3). The x-axis records the utility weight placed on summative ratings, while the y-axis records the optimal ratings weight on the composite measures based on OLS models defined in Section 4.

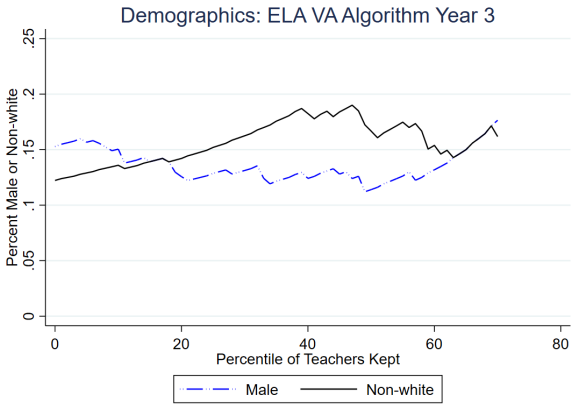
Panel A: Average Summative Ratings



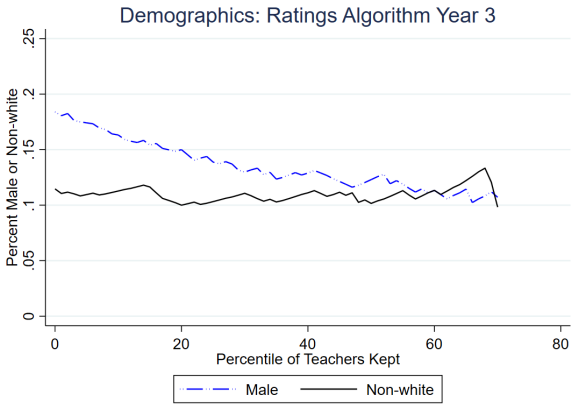
Panel B: Math Value-Added Prediction Model



Panel C: ELA Value-Added Prediction Model Model



Panel D: Summative Ratings Prediction Model



Panel E: 50% Summative Rating Weight Composite

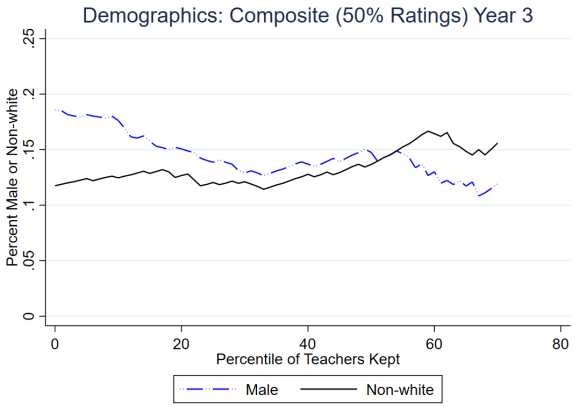
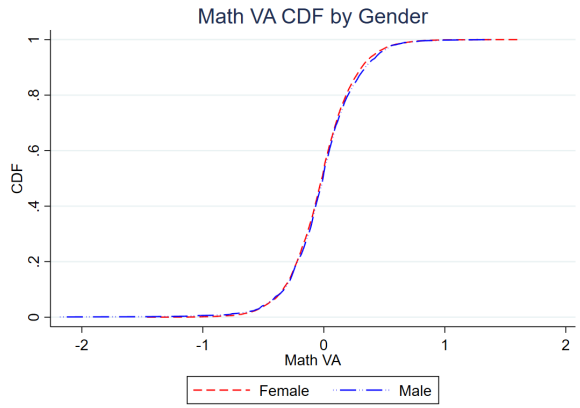


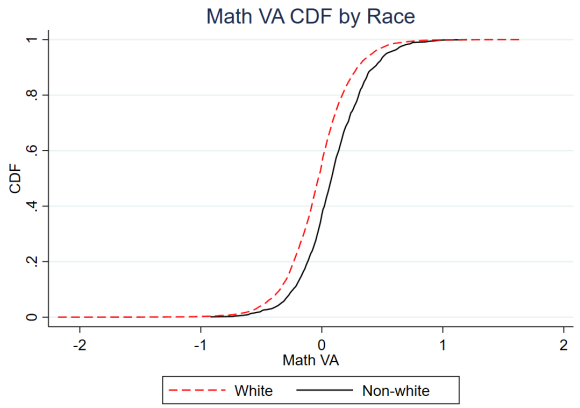
Figure 7: Mean Demographics by Percentile using ML

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 OLS models defined in Section 4 based on three years of data. Panel E uses the composite measure defined in Section 4 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.

