Informing Teacher Tenure Decisions using Value-Added Modeling:
The Case of New York City

by

CARL JEPSEN MCPHERSON

Lara Shore-Sheppard, Advisor

A thesis submitted in partial fulfillment of the requirements for the Degree of Bachelor of Arts with Honors in Economics

WILLIAMS COLLEGE

Williamstown, Massachusetts

May 2014
Abstract

In recent years, many states and school districts have enacted policies that weaken tenure protections for public school teachers. Value-added (VA) models for measuring teacher effectiveness have also become increasingly popular. Although there has been much conjecture over the impacts of both trends, there have been few opportunities to study their effects empirically. I examine the effects of a policy combining elements of both ideas that was implemented by New York City in 2010. I show that the policy did not induce less effective teachers to leave, but a positive effect of the program may be undetectable due to data issues. The policy had no impact on teacher effort as measured by the frequency of absences or the rate at which teachers improved, and it also had no obvious effect on teacher supply.
Acknowledgments

My thesis would not have been possible without the help of many people. I would like to thank, in particular, Professor Lara Shore-Sheppard, who generously and bravely agreed to be my advisor before ever meeting me. Her advice and commentary, along with that of Professors David Zimmerman and Jon Bakija, has greatly improved the quality of this work. I am also grateful to Professor Sarah Jacobson, who helped me craft my initial project proposal, and the many other dedicated professors and teachers that have helped me at some time or another.

I also want to thank my family, whose love and support play an immeasurable role in every one of my accomplishments. Much thanks is also due to my friends at Williams and beyond, for making my life richer, easier and more enjoyable. I am grateful to the Barrett family, who opened their doors to me when I needed a place to stay while doing research this past summer.

Finally, I would like to thank The Williams Economics Department for their generous financial support through the Carl Van Duyne Prize, and the New York City Department of Education, for providing the data for this thesis. This paper tends to focus on the shortcomings of this data, but readers should not forget the generosity required to share such a large amount of information with an unexceptional undergraduate researcher.

All errors are my own. I intend no disrespect by my typos, and I encourage feedback on my work.

Contact: cjm4@williams.edu
## Contents

1 Introduction .......................... 1

2 Background .......................... 3
   2.1 Tenure Law in the United States .................................. 3
   2.2 The Tenure Debate ................................................. 4
   2.3 Teacher Evaluation .................................................. 6
   2.4 Selective Dismissal ................................................. 10
   2.5 The Case of New York City ......................................... 12

3 Theory .................................. 16

4 Data .................................... 18
   4.1 Data Structure ...................................................... 18
   4.2 Population Restrictions ............................................ 20
   4.3 Limitations and Concerns ........................................... 22

5 Methodological Overview ............... 27

6 Specific Methods and their Results ..... 28
   6.1 City-wide Trends ................................................... 28
   6.2 Value-added Model Estimation .................................... 29
   6.3 Deselection .......................................................... 31
   6.4 Effort ............................................................... 33
   6.5 Teacher Supply ..................................................... 37

7 Conclusions ............................ 38

8 Tables ................................... 41

9 Figures ................................... 53

10 References ............................. 65

11 Appendix A: Primary Sources ........... 68
List of Figures

1  Comparison of Overall Turnover Trends ........................................... 53
2  Comparison of Turnover Trends for Teachers with < 3 Years of Experience ................................................................. 54
3  Percent of Teachers with < 5 Years of Experience .......................... 54
4  Tenure Decisions over Time ............................................................. 55
5  Average 4th and 8th Grade Math Scores on NAEP and NYS Exams ... 56
6  Average 4th and 8th Grade ELA Scores on NAEP and NYS Exams ... 56
7  Distributions of Math VA Scores ....................................................... 57
8  Distributions of ELA VA Scores ....................................................... 58
9  Mean Math VA Score by Year in Career ........................................... 59
10 Mean ELA VA Score by Year in Career ............................................ 60
11 Early gains in VA for Math Teachers ............................................... 61
12 Early gains in VA for ELA Teachers ............................................... 62
13 Time-Variant VA Score in First Year .............................................. 63
14 Persistent VA Score in First Year ................................................... 64
15 The Principal’s Task .......................................................................... 68
16 The Evaluation Rubric ...................................................................... 69

List of Tables

1  Summary of APPR Rubric (2010) ....................................................... 41
2  Data Structure and Sources ............................................................. 42
3  Mock-up of Selected Entries and Variables in Main Dataset ............... 43
4  Selected Summary Statistics ............................................................ 44
5  Years Experience at time of Tenure Decision ................................... 45
6  The Number of Tenure Decisions Made Annually .............................. 45
7  Predicting Induced Turnover for Math Teachers ............................... 46
8  Predicting Induced Turnover for ELA Teachers ............................... 47
9  Teacher Absences ........................................................................... 48
10 Year-to-Year Change in Math M.E. VA ........................................... 49
11 Year-to-Year Change in Math F.E. VA ............................................. 50
12 Year-to-Year Change in ELA M.E. VA ............................................ 51
13 Year-to-Year Change in ELA F.E. VA ............................................. 52
1 Introduction

Between 1995 and 2009, real per pupil spending in the United States increased 35% at the primary and secondary levels, and, although student achievement did rise over this same period, the improvement was mediocre in comparison to other OECD countries (Hanushek et al. 2012). Educators and policy makers tried and are still trying a variety of methods to improve student performance, from increasing the number of charter schools to decreasing class size (U.S. Department of Education 2012). In 2009, the U.S. Department of Education launched President Obama’s “Race to the Top” program, which offered $4 billion in funding to the states that scored best on a rubric that emphasized data-driven approaches and holding schools and educators accountable for student performance. Since then, states have rapidly adopted a number of different policies to quantitatively measure and systematically increase teacher quality (Jacobs et al. 2012).