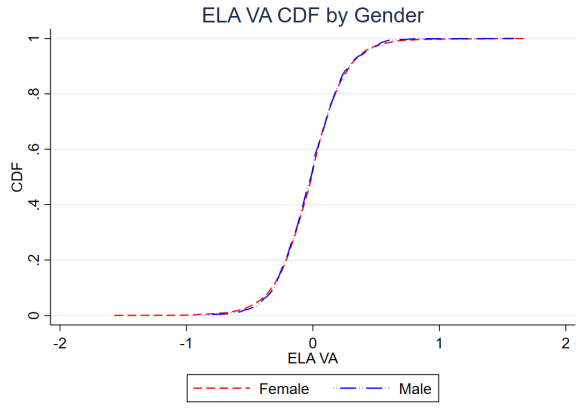
Panel A: Math Value-Added by Gender



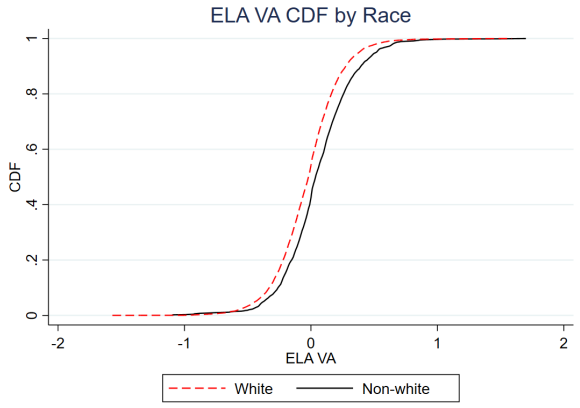
Panel B: Math Value-Added by Race



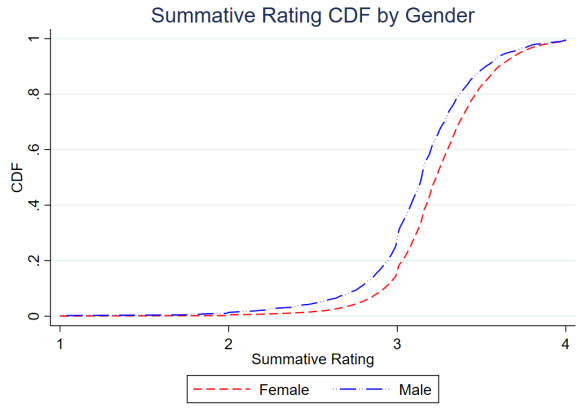
Panel C: ELA Value-Added by Gender



Panel D: ELA Value-Added by Race



Panel E: Summative Ratings by Gender



Panel F: Summative Ratings by Race

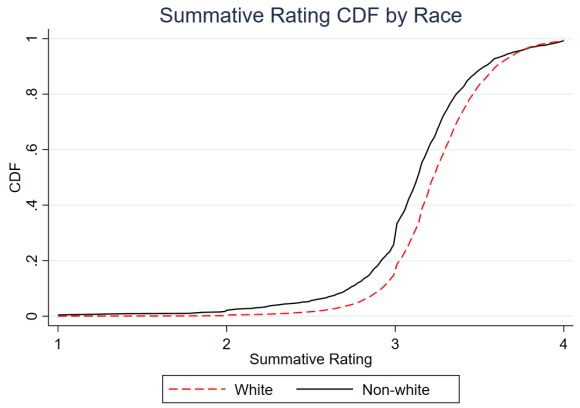
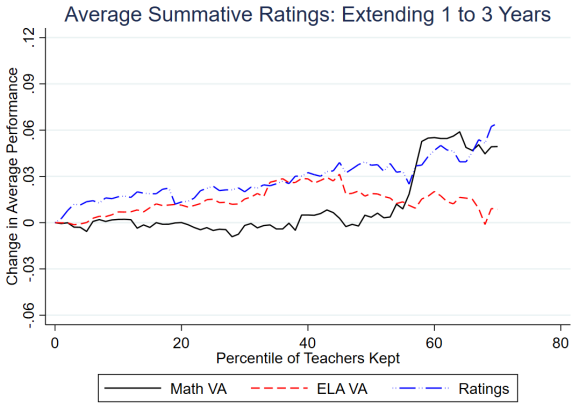


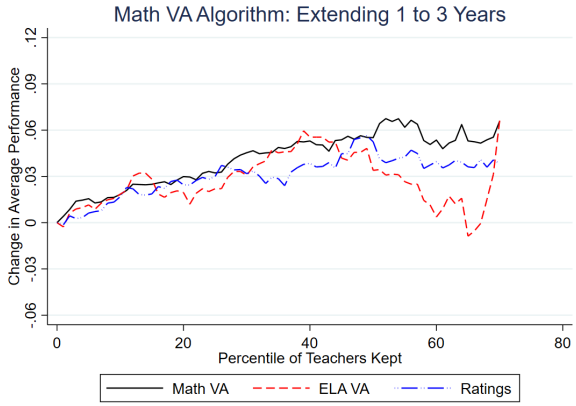
Figure 8: Performance CDF by Gender and Race

Notes: This figure shows the cumulative density of performance by gender (Panels A, C, and E) and race (Panels B, D, and F) measured in student test score standard deviations or summative rating points. 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 depict non-white teachers.

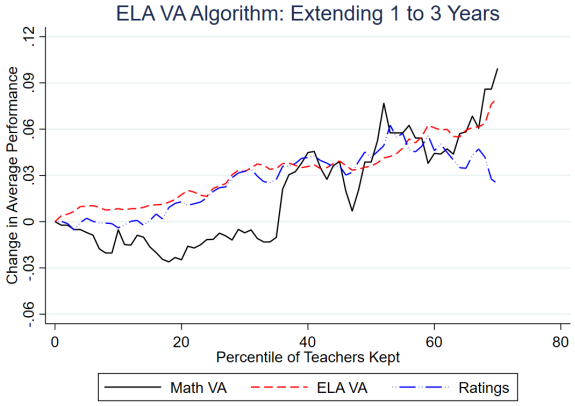
Panel A: Average Summative Ratings



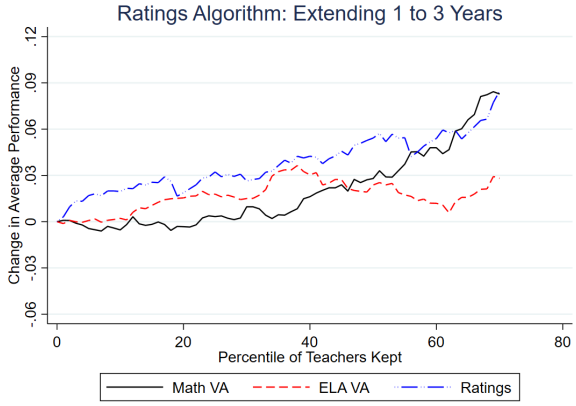
Panel B: Math Value-Added Prediction Model