One natural approach to improving overall teacher quality might be to institute teacher training or mentoring programs, and many school districts around the country are experimenting with these. New York City, in the 2013-14 school year, implemented a program that provides teachers with detailed evaluations and advice on the techniques they use in the classroom (New York City Department of Education 2012). Another approach might be to institute more selective hiring practices, but determining teacher quality using characteristics observable at the time of hire is difficult. The average teacher with a master’s degree in education, for instance, has not been found to be more effective than the average teacher with a bachelor’s degree (Staiger and Rockoff 2010). Because of this difficulty, many researchers and policymakers have turned their attention to measuring teacher quality using on-the-job performance.

A number of options for assessing quality, such as in-classroom observations or student survey data, immediately come to mind, but, perhaps motivated by Federal emphasis on standardized testing and quantitative measurement or by recent research, many states are experimenting with value-added models (VAMs) of teacher effectiveness. VAMs will be treated in more detail later, but their goal is to determine teacher quality using only student achievement on standardized tests. Teachers with high VA scores are taken to be more effective than teachers with low VA scores.

The accuracy of VAMs is still debated, but, even assuming that principals could measure teacher effectiveness perfectly, the question of what to do with that information would still remain. Some school districts have tried offer bonuses to encourage effective teaching and retain good teachers, but the salary step structure of most
union contracts is not particularly congenial to this type of reform (Dee and Wyckoff 2013). Other districts employ professional development programs or mentoring programs. Although the lack of empirical evidence for the benefit of a Masters in Education speaks to the difficulty of this project, some programs seem to have positive results (Taylor and Tyler 2012). A third option is “deselecting” ineffective teachers, removing them from classroom duty by firing them, transferring them to other jobs within the school system or encouraging them to leave voluntarily.

Rightly or wrongly, attempts at programs that encourage deselecting potentially ineffective teachers are often hindered by the employment protection provided by tenure. Efforts for more rigorous teacher evaluation, therefore, are often coupled with efforts to weaken employment protections. In 2011, Idaho became the first state to explicitly reject a tenure system in its legal code. Thirty-two states enacted some type of tenure reform between 2008 and 2011 (Christie and Zinth 2011). As of 2013, twenty states explicitly tied teacher evaluation to student growth, as measured by VAMs or average test score benchmarks, compared to four states in 2009 (Doherty and Jacobs 2013). Many researchers have estimated the VA for all the teachers in a school district and projected the effects that a deselection program might have in theory (Rothstein 2013). In spite of their growing popularity and importance in both teacher evaluations and tenure decision-making, however, there has not been much empirical research into the costs and benefits of such programs.

New York was an early mover in both these policy trends, and provides one of the first opportunities to study a large-scale implementation of the newly popular practices. In 2010, its state legislature enacted a plan to evaluate teachers using VA and other measures in order to determine tenure decisions and make tenured teachers easier to fire. The resultant political battle in New York City (NYC) between the Mayor’s Office and the teachers’ unions prevented this law from taking effect until after the 2012-2013 (henceforth referred to as 2013) school year. This paper focuses on the intervening years, when a weaker version of the policy was in place. The Mayor’s office required principals to justify tenure decisions based on an evaluation rubric, for which VA scores were a suggested form of evidence (see Appendix A for the rubric).

In this paper, I primarily use state test score data in math and English language arts (ELA) for NYC students and their teachers for grades 3 through 8 between 2006 and 2012 to examine the effects of the policy shift on the average effectiveness of teachers as measured by VAMs, teacher effort and the quality of new hires. I find no convincing evidence of a significant positive or negative impact of the policy in
any case. Although this may be an artifact of a number of issues with the data, the teachers selected for tenure under the new policy did not seem to be any more effective than those who were selected under the previous policy. Effort level did not seem to increase, in terms of the frequency of absences or the return to experience. Teacher supply did not apparently change after the policy change.

Section II provides the background for this paper. It reviews tenure law in the United States as well as the current debate on tenure and teacher evaluation, and the recent policy changes regarding tenure and teacher evaluation in NYC. Section III provides a theoretical framework for the effects of the policy on total teacher output. Section IV details the data and its limitations. Section V gives an overview of the methodology. Section VI is the results and Section VII concludes.

2 Background

2.1 Tenure Law in the United States

Tenure, by a very simple definition, is a set of legal protections that prevent a teacher from being fired suddenly, arbitrarily or without due process of law. Almost every public school in the United States has a tenure system in place, and has had one of some kind or another for over half a century. Because of the federalist nature of the U.S. education system, the specific details of teacher tenure vary greatly across the country. Almost every state passes guidelines for how tenure is granted and what protection it provides and leaves some of the implementation up to collective bargaining between school districts and teachers’ unions. There are some exceptions: Hawaii leaves tenure agreements entirely up to the districts, and Idaho began phasing out tenure in 2011 (Christie and Zinth 2011).

Little idiosyncrasies and some exceptions aside, however, tenure law can be broken down into two parts: due process protections and how these protections are gained. In most states, due process entails providing notice to a teacher who will be fired, giving him or her some opportunity to improve, and an appeals process if the teacher believes that he or she was fired unfairly. Usually, fair reasons to fire a teacher are immorality, neglect of duty and incompetence with further allowances for other good and just causes (Christie and Zinth 2011).

Traditionally, teachers earn tenure during a “probationary period,” spanning the early years of their careers, during which they have very few employment protections. Probationary periods differ from state to state, but the average period of probation is
three years. Mississippi has the shortest defined probationary period at one year, and five states have the longest at five years. In some states, completing this probationary period is still the only requirement for receiving the security of tenure, but most states require that new teachers be deemed effective in some sort of evaluation during this time period (Jacobs et al. 2012).

The customary method for teacher evaluation is one or more in-classroom observations performed by a school administrator or fellow teacher. The adequacy of these evaluation procedures for identifying good teachers and assisting in professional development has been increasingly questioned in recent years (see Kane et al. 2012, for example). The growing dissatisfaction with traditional subjective evaluations is partly motivated by federal education reforms, such as No Child Left Behind (2003), which emphasized a scientific and quantitative approach to improving education, and Race to the Top (2009), which distributed billions of dollars towards developing data systems and effective faculty. Between 2008 and 2011, 32 states altered their teacher-evaluation policies, and almost all of these reforms tried to incorporate more objective measures of teacher quality (Christie and Zinth 2011). Tenure policy in the U.S. is changing rapidly, and the focus is largely on determining who gets tenure. These changes will have important implications for teachers and students, so it is important that we understand them well.