Panel C: ELA Value-Added Prediction Model Model



Panel D: Summative Ratings Prediction



Panel E: 50% Summative Rating Weight Composite

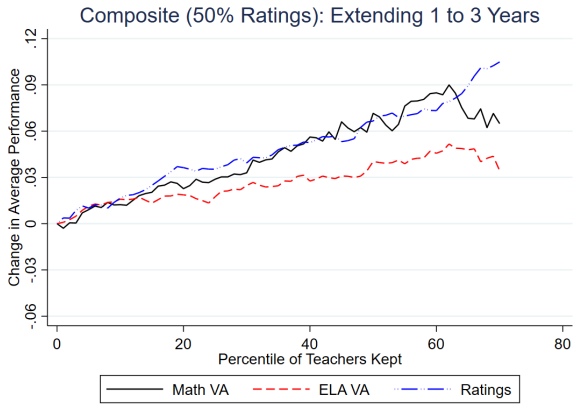
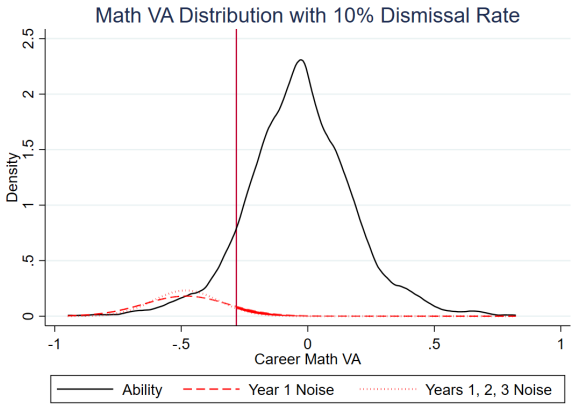


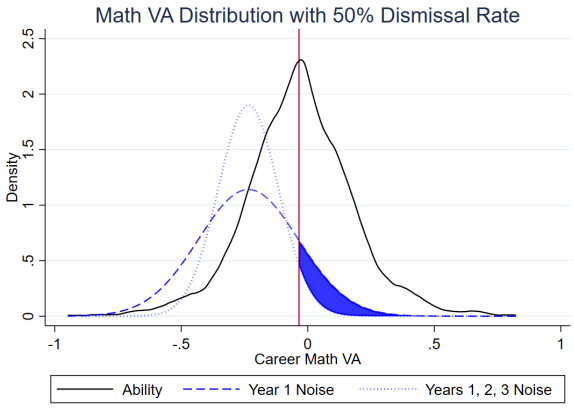
Figure 9: Extending Pretenure Period from 1 to 3 Years using Composite Ranking

Notes: This figure plots the change in average subsequent performance when using 3 years of data rather than 1 year of data measured in student test score standard deviations or summative rating points. Panel A uses mean summative ratings. Panels B–E use the OLS models defined in Section 4. Panel E uses the composite measure defined in Section 4 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.

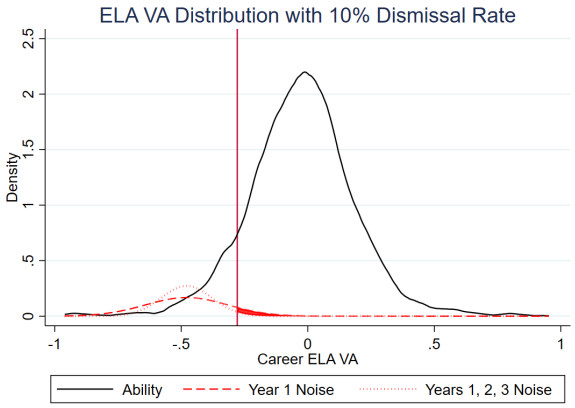
Panel A: Math Value-Added 10% Dismissal



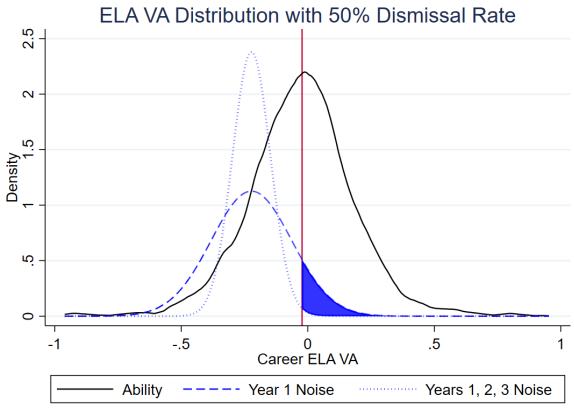
Panel B: Math Value-Added 50% Dismissal



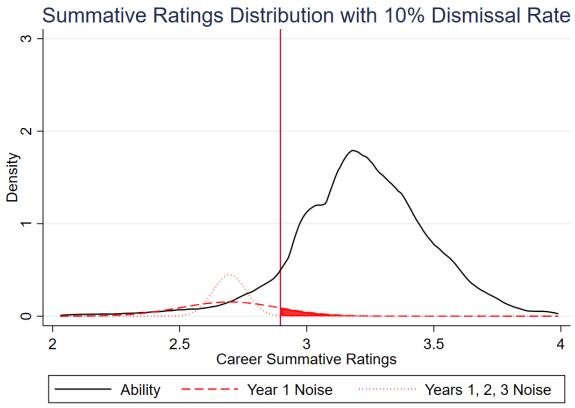
Panel C: ELA Value-Added 10% Dismissal



Panel D: ELA Value-Added 50% Dismissal



Panel E: Summative Ratings 10% Dismissal



Panel F: Summative Ratings 50% Dismissal

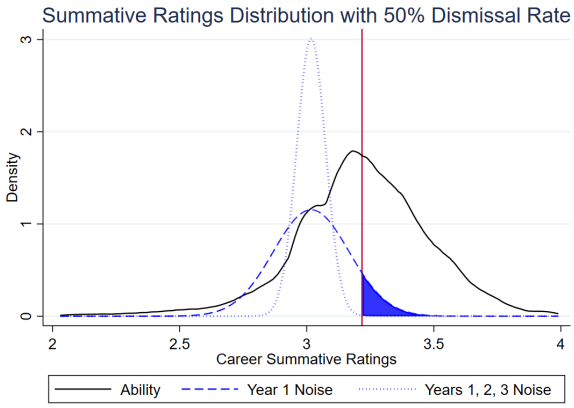


Figure 10: Performance Distribution and Noise

Notes: This figure plots performance kernel densities measured in student test score standard deviations or summative rating points. 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 student test score standard deviation 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 student test score 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 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.2 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 A9 shows 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 0.43, while the correlation between 2014 NJASK and 2015 PARCC math value-added is 0.42. 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.3 Imputing Missing Data

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 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:

$$E[y_j|x_j] = E[y_j|x_j, r_j = 1]. \tag{4}$$

To evaluate this assumption, I leverage district dismissal residuals. Districts retain some discretion when dismissing low-performing teachers, particularly for pretenured teachers with annual contracts. 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 (4) to hold, I must show that $E[g(s_j)|r_j = 1] = 0$. First, I partition the sample into low-dismissal districts that only rely on observed characteristics and high-

dismissal districts that also consider unobserved characteristics. For example, low-dismissal districts retain all teachers near the margin of ineffective teaching, while high-dismissal 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 low-dismissal districts, $r_j = 1$ if $F(x_j) > 0$ because they only rely on observed characteristics. In high-dismissal districts, $r_j = 1$ if $F(x_j) + G(s_j) > 0$ because they rely on both observed and unobserved characteristics.

I estimate a model on high-dismissal districts to estimate $E[y_j|x_j, r_j = 1] = f(x_j) + E[g(s_j)|r_j = 1]$. A prediction estimated on high-dismissal districts would incorporate any positive selection generated by these districts that select on unobserved characteristics. In comparison, the actual performance in low-dismissal districts provides an estimate of $E[y_j|x_j] = f(x_j)$ because these districts ignore unobserved characteristics.[52] By comparing predicted performance ($E[y_j|x_j, r_j = 1]$) to actual performance ($E[y_j|x_j]$) in the low-dismissal 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 high-dismissal and low-dismissal halves using the following regression:

$$r_{jt} = \beta x_{jt} + \delta_t + \varepsilon_{jt}. \tag{5}$$

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 dismissal residuals, 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 retention residuals 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 A2 plots the positive relationship between leave-one-out mean district retention residual and teacher retention. Thus, I parti-

[52] For low-dismissal districts, $E[g(s_j)|r_j = 1] = E[g(s_j)|F(x_j) > 0] = E[g(s_j)] = 0$.

tion the leave-one-out mean residuals into high-dismissal (below median) and low-dismissal (above median) halves.[53] Since high-dismissal districts retain fewer teachers conditional on ratings, I assume that these districts rely on both summative ratings (observed) and unobserved characteristics, while low-dismissal districts only rely on summative ratings. This partition also relies on a monotonicity assumption. I assume that any teacher retained in a high-dismissal district also would have been retained in a low-dismissal district.[54]

Ideally, I would focus on involuntary dismissals to identify high-dismissal and low-dismissal districts. Unfortunately, I cannot distinguish between voluntary and involuntary turnover.[55] Therefore, I may misclassify low-dismissal districts as high-dismissal 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 high-dismissal and low-dismissal districts in Table A10. Relative to high-dismissal districts (second column), low-dismissal 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 third column. This suggests that low-dismissal districts actually have the characteristics of hard-to-staff districts. With high voluntary attrition rates but low turnover rates conditional on summative ratings, the low-dismissal 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 high-dismissal and low-dismissal districts in equation (5).

After dividing the sample, I estimate an OLS model on the high-dismissal districts. I use the first three years of math value-added and summative ratings to predict subsequent math

[53] The results are similar when using different deciles to partition the sample into high-dismissal and low-dismissal districts (not shown).

[54] To primarily focus on pretenure dismissal residuals, this exercise relies on the same sample of novice teachers as the main analysis.

[55] 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.

value-added and repeat the process for ELA value-added. I then use the first three years of summative ratings and the first three years of average non-missing value-added to predict subsequent summative ratings.

Figure A3 plots the relationship between predicted and actual performance using prediction models estimated on high-dismissal districts and applied to low-dismissal 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 -0.011 to 0.022 student test score 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 (4) holds.[56] I do not find any evidence that districts use unobserved characteristics that impact subsequent performance to selectively retain teachers.[57]

A.4 Machine Learning Algorithm

In this section, I estimate the same models as described in Section 4 but use machine learning techniques rather than OLS. 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.

Specifically, I use random forests, which generate algorithms that sort teachers into bins of predicted performance. Although the coefficients lack causal interpretations, the algorithms account for nonlinear relationships to effectively predict outcomes (Mullainathan & Spiess,

[56] All results are robust to including district retention residuals as a predictor to impute performance (not shown).

[57] I also must account for this imputation when calculating standard errors. To do so, I bootstrap each sample prior to imputing missing data. This incorporates imputation error into the calculation of standard errors.

2017; Kleinberg et al., 2015).

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. I continue to use 40% of the sample to impute missing performance data, another 40% to estimate the algorithm, and the remaining 20% to conduct the analysis.

Table 2 shows the baseline results comparing no dismissals to 10% dismissals using mean summative ratings (top row) and changes relative to the current system (remaining rows). The random forest estimates are very similar to the OLS results in Table 1.

A.5 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, 2022; Chi, 2021; Grissom & Bartanen, 2022). 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 prediction models. In Table A11, I estimate the improvements in average teacher performance when using a 10% dismissal rate, three years of performance data, and demographics.[58] With nearly identical results to Table 1, Table A11 shows demographic data do not improve the model's prediction accuracy. For example, the composite measure with 50% weight on summative ratings shows ratings do not change, while value-added increases by 0.0129–0.0140 student test score standard deviations relative to the current system. Since these values are identical to those from Table 1, demographic data do not appear to affect the prediction models.

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

Next, I test for differences in residuals generated by the OLS models in Table A12. 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 model 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 model underpredicts male and non-white performance by 0.0151 and 0.0192 student test score standard deviations, respectively. However, the ELA value-added model overpredicts male and non-white performance by 0.0174 and 0.0186 student test score standard deviations, respectively.[59]

Although I find summative rating disparities by gender and race in Table 3, these tests do not provide any conclusive 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 OLS models, I cannot completely eliminate this possibility.

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

A.6 Appendix Tables

Table A1: 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, SGOs and mSGPs are acronyms for Student Growth Objectives and median Student Growth Percentiles, respectively.