2.2 The Tenure Debate

The recent torrent of changes to U.S. tenure law is also the product of an ongoing debate over how strong tenure protections ought to be and how difficult they should be to earn. One camp wants to water down protections and make them harder to attain, the other believes tenure is working reasonably well as it is, and that there are more pressing problems within the education system. At issue is whether tenure protections are so strong that it is difficult to fire ineffective teachers, and if evaluations are rigorous enough to prevent ineffective teachers from receiving tenure or provide evidence for the dismissal of a tenured teacher.

Tenure protection, according to much of the popular media, is extremely powerful. Every so often, a horror story is published about how a clearly unfit teacher remains in the classroom due to tenure (e.g. Song 2009). Union leaders respond that these cases are exceptional, that tenure protects only good teachers, and that it does not prevent the prompt dismissal of unsatisfactory ones (e.g. Weingarten 2010).

Unbiased data on the strength of tenure is difficult to come by, since it is often
difficult to determine from administrative records why and how a teacher left a school. Data from the national Schools and Staffing Survey (SASS), administered by the Census Bureau, shows that principals report dismissing a far smaller percentage of tenured teachers for poor performance than probationary teachers, but that could mean either that tenure protects bad teachers or that the tenure approval process is effective at identifying good teachers.

Although it was not looking at tenured teachers, an interesting study by Jacob (2010) examines whether more teachers get fired when employment protections are weakened and if the teachers who are dismissed are systematically different from the ones that are not. After a collective bargaining agreement in Chicago schools was altered so that principals could fire probationary teachers without evidence of poor performance, the separation (dismissal and voluntary departure) rate for probationary teachers nearly doubled. Using a logit model, Jacob determines that separation is predicted by poor observable characteristics. He finds that a standard deviation increase in the average test score gains achieved by a teacher’s students reduced his or her chances of separation by about seven percent. He also found that high absenteeism increased the likelihood of dismissal while high subjective evaluations decreased a teacher’s chances of being fired.

Jacob’s study indicates that if employment protections are weakened, principals will fire more teachers on the lower end of the performance spectrum. This seems to suggest that tenure keeps dismissal rates too low, but that is only part of the story. The dismissal rate of teachers may be affected less by legislation or explicit employment policies than by social norms and tacit understandings. Although most Chicago schools dismissed between nine and thirteen percent of probationary teachers under the new collective bargaining agreement, over thirty percent of schools did not fire any probationary teachers in any year of the study. This undermines the assertion that tenure alone is keeping dismissal rates too low.

It is also worth mentioning that over fifty percent of the probationary teachers that principals dismissed were later rehired into the Chicago school system. This may reflect a belief that teacher effectiveness is heavily dependent on the situation or it may hint at a supply shortage. From both of these possible conclusions come arguments for focusing on teacher mentoring and development rather than outright dismissal.

Questions like these have led people on both sides of the debate to push for better evaluations of teacher effectiveness. It is only in understanding how many (if any) teachers are persistently underperforming that one can determine how many
and which teachers would ideally be dismissed or granted tenure in the first place. Better evaluations would also help dismiss any ineffective teachers if they were found. Because tenure is essentially a due process law, it requires that the grounds for dismissal be substantiated by evidence. If an evaluation cannot serve as evidence for incompetence, tenure effectively prevents bad teachers from being fired.

2.3 Teacher Evaluation

The traditional method of teacher evaluation in the United States has often been characterized as a principal, or some other administrative figure, sitting in on a teacher’s class for some minutes and taking notes. This sort of evaluation might happen once or twice a year, and would result in a usually binary satisfactory-unsatisfactory rating. Whether evaluation works is a matter for debate. Weisburg et al. (2009) review teacher evaluation records from eleven large school districts around the country and find that one percent or fewer of all teachers evaluated were deemed “unsatisfactory” after in-classroom observations. If it is the case that only one percent of teachers really are “unsatisfactory,” the results from Weisburg et al. (2009) are evidence that these evaluation systems work well, and provide enough evidence to remove incapable educators effectively.

Many observers, however, have taken these almost universally positive performance reviews as evidence of leniency bias. In order to distinguish between the best and worst in the profession, some lawmakers and educators have looked for a more standardized and data-driven system. The foremost candidates among these evaluation systems are VAMs, which are now the most important component of a teacher’s performance review in twelve states (Jacobs et al. 2012).

The exact specifications of VAMs differ from district to district and state to state, but each purports to measure the average change in student test scores caused by the teacher, isolating the contribution of a teacher to a student’s improvement in test performance from individual heterogeneity and random shocks. Essentially, VAMs can be thought of as using past student test scores and often student and peer characteristics to predict how a student ought to perform on later tests. The difference between the student’s predicted score and his actual score is attributed to the influence of the teacher in the year that he actually took that test. These differences are averaged over all of a teacher’s students and called a VA score. Another way of thinking about it is that VAMs estimate which teachers are effective by comparing the test score changes of their students to the score changes of similar students taught
by different teachers. Certainly, these methods avoid the problem of giving teachers universally positive ratings. Indeed, since VAMs measure effectiveness by comparison to other teachers, if there is any variance at all, they cannot fail to find many teachers below average. VAMs, however, have problems of their own.

The problems with VAMs can be broken down in much the same way as the problems that have been identified with standardized testing. At issue is, first, whether they are accurate enough to be used in life-changing decisions such as promotion and firing, and, second, if they are accurate, whether using them creates perverse incentives and negative environments that do more harm than the information they provide does good. In the case of standardized tests, it is clear that scoring well on any given test and truly understanding the important material in the wider curriculum are not necessarily the same things. Test scores are affected by more than how well students know the material and the design of the test; they are often influenced by random factors, such as the infamous dog barking outside the window on test day.