Table A2: Summary Statistics

	Students	Teachers
Female	0.484	0.818
	(0.500)	(0.386)
Black	0.197	0.062
	(0.398)	(0.242)
Hispanic	0.271	0.078
	(0.445)	(0.268)
Non-white	0.402	0.135
	(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		2.901
		(1.436)
Years in District		2.734
		(1.452)
Summative Rating		3.268
		(0.312)
Obs	12405063	24012
Unique Obs	2164750	10329

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 A3: Sample Restrictions

	Math VA	ELA VA	Ratings
Teachers with Non-Missing Data	40,484	45,686	154,671
Has Both VA and Ratings	33,443	37,487	51,634
Has Year 1 Performance	2,865	3,162	4,714
Has Performance up to Year 3	1,051	1,160	1,904
Estimating Sample	428	487	785
Imputing Sample	417	444	744
Holdout Sample	206	229	375

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 record the number of observations in the estimating, imputing, and holdout samples. These samples represent approximately 40%, 40%, and 20% of the remaining sample, respectively.

Table A4: Difference in Performance using OLS

	Math VA	N	ELA VA	N	Ratings	N
Mean Ratings (Baseline)	0.0122	69	0.0147	69	0.0374***	69
	(0.0125)		(0.0126)		(0.0137)	
	[0.0394]		[0.0514]		[0.1162]	
Math using Math	0.0169	69	-0.0043	69	-0.0154	69
	(0.0125)		(0.0126)		(0.0137)	
	[0.0549]		[-0.0151]		[-0.0478]	
ELA using ELA	-0.0167	69	0.0029	69	-0.0422*	69
	(0.0253)		(0.0164)		(0.0217)	
	[-0.0540]		[0.0103]		[-0.1311]	
Ratings using Ratings	0.0022	69	-0.0010	69	-0.0057	69
	(0.0123)		(0.0118)		(0.0121)	
	[0.0072]		[-0.0033]		[-0.0176]	
Composite using						
10% Ratings	0.0153	69	0.0044	69	-0.0138	69
	(0.0146)		(0.0119)		(0.0160)	
	[0.0495]		[0.0153]		[-0.0429]	
30% Ratings	0.0200	69	0.0113	69	-0.0098	69
	(0.0129)		(0.0108)		(0.0124)	
	[0.0649]		[0.0396]		[-0.0304]	
50% Ratings	0.0159	69	0.0109	69	-0.0017	69
	(0.0125)		(0.0106)		(0.0117)	
	[0.0515]		[0.0381]		[-0.0054]	
70% Ratings	0.0167	69	0.0149	69	-0.0017	69
	(0.0111)		(0.0095)		(0.0114)	
	[0.0541]		[0.0522]		[-0.0051]	
90% Ratings	0.0010	69	-0.0008	69	-0.0079	69
	(0.0120)		(0.0114)		(0.0123)	
	[0.0033]		[-0.0028]		[-0.0246]	

Notes: This table estimates the change in performance generated when dismissing the bottom 10% of teachers using three years of data measured in student test score standard deviations or summative rating points. These models use OLS regressions defined in Section 4. I restrict the sample to only teachers with non-missing math value-added, ELA value-added, and summative ratings to keep samples constant across rows and columns. The row headers define the model'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 comparison group changes in the remaining rows. These rows record changes relative to the *first row* using the models 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 (teacher-level standard deviations) in the dataset are included in brackets.

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

Table A5: Difference in Performance using All Teachers

	Math VA	N	ELA VA	N	Ratings	N
Mean Ratings (Baseline)	0.0060***	3,367	0.0036***	3,791	0.0395***	5,967
	(0.0012)		(0.0010)		(0.0012)	
	[0.0193]		[0.0125]		[0.1229]	
Math using Math	0.0187***	3,367	0.0128***	1,304	-0.0322***	3,367
	(0.0015)		(0.0029)		(0.0022)	
	[0.0607]		[0.0445]		[-0.1001]	
ELA using ELA	0.0121***	1,304	0.0193***	3,791	-0.0271***	3,791
	(0.0028)		(0.0014)		(0.0022)	
	[0.0394]		[0.0673]		[-0.0841]	
Ratings using Ratings	0.0007	3,367	-0.0005	3,791	0.0021***	5,967
	(0.0009)		(0.0008)		(0.0007)	
	[0.0023]		[-0.0017]		[0.0066]	
Composite using						
10% Ratings	0.0174***	3,367	0.0190***	3,791	-0.0258***	5,854
	(0.0017)		(0.0015)		(0.0017)	
	[0.0565]		[0.0663]		[-0.0800]	
30% Ratings	0.0163***	3,367	0.0158***	3,791	-0.0134***	5,854
	(0.0014)		(0.0013)		(0.0014)	
	[0.0527]		[0.0552]		[-0.0417]	
50% Ratings	0.0115***	3,367	0.0114***	3,791	-0.0037***	5,854
	(0.0012)		(0.0012)		(0.0011)	
	[0.0372]		[0.0400]		[-0.0114]	
70% Ratings	0.0081***	3,367	0.0064***	3,791	0.0003	5,854
	(0.0010)		(0.0010)		(0.0009)	
	[0.0262]		[0.0225]		[0.0010]	
90% Ratings	0.0035***	3,367	0.0013	3,791	0.0017**	5,854
	(0.0009)		(0.0008)		(0.0007)	
	[0.0113]		[0.0045]		[0.0053]	

Notes: This table shows the change in performance generated when dismissing the bottom 10% of teachers using OLS models defined in Section 4 and three years of data measured in student test score standard deviations or summative rating points. 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 comparison group changes in the remaining rows. These rows record changes relative to the *first row* using the models 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 (teacher-level standard deviations) in the dataset are included in brackets.