The second problem with tests comes with making them more consequential. Making funding contingent on student test scores, such as in No Child Left Behind, has led to negative outcomes at some schools, such as teaching to the test, triaging students to focus only on those who could improve their scores and even outright cheating (Supovitz 2010). Making tests more important provides increased incentives to distort their accuracy, and these distortion efforts can hurt student learning.

Since VAMs make use of standardized test scores, they inherit all of the problems associated with them, but the additional assumptions attached to VAMs further complicate these issues. Making the results of value-added modeling too consequential encourages additional forms of bad behavior, such as competing for the students most likely to make large achievement gains. There has been an explosion of economic research on VAMs in education in recent years, but most of it focuses on their usefulness, biases and predictive power. As with selective dismissal policies in general, researchers still do not know the overall impact of using VAMs to make crucial staffing decisions (Rothstein 2013).

There is also the question of whether or not VAMs are accurate enough to be used in making staffing decisions. The largest problem is the assumption that all of the variation unaccounted for by the model is due to the teacher. In spite of teachers’ substantial effect on student outcomes, their value-added score may ultimately depend on more important factors outside of their control, such as a student’s home life.

Depending on the model specification, previous studies have found that teachers can only account for about 3–18% of a student’s achievement gains, and thus it can
be hard to detect the signal through the noise (McCaffrey et al. 2004). It would be impossible to say how much of the variance in student gains is due to teacher effects without controlling for every relevant variable, but generally, as the number of included variables increases, so does prediction bias. It is a difficult balancing act. If one excludes a predictor like race from the model, then, based on the differing racial makeup of various classrooms, one is attributing the achievement differences between races to differences in effectiveness between teachers. If, however, one includes too many predictor variables, then one risks over-fitting the model, in which case it will also fail to measure teacher effectiveness. [There are also non-statistical issues with deciding how to specify VAMs. Some school districts, such as Dallas, have opted not to control for characteristics such as race, even if they have predictive power, because they do not want to do anything that seems like it allows teachers to hold certain students to lower standards (McCaffrey et al. 2004).]

The ideal situation for isolating teacher effects would be to randomize the assignment of both students and teachers to classes, but this is not what happens. Parents’ choices about where to live are influenced by school quality. Students are often grouped into classes by ability, and sometimes exercise choice over their own schedules. Teachers are also paired with classes based on their particular skill sets. Rothstein (2010) using panel data on 90,000 middle school students in North Carolina, uses 4th grade test score growth as the outcome variable of a number of different VAMs that were designed to predict 5th grade test score growth and finds that 5th grade teachers have large “effects” on their (future) student’s fourth grade test scores. Rothstein determines that these effects were caused by a tendency for students with similar test score gains in 4th grade to be placed in the same classroom in 5th grade, and by the fact that test score gains were serially correlated. Certainly, which teacher a student is assigned to is non-random.

This leads to questions of what VAMs are actually measuring. Chetty, Friedman and Rockoff (2013) put together a huge dataset, linking a decade’s worth of testing data from a large urban school district to tax records, to determine, among other things, the magnitude of selection bias. They are able to replicate Rothstein’s results, finding small but significant bias because of grouping on lagged test scores, but they go on to show that there is probably not a lot of additional bias to worry about.

One method they use to look at this is the effect of staffing changes. The argument is that, if a teacher is new to a school, parents and students will not know enough about that teacher to try and get into or avoid her classes, and that students are unlikely to switch schools because one teacher leaves. Furthermore, if VA is a good
measure of teacher effectiveness, that teacher should have a similar impact on her new class. Chetty, Friedman and Rockoff find that schools that gain a teacher with a VA score in the top 5% experience test score gains on par with what would be predicted based on the change in the mean VA score for that school. Unsurprisingly in these cases, the VA score of the incoming teacher had no measured impact on their students’ test scores in the previous year. In the school the high VA teacher left, test scores fell at a level indistinguishable from what would be predicted by changes in mean VA. The opposite held true for teachers with VA scores in the lowest 5%.

Chetty et al. (2014), using the same data, go on to examine the relationship between a variety of life outcomes for students and the VA of their teachers in primary school. They find that students whose teachers had higher cumulative VA scores are more likely to attend college, less likely to have children as teenagers, and have higher earnings and savings rates, at least up until age 28. To combat the critique that the same student characteristics that cause better life outcomes also cause higher test score gains, they establish these relationships with a “leave-year-out” VAM, meaning that if the life outcomes of a certain student are the outcome variable, the VA score for the teacher in question was calculated using only the years in which that teacher did not teach that student. They also use the staffing-changes method employed in their previous paper. Evidence of the long-term impact of high VA teachers is important given evidence from papers like Jacob, Lefgren and Sims (2010), which show that teacher-induced gains on test scores fade out rapidly. It is also important because a positive impact on true learning should manifest itself in more than just a few years of higher test scores.

One final source of encouragement for those who advocate using VAMs in teacher evaluation is that they may also perform better than other measures of teacher effectiveness at predicting which teachers will induce test-score gains in their students. The $50-million “Measures of Effective Teaching” (MET) study funded by the Gates Foundation compared the results from five different evaluations based on in-classroom observation to VA scores as estimated on a variety of high- and low-stakes tests as well as student surveys. Teachers were recorded in front of the classroom four to eight times over the course of two years, with all teachers in the evaluation teaching similar material. Multiple trained evaluators for each evaluation system rated the videos from each year. Students took both annual state exams and exams designed to assess higher cognitive abilities. The MET study found that using teachers’ value-added scores alone was better at predicting student achievement on a variety of tests than any of five systems for observational evaluation, teacher education level, teacher
experience or student surveys. VA scores were also less volatile measures than scores from in-class evaluations (Kane et al. 2012).

In spite of a long list of theoretical and technical problems, many researchers today suggest including VA scores balanced by traditional in-class observations when evaluating teachers (e.g. Rockoff and Speroni 2010; Kane et al. 2012). The conclusion here is similar to the majority opinion on standardized-test scores: value-added scores are imperfect, but they are useful, and better when combined with multiple assessments.

Although this is probably the prevailing attitude among economists, it is not an easy consensus. There is still considerable debate over the validity of VAMs and, especially, the best specification for VAMs. One possible critique of the Chetty, Friedman and Rockoff papers, however, has nothing to do with the models and everything to do with their external validity. Following the lead of Hanushek (2011), the authors run a simulation of a deselection program in which teachers with VA-scores in the bottom 5% are fired and replaced with average teachers. It is not obvious, however, that such a deselection program would be possible. For one thing, it would make testing high-stakes. The data used in the Chetty, Friedman and Rockoff studies come from the 1990s, before the era of high-stakes testing. Such a program might also alter teacher supply and teacher effort, and it certainly involves increased hiring costs. Economists and educators know little about the consequences of making employment conditional on VA scores. As many states have already attached importance to VA scores in terms of promotion and retention, more research on the implementation of selective dismissal programs is sorely needed.

2.4 Selective Dismissal

Any system that proposes to improve the general quality of instruction by firing teachers based on evaluations hinges on two assumptions: first, that teacher quality is a significant determinant of student outcomes, and second, that there are significant variations in teacher quality. Research has largely validated these two assumptions. In their meta-analysis of ten papers that use VAMs, Hanushek and Rivkin (2010) find that the average estimated standard deviation in teacher VA is 0.13 in English and 0.17 in math. The difference between a top-quartile and bottom-quartile teacher in terms of value-added scores implies a 0.2 standard deviation increase in student test scores.

To put that number in perspective, note that a nine-month school year is asso-
associated with roughly a 0.25-0.4 standard deviation gain in test performance (Kane et al. 2012). The black-white achievement gap is between 0.7-1.0 standard deviation, and reducing class size by ten students results in gains of between 0.1-0.3 standard deviation in test performance (Hanushek and Rivkin 2010).

Results outside of VAM confirm these same two assumptions of teacher influence on students and heterogeneity in skill. The MET study determined that 14% to 37% of the variance in scores across classes, lessons and raters was attributable to persistent teacher effects. In this study, the difference between a top-quartile and bottom-quartile teacher in terms of observation scores meant increased performance on low-stakes tests by 0.07 standard deviations. High observation scores were also highly correlated with student feedback scores (Kane et al. 2012).

In light of the evidence, it seems reasonable to assert both that teacher capability is heterogeneous and that this heterogeneity explains some of the difference in student performance, even if this difference is small compared to factors outside of their control. This alone, however, only suggests that it is better for students to be taught by teachers that score high on evaluations.

It does not say anything about what to do with the information garnered from evaluations. Perhaps it is best to use it to target training programs towards ineffective teachers. Perhaps it is in fact best to fire the teachers, but how many teachers? A theoretical model put forth by Staiger and Rockoff (2010) in which 60% of the value-added score was not actually attributable to the teacher, and fired teachers could be replaced by average ones, predicts that principals would maximize student scores if they fired all but the top quintile of first-year teachers provided they had to make the decision using one year of data. This massive dismissal rate might suggest that schools should fire more teachers, but there are also potentially negative side effects of such a high turnover rate, such as a teacher supply shortage caused by increased risk and stagnant wages. Staiger and Rockoff attempt to address many of some potential concerns, and they never suggest that this particular policy be implemented, but it is clear that more empirical research needs to be done to determine the overall effects of using evaluations for high-stakes teacher selective dismissal policies.

Rockoff et al. (2012) is one of the few papers to examine the effects of incorporating value-added information in a teacher evaluation policy. Between 2007 and 2008, NYC ran a pilot program for 233 principals that volunteered. About half of these principals were randomly selected to receive value-added information on their teachers. In the year of the program, treated schools experienced slight (.044 of a standard deviation) and marginally significant \( p < 0.1 \) gains in math scores and insignificant gains in
English over control schools. Rockoff et al. also found that VA scores had an impact on the principal’s evaluation of math teachers (but not English teachers), and that more precise VA measurements had more of an impact on evaluations.

Unfortunately, the trial ended after just one school year, so Rockoff et al. could not look at the long-term effects of the program. There is also reason to believe that the program might have a different effect if it were mandated rather than voluntary, and city-wide rather than on less than 10% of schools, and if tenure decision-making and dismissal were further emphasized.

One important additional consideration when assessing the effects of dismissal programs based on evaluations is that evaluations can have positive effects on student outcomes even without dismissing bad teachers. Evidence from Taylor and Tyler (2012) suggests that giving a teacher feedback after an evaluation will cause a lasting improvement in his or her VA score. Using six years of panel data on mid-career math teachers in Cincinnati, they compare within-teacher variation in student test score gains over time to determine that, after the year of evaluation (which was not the same for all teachers, even with the same level of experience), students with treated teachers score about 0.1 of a standard deviation higher than their peers in classes taught by unevaluated teachers. A similar evaluation feedback program is currently being implemented in NYC, although it was not in place during the period under study.

2.5 The Case of New York City

The program that NYC implemented in 2010 was focused instead on selective dismissal. Figure 13 in Appendix A, which advises principals to “counsel out” less effective teachers, is evidence of this ethos. The DOE overhauled the evaluation system, requiring principals had to marshal specific evidence of teacher performance, and to use this evidence to justify tenure decisions. The DOE nudged principals to make VA scores the most important piece of evidence.

This change was precipitated by a legal reversal of the previous state-wide prohibition on using student test score growth to evaluate teachers. On July 10th, 2010, as a part of its Race to the Top application, the State of New York passed a law altering the structure of the Annual Professional Performance Review (APPR) that all teachers, tenured and probationary, must undergo each year. Since 2001, the APPR had evaluated teachers across a number of categories, including mastery of content, quality of instruction and classroom management using a binary unsatisfactory-satisfactory
system. It explicitly rejected using student testing data to evaluate teachers, and, since 2008, making tenure decisions about new teachers that were informed by student growth on tests was illegal. This prohibition was championed by teachers unions, which were incensed by the pilot discussed in Rockoff et al. (2012), and the subsequent decision by NYC to make teacher VA scores visible to all principals.