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

Table A6: Difference in Composite Performance using OLS with Flexible Inputs

	Math VA	N	ELA VA	N	Ratings	N
Composite using						
10% Ratings	0.0236***	206	0.0119*	229	-0.0159**	366
	(0.0082)		(0.0064)		(0.0077)	
	[0.0766]		[0.0417]		[-0.0495]	
30% Ratings	0.0192***	206	0.0143**	229	-0.0090	366
	(0.0072)		(0.0056)		(0.0059)	
	[0.0621]		[0.0500]		[-0.0279]	
50% Ratings	0.0153**	206	0.0115**	229	-0.0033	366
	(0.0069)		(0.0054)		(0.0054)	
	[0.0496]		[0.0403]		[-0.0103]	
70% Ratings	0.0132**	206	0.0095*	229	0.0018	366
	(0.0059)		(0.0053)		(0.0046)	
	[0.0428]		[0.0333]		[0.0056]	
90% Ratings	0.0080	206	0.0042	229	0.0071**	366
	(0.0050)		(0.0045)		(0.0034)	
	[0.0259]		[0.0147]		[0.0221]	

Notes: This table estimates the change in performance generated when dismissing the bottom 10% of teachers using three years of data measured in student test score standard deviations or summative rating points. These models use OLS regressions defined in Section 4, but regress the composite measure in each row on summative ratings and average value-added in each of the first three years. The values record the change in performance when dismissing the bottom 10% of teachers using the given model, compared to the average performance when dismissing the bottom 10% of teachers using mean summative ratings. 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 (teacher-level standard deviations) 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: 10% Dismissal Rate

	Math VA	N	ELA VA	N	Ratings	N
Mean Ratings	0.0066*	324	0.0082**	358	0.0054	567
	(0.0037)		(0.0039)		(0.0047)	
	[0.0214]		[0.0288]		[0.0167]	
Math using Math	0.0012	324	0.0034	124	-0.0035	324
	(0.0106)		(0.0255)		(0.0142)	
	[0.0039]		[0.0117]		[-0.0108]	
ELA using ELA	-0.0076	124	0.0016	358	-0.0004	358
	(0.0245)		(0.0110)		(0.0137)	
	[-0.0247]		[0.0057]		[-0.0013]	
Ratings using Ratings	-0.0003	324	0.0060	358	0.0060	567
	(0.0090)		(0.0092)		(0.0055)	
	[-0.0010]		[0.0208]		[0.0186]	
Composite using						
10% Ratings	0.0033	324	0.0092	358	0.0016	558
	(0.0102)		(0.0106)		(0.0085)	
	[0.0107]		[0.0321]		[0.0050]	
30% Ratings	0.0125	324	0.0065	358	0.0021	558
	(0.0096)		(0.0102)		(0.0082)	
	[0.0405]		[0.0225]		[0.0065]	
50% Ratings	0.0062	324	0.0055	358	0.0041	558
	(0.0095)		(0.0101)		(0.0075)	
	[0.0200]		[0.0193]		[0.0128]	
70% Ratings	0.0024	324	0.0132	358	0.0010	558
	(0.0094)		(0.0093)		(0.0069)	
	[0.0079]		[0.0462]		[0.0031]	
90% Ratings	-0.0006	324	0.0053	358	0.0066	558
	(0.0092)		(0.0092)		(0.0063)	
	[-0.0020]		[0.0184]		[0.0204]	

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 measured in student test score standard deviations or summative rating points. I use the OLS models defined in Section 4. 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 (teacher-level standard deviations) 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: 10% Dismissal Rate

	Math VA	N	ELA VA	N	Ratings	N
Mean Ratings	-0.0047	206	-0.0012	229	0.0072	375
	(0.0048)		(0.0046)		(0.0061)	
	[-0.0154]		[-0.0042]		[0.0223]	
Math using Math	0.0132	206	0.0104	69	0.0129	206
	(0.0143)		(0.0335)		(0.0177)	
	[0.0428]		[0.0363]		[0.0401]	
ELA using ELA	0.0111	69	0.0099	229	0.0055	229
	(0.0337)		(0.0124)		(0.0191)	
	[0.0358]		[0.0344]		[0.0170]	
Ratings using Ratings	-0.0043	206	-0.0027	229	0.0108	375
	(0.0131)		(0.0111)		(0.0088)	
	[-0.0140]		[-0.0095]		[0.0335]	
Composite using						
10% Ratings	0.0099	206	0.0035	229	0.0146	366
	(0.0140)		(0.0124)		(0.0115)	
	[0.0322]		[0.0123]		[0.0453]	
30% Ratings	0.0063	206	0.0088	229	0.0165	366
	(0.0139)		(0.0117)		(0.0113)	
	[0.0203]		[0.0308]		[0.0512]	
50% Ratings	0.0025	206	0.0074	229	0.0147	366
	(0.0136)		(0.0115)		(0.0110)	
	[0.0081]		[0.0258]		[0.0458]	
70% Ratings	0.0088	206	-0.0029	229	0.0194*	366
	(0.0131)		(0.0113)		(0.0103)	
	[0.0287]		[-0.0101]		[0.0602]	
90% Ratings	0.0058	206	0.0044	229	0.0148	366
	(0.0135)		(0.0111)		(0.0102)	
	[0.0188]		[0.0152]		[0.0461]	

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 measured in student test score standard deviations or summative rating points. I use the OLS models defined in Section 4. 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 (teacher-level standard deviations) in the dataset are included in brackets.

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

Table A9: Annual VA Correlation by Year

Panel A: Math Value-Added

	2013	2014	2015	2016	2017	2018
NJASK:2013	1.00					
NJASK:2014	0.46	1.00				
PARCC:2015	0.36	0.42	1.00			
PARCC:2016	0.33	0.41	0.43	1.00		
PARCC:2017	0.34	0.39	0.39	0.47	1.00	
PARCC:2018	0.30	0.37	0.38	0.46	0.50	1.00

Panel B: ELA Value-Added

	2013	2014	2015	2016	2017	2018
NJASK:2013	1.00					
NJASK:2014	0.37	1.00				
PARCC:2015	0.25	0.25	1.00			
PARCC:2016	0.26	0.28	0.37	1.00		
PARCC:2017	0.23	0.25	0.32	0.41	1.00	
PARCC:2018	0.23	0.25	0.33	0.39	0.43	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.

Table A10: Summary Statistics by District Dismissal Residual

	Low-Dismissal	High-Dismissal	Difference
FRPL	0.452	0.296	0.156***
	(0.297)	(0.256)	(0.006)
ELL	0.079	0.046	0.033***
	(0.079)	(0.064)	(0.002)
Math Proficient	0.476	0.565	-0.088***
	(0.164)	(0.159)	(0.004)
ELA Proficient	0.521	0.622	-0.101***
	(0.175)	(0.155)	(0.004)
Black	0.196	0.160	0.036***
	(0.187)	(0.208)	(0.005)
Hispanic	0.339	0.220	0.119***
	(0.284)	(0.212)	(0.006)
Observations	3,826	3,826	

Notes: This table provides summary statistics for low-dismissal and high-dismissal districts calculated at the district level. I define low-dismissal (high-dismissal) districts as those with above (below) median leave-one-out average residuals from equation (5). The row headers define the variable. The first column provides statistics for low-dismissal districts, while the second column provides statistics for high-dismissal districts. The standard deviations of each value are listed in parentheses below the means. 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 A11: Difference in Performance with Demographics: 10% Dismissal Rate

	Math VA	N	ELA VA	N	Ratings	N
Mean Ratings (Baseline)	0.0031	206	0.0116**	229	0.0343***	375
	(0.0065)		(0.0052)		(0.0052)	
	[0.0101]		[0.0404]		[0.1066]	
Math using Math	0.0265***	206	-0.0015	69	-0.0098	206
	(0.0064)		(0.0115)		(0.0074)	
	[0.0858]		[-0.0051]		[-0.0305]	
ELA using ELA	-0.0097	69	0.0137**	229	-0.0363***	229
	(0.0232)		(0.0065)		(0.0101)	
	[-0.0316]		[0.0477]		[-0.1128]	
Ratings using Ratings	-0.0074	206	-0.0048	229	0.0027	375
	(0.0058)		(0.0041)		(0.0036)	
	[-0.0241]		[-0.0168]		[0.0085]	
Composite using						
10% Ratings	0.0246***	206	0.0141**	229	-0.0161***	366
	(0.0071)		(0.0058)		(0.0062)	
	[0.0797]		[0.0494]		[-0.0499]	
30% Ratings	0.0200***	206	0.0154***	229	-0.0091*	366
	(0.0070)		(0.0058)		(0.0052)	
	[0.0647]		[0.0537]		[-0.0283]	
50% Ratings	0.0129**	206	0.0140***	229	-0.0017	366
	(0.0061)		(0.0052)		(0.0050)	
	[0.0420]		[0.0488]		[-0.0054]	
70% Ratings	0.0075	206	0.0043	229	0.0055	366
	(0.0046)		(0.0043)		(0.0034)	
	[0.0244]		[0.0149]		[0.0172]	
90% Ratings	-0.0014	206	0.0000	229	0.0074**	366
	(0.0048)		(0.0040)		(0.0035)	
	[-0.0047]		[-0.0000]		[0.0229]	

Notes: This table shows the change in performance generated when dismissing the bottom 10% of teachers using three years of data measured in student test score standard deviations or summative rating points. These models add race (white or non-white) and gender as predictors to the OLS models defined in Section 4. 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 comparison group changes in the remaining rows. These rows record changes relative to the *first row* using the models 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 (teacher-level standard deviations) in the dataset are

included in brackets.

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

Table A12: Residual Difference by Group Using 3 Years of Data

	Female - Male	N	White - Non-white	N
Math using Math	-0.0151	212	-0.0192	212
	(0.0298)		(0.0415)	
	[-0.0489]		[-0.0621]	
ELA using ELA	0.0174	232	0.0186	232
	(0.0278)		(0.0245)	
	[0.0606]		[0.0650]	
Ratings using Ratings	-0.0044	390	-0.0139	390
	(0.0200)		(0.0237)	
	[-0.0137]		[-0.0433]	

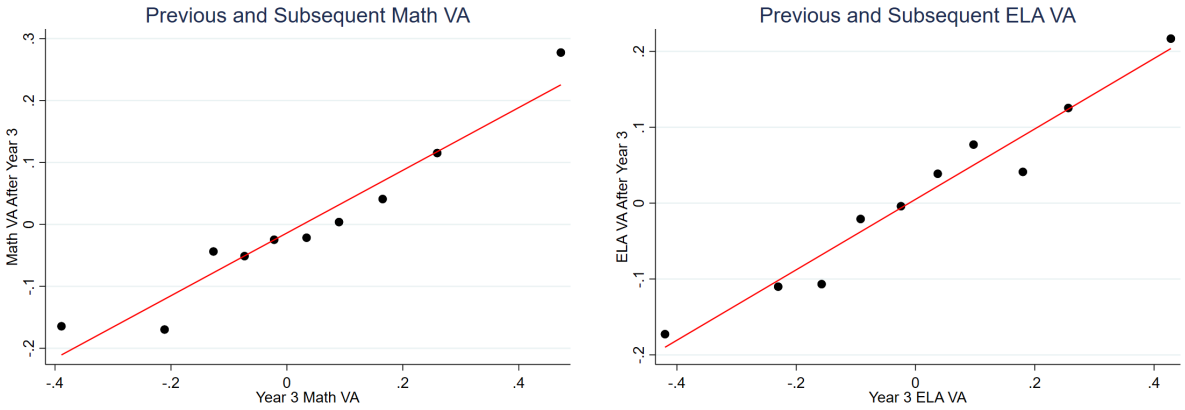
Notes: This table compares group-wide average residuals generated by the OLS models described in Section 4 measured in student test score standard deviations or summative rating points. The row headers define the model's outcome and predictors. The first two columns show the difference between male and female teachers, while the second two columns show the difference between white and non-white teachers. In the female-male (white-non-white) comparison, a negative value suggests that the model 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 (teacher-level standard deviations) in the dataset are included in brackets.