Under the new law, teachers were rated on a four-tier scale: “Highly Effective,” “Effective,” “Developing” and “Ineffective.” Twenty percent (twenty points) of the score is supposed to be determined by a state-developed student growth measure. For English Language Arts (ELA) and Math teachers for grades 4 through 8, this measure is based on the changes in the New York State Exams, the standardized tests administered annually by the state. The remaining point distribution is left up to individual school districts, but twenty of the points must be determined by multiple quantitative student growth measures and sixty points must be decided by direct measures of the teacher, such as multiple classroom observations, lesson plans or student feedback. If a teacher receives poor ratings for two consecutive years, he or she may be denied tenure or dismissed (NY State Education Law 3012-c).

From the point distribution alone, which is summarized in Table 1, it may seem that quantitative measures are weighted less than more traditional methods of teacher evaluation, which comprise 60% of the points. It needs to be emphasized, however, that teachers cannot be rated higher than their rating in either student category. For example, if a teacher receives only eight points from the state measure, the teacher will be rated “Developing” even if he or she receives 100% of the local points. In other words, state VA scores were meant to be much more consequential than the twenty-point allotment would make them seem.

This was an abrupt reversal of the previous law was not well received by many in the education community. Both major teachers’ unions in New York immediately sued the state, and thousands of principals petitioned lawmakers to reconsider. The courts did not overrule the law (although some sections were clarified following a 2012 case), the legislature did not budge and the law went into effect September 2010.

The struggle then turned to the implementation, which in NYC became a battle between the United Federation of Teachers (UFT) and the Mayor’s Office through the NYC Department of Education (DOE). State law stipulated that the new APPR would take effect for ELA and Math teachers for grades 4-8 in 2011 and all other teachers in 2012. In 2014, the state measure would begin to constitute 25% of the score, and would be based on a VAM. The individual school districts, over this period, had to decide exactly what the local measures would be.
Although unions failed to kill the new law in the courts, conflicts between the DOE and UFT prevented NYC from completely implementing a state-approved APPR until May 2013, when the State Commissioner intervened (King 2013). For failing to implement the program, NYC was denied $250 million in state funding (Slentz and D’Agati 2013).

This paper focuses on the policy change that did take place in 2010, which, if more limited than state law required, was enabled by this law and persisted until state intervention in 2013. Mayor Michael Bloomberg used the new freedom afforded by the law to go around the UFT and directly address principals, who make tenure recommendations, and superintendents, who make tenure decisions. A month after the new law was passed, he announced “the end of tenure as we know it” (Matthews 2010). Without reaching an agreement with the unions, he updated the Tenure Notification System (TNS), which, since 2006, had been an online resource to help principals and superintendents handle tenure decisions. Specifically, it tracked impending tenure-decision deadlines and notified principals through email when action was required. Among other things, this initial reform was meant to target tenure by estoppel, the automatic receipt of tenure in the absence of a timely decision. Beginning in 2010, the TNS also flagged teachers with low (city-calculated) VA scores as “tenure in doubt,” and teachers with high VA scores as “tenure likely.” It also reported other performance indicators, such as previous APPR ratings and attendance.

As an additional reform measure, principals were required to evaluate teachers on the new four-tier scale by combining evidence on student learning, instructional practice and professional contribution (the rubric is in Appendix A). This evidence now had to be marshaled and presented in a Tenure Recommendation Form. Principals were encouraged to extend the probationary period of teachers whom they were unsure whether to recommend for as many years as they felt necessary and the TNS was changed to allow them to process these extensions more easily. Finally, Bloomberg revoked some of the hiring freeze that had been in place since the end of the May 2009. This hiring freeze, imposed in the wake of recession-motivated budget cuts, required principals (except in special circumstances) to hire new teachers from the Absent Teacher Reserve (ATR). NYC places teachers with tenure in the ATR if, due to factors like decreased enrollment, there is no longer a position for them at their schools. The city wanted to reduce the size of the ATR because its teachers receive full pay without having classroom duties. Some principals claim that, if a teacher has been in the ATR for over a year, it is typically a sign that no one wants him or her in a classroom (Brill 2009). If that is the case, then the freeze may have made
principals may have been reluctant to deny tenure.

Starting in the 2011 school year, provided they had the budget for it, principals could hire new teachers to replace old ones, but, should they remove a recently tenured teacher from permanent classroom assignments, the other schools in their network would have to try and hire that teacher. It seems like Bloomberg was trying to incentivize denying tenure to ineffective teachers in two ways, by making them easier to dismiss before tenure, and more difficult to dismiss afterwards.

One final thing to consider is that, in 2010, New York State changed the exam schedule so that ELA and Math exams were taken one month later in the school year. Because this change allowed the tests to draw on a wider range of material, test questions also became less predictable. Although test scores are scaled to be comparable from year to year, and, furthermore, VA measures are comparative within years, this may add some minor noise to measurements of VA score to the extent that some teachers may have altered their behavior to teach to the new test, and the rate at which they adapted was uncorrelated with their innate effectiveness.

To understand the effects of all these political maneuverings and simultaneous changes, it is useful to focus specifically on the principal. In the 2010 school year, when a teacher was up for tenure, a principal would receive an email through the TNS notifying him or her that it was necessary to recommend this teacher for approval or denial. The principal would be able to see some information about the teacher, and would have access to that teacher’s VA score information, but the principal could not use that information to make his or her recommendation. If the principal decided to approve a teacher for tenure, the principal did not have to justify that decision.

In 2011, the principal again receives an email this time with more information linked to it. She is not prevented from using VA scores to evaluate the tenure candidate, but is encouraged to take it very seriously. If the teacher has a particularly high or low VA score, the system flags that teacher as likely or unlikely based on those scores. Whatever decision the principal makes, she must back up that position with evidence. (Every APPR, for every teacher, takes longer because of these new, more stringent requirements.) If the principal award tenure to a teacher that does not deserve it, then the school will likely be stuck with that teacher, but if she denies teacher tenure, there is the hope that she can hire a more effective replacement.
3 Theory

How would one expect principals and teachers to respond to this new policy? Here I consider the implications of a modified version of the model put forth by Staiger and Rockoff (2010). The intellectual ancestor of this model is Jovanovic (1979). In this model, the performance of teachers is not known before they begin working. Instead, principals must make decisions about retention and dismissal based on teachers’ observed performance $v_{jt}$, which is determined for teacher $j$ in year $t$ by a combination of that teacher’s persistent impact on students $\mu_j$, which might be conceptualized as innate ability, some time-variant component of teacher performance $\theta_{jt}$, which would capture factors like effort and the returns to effectiveness from experience, or shirking after receiving the job security of tenure, and $\varepsilon_{jt}$, which is caused by differences in student demographics, class size, testing conditions and whatever else varies from year to year and is outside of the teacher $j$’s control.

$$v_{jt} = \mu_j + \theta_{jt} + \varepsilon_{jt} \text{ where } \varepsilon \sim N(0, \sigma_\varepsilon^2)$$ (1)

Equation 1 is fairly flexible, but one thing it does not include is spillover effects. For the sake of simplicity, it does not allow observed teacher effectiveness to be influenced by the other teachers in a school, which is often an argument made in favor of strong tenure protections. If new teachers learn from their more experienced colleagues, then $v_{jt}$ will not capture the full benefit that experienced teachers have on students. Although it ignores some discounting, this omission might be justified in the tenure debate on the grounds that the effect is vanishing. If weakened tenure protections allow principals to fire more ineffective teachers, then, eventually, new teachers could expect to be surrounded by more skilled coworkers. To the extent that the ability of teachers to help their colleagues be more effective in the classroom is correlated with their impact on their own students, selecting teachers on $v_{jt}$ would lead to the highest positive spillover years later.

Although my empirical strategy allows for $\theta_{jt}$ to move in various ways with respect to time, it may be useful to consider here a functional form that a principal may expect. The two main contributors to $\theta_{jt}$ would likely be effort and experience. In a typical career concerns model, effort is determined in part by perceived job security, which, for teachers, is determined by seniority and tenure status, and is thus likely linear with a discontinuity upon tenure receipt. Experience would generate some marginally decreasing returns to teacher ability, thus $\theta_{jt}$ would take a functional form something like Equation 2.
\[ \theta_t = \begin{cases} 
\alpha + \beta t + \kappa t^\lambda & : \ t \leq T \\
\gamma + \delta t + \psi t^\omega & : \ t > T 
\end{cases} \]  
(2)

where \( T \) is the year of tenure receipt. The subscript \( j \) is repressed throughout.

Taking \( \lambda, \omega < 1 \) to generate decreasing marginal returns (though this property can be achieved in other ways using this functional form), the terms of \( \theta_t \) split up such that the first two terms in each interval of the function account for experience and the final terms account for experience. During the probationary period, a teacher has some baseline level of effort \( \alpha \), which likely decreases over time (\( \beta < 0 \)) because of ‘first in, first out’ staffing polices. Simultaneously, the teacher is gaining experience that improves his or her performance (\( \kappa > 0 \)) at a decreasing rate (\( \lambda < 0 \)).

Notice that the returns to experience may change before and after tenure (it is possible that \( \lambda \neq \omega \) and \( \kappa \neq \psi \)). This is because I expect the returns to experience to depend, in part, on the amount of effort exerted. The teachers that put the most time and care into their classes should improve more in a year of teaching than those who put in less work. One might even take \( \lambda \), for instance, to be a function of \( \alpha \) and \( \beta \). Equation 2, assuming \( \alpha > \gamma, \psi < \beta < 0 \) and the initial returns to experience are relatively large, says that \( \theta_{jt} \) will increase at the beginning of a teacher’s career, drop abruptly after receiving tenure, and start shrinking or even becoming increasingly negative late in a teacher’s career.

In NYC, principals have no power to grant bonuses, and the teaching profession in general does not offer many opportunities for promotion, so principals have to be creative if they want to influence \( \theta_{jt} \) independently of the city-wide policy change. I assume, therefore, that principals attempt to maximize student learning by selecting teachers with the most positive persistent impact on their students. Although they cannot observe \( \mu_j \) directly, they can get a better idea of the true value by repeatedly observing \( v_{jt} \) over multiple years. The greater the variance in \( \varepsilon_{jt} \) or the larger proportion of \( v_{jt} \) that \( \varepsilon_{jt} \) represents, the more difficult it will be for the principal to determine which teachers to retain. It is worth noting that, empirically, observations of persistent teacher effects are very noisy, as \( \varepsilon_{jt} \) is at least twice the size of \( \mu_j + \theta_{jt} \) (McCaffrey et al. 2004.)

After determining the expected value of \( \mu_{uj} \), the principal attempts to choose some effectiveness level cutoff which teachers must attain in order to receive tenure. She then maximizes expected student learning by awarding the tenure to only those teachers whose expected value of \( \mu_{uj} \) is above the cutoff in the year they complete their probationary period. Assuming that the principal must always have the same
number of teachers on staff, the optimum level of student learning is achieved by choosing a cutoff such that the output of the average teacher equals the marginal productivity of teachers approved for tenure. If the marginal teacher were less effective than average, then the principal would raise the cutoff to exclude that teacher. If the marginal teacher were more effective than average, then the principal could lower the cutoff and add another above-average teacher.

Picking the exact cutoff, of course, is a very difficult task, depending on the applicant pool, exogenous turnover and the returns to experience, among other things. The new tenure policy in NYC should have a positive impact on students if being able to examine VA scores improves information on teachers such that principals can estimate $\mu_j$ or set the cutoff with greater accuracy. It will also benefit students if principals could not previously deny tenure to all teachers whose effectiveness fell below the optimal cutoff. Furthermore, if ineffective teachers feel they are more likely to be identified and denied tenure, they may increase effort-levels, at least until after they receive the protection of tenure.

Conversely, if the policy decreases perceived job security, and inclusion error is significant, then the new policy may hurt student learning. In this situation, the most skilled teachers, with job options elsewhere, may choose not to work in NYC. Wages have certainly not risen to compensate them for their risk. If $\sigma^2$ is so large that $\mu_j$ cannot be estimated accurately, there may be no effect, even if denial rates increase. The theory suggests many possible outcomes, so I turn to the data.