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

A.7 Appendix Figures

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



Panel C: Summative Ratings

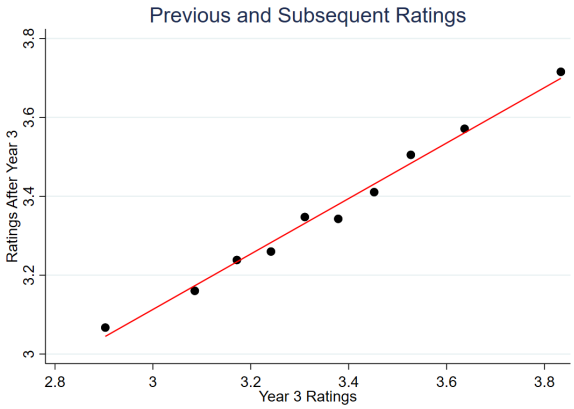


Figure A1: Relationship Between Previous and Subsequent Performance

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.

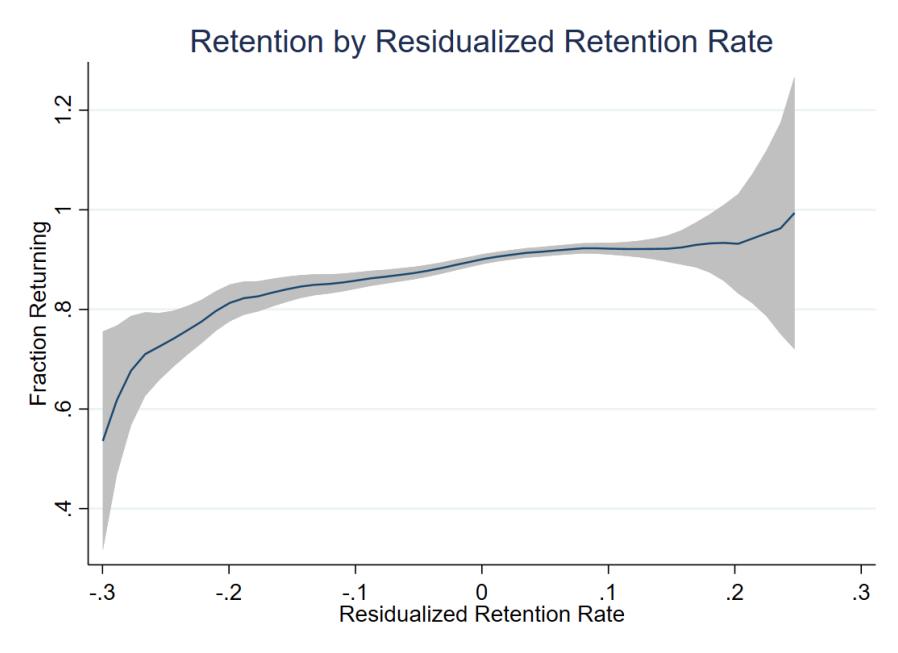
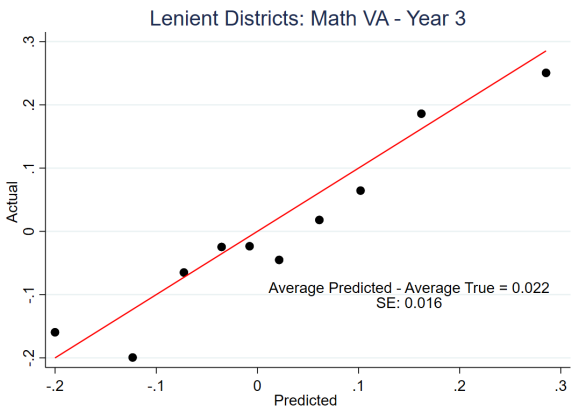
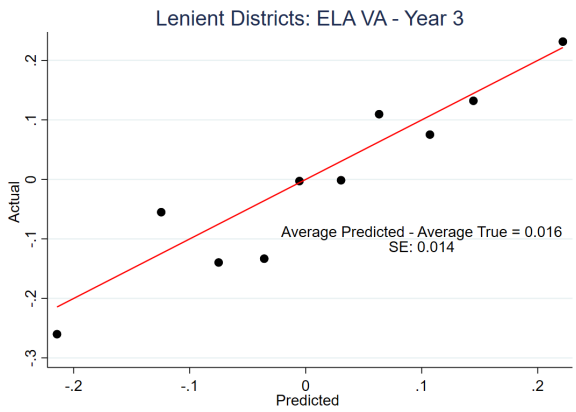


Figure A2: Retention by Residualized Retention Rate

Notes: This figure plots a local quadratic regression of the retention rate against the leave-one-out mean residual from equation (5). 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 sparse and generate noisy estimates.

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





Panel C: Summative Ratings

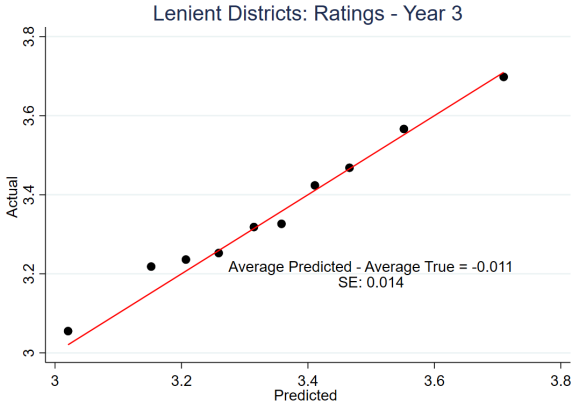


Figure A3: Actual and Predicted Performance in Low-Dismissal Districts Estimated on High-Dismissal Districts

Notes: This figure shows the relationship between predicted and actual subsequent performance in low-dismissal districts based on models estimated in high-dismissal districts. I use OLS models defined in Section 4 based on three years of data. The performance measure of interest is labeled in each graph. The x-axis records the mea predicted performance in 10 equal-sized bins, while the y-axis records the average actual performance within that bin. I define low-dismissal (high-dismissal) districts as those with above (below) median leave-one-out average residuals from equation (5). In each graph, I include 45° lines, the mean deviation, and the standard error of the deviation.