4 Data

4.1 Data Structure

The data for this study come from a variety of restricted-use and publicly available datasets, organized at a variety of different levels. The sources, years and level of organization are summarized in Table 2.

The bulk of the analysis uses the student, linkage and teacher datasets. The other datasets are used for specific calculations, and I will discuss them each in turn as I use them. The student, linkage and teacher datasets were amassed by the NYC DOE for administrative purposes. The Student and Linkage data were collected using software called Automate the Schools (ATS). Schools are responsible for entering the data and for its veracity. Since 2009, the ATS has alerted schools to various problems in their data, such as missing values or inconsistencies, through a number of automatic checks.
The teacher data is maintained by NYC DOE Human Resources, mostly for payroll purposes.

Student data contains information on gender, ethnicity, free lunch status, English language learner status, attendance and Math and ELA New York State Exam scores for all test-taking students in grades 3 through 8 annually from 2001 to 2012. The DOE changed how they defined and recorded many of these variables of this period, so I recoded all of the relevant variables so that they matched, as best as possible, over each year. If a student was ever categorized as having a disability that makes learning more difficult, I treated that student as if they had that disability in each period. I did the same thing with English language learner status. If attributes that one would expect to be static, such as ethnicity or sex, changed over time, I set those characteristics to missing for that student. This happened less than 1% of the time.

The teacher data contains each teacher’s sex, ethnicity, salary step, school, absences, and years of experience as a teacher in NYC annually from 2007 to 2013. It also includes each teacher’s start date, and probation completion date (projected or realized) as of the end of the 2012 school year. I imputed each teacher’s education from their salary step, which is based on number of college and university credits and teaching experience. I impute experience levels for 2006 by subtracting one from teacher experience in 2007 if the teacher is still present. I use the 2013 data, which cannot be matched to students, only to help determine turnover rates in 2012.

The linkage data matches teachers to students, and contains information on the subject and grade taught and whether there were any assistant teachers. It does not always allow me to see if two students who were taking English in 7th grade in 2012 with the same teacher were in the same class. The linkage files for 2011 and 2012 are arranged and coded differently from the linkage files in previous years and from each other. These differences ought to be merely superficial, having nothing to do with how much or what type of data was been collected, but nevertheless I have been able to link more teachers to students in 2011 than in other years, and fewer than average in 2012. There appear to be some mistakes, which I discuss later, in the 2012 linkage data.

I can link data for each year from 2006 to 2012, so, if I do not include data on teacher characteristics in my models, I can calculate VA scores for each teacher from 2006 to 2012. I use the student and teacher identification codes in the linkage data to merge the student and teacher datasets and combine each year of data to form one dataset from 2006 to 2012 in which the unit of observation is student-subject-year (equivalently student-teacher-subject-year). I am also able to see which teachers
were still employed by the DOE and their characteristics for 2013. This is my main dataset.

Table ?? is a mock-up of what data appears in which years and how it is matched. It shows the final dataset, demonstrating that I have multiple years for each student and teacher, and that I can follow them throughout their time in the NYC school system. It also shows that I fill in all static characteristics for student or teachers who appear in the data without those characteristics in the raw data in any given year, such as every teacher does in 2006 because I do not have a teacher data file for that year. Along the bottom, boxes for each raw dataset appear below what columns they contributed to the final dataset. Data from the year 2005 is only used to get prior-year test scores for students in 2006. Data from 2013 is only used to help determine which teachers left the NYC DOE in 2012.

4.2 Population Restrictions

The main dataset, in its raw form, contains seven years of observations, 1,024,450 unique students, 126,906 distinct NYC DOE employees and 2,236 different schools comprising 5,887,605 rows of data, which should be roughly the population of test-taking students in grades 3 through 8 appearing once for ELA and once for math over these seven years and their teachers.

However, not all of these data are usable, consistent and relevant, so I have to eliminate some observations. I begin by dropping observations without important identification information. I eliminate 498,676 (8.5% of the raw data) observations because the student identification number is missing. This may seem high, but this number is largely driven by NYC DOE employees (principals, high school teachers, special education assistants and so on) who were not matched to primary school students taking ELA and Math courses. It is difficult to tell exactly how many employees should have been matched to students based on the data available, but, ignoring preschool teachers, district administrators and others who should clearly not be matched, it seems that only 2.6% of observations lack a student identifier.

Next I drop the 973,382 (18.0% of the remaining data) observations without an identified teacher. This is a large number, which I cannot fully explain. Around 1% out of that 18% (that is, 1% of the entire dataset, or 5.5% of the observations without an identified teacher) comes from those observations noted to be missing in the linkage data, the remaining 17% comes from unmatched observations in the student data. Five-percent of students (again, in the entire dataset) are never once matched to a
teacher. Another 1.5% of students are justifiably not matched to a teacher because there is no evidence they took Math or ELA tests. A further 2% were absent from school for 50 days or more.

I proceed to drop 66,758 (1.5%) observations that lack any testing data. I then consider students who have multiple observations in the same subject in the same year, which account for 112,387 (2.6%) observations. In most years, they represent less than 1% of the data, but they represent 20% of the observations from 2012. The duplicates in 2012 are particularly troubling because they come almost exclusively from linking the same students to multiple teachers. If it were simply a case of the same information being entered twice, I could perhaps keep one entry, but without knowing which entries are accurate, I am forced to drop all of them. Finally I drop 79 (0.0%) observations from students who were outside of the appropriate grade range. This leaves me with 4,236,323 observations, or 71.9% of my original dataset.

I now move on to more subjective choices. I eliminate all 17,438 (0.4%) observations from charter schools because those schools often have different hiring and retention practices than other NYC public schools. I also drop the 6,309 (0.1%) observations matched to non-standard teachers, such as principals, who although they were paired with students, seem to differ from regular teachers in ways that affect both performance expectations and staffing decisions. I drop 128,156 (3.0%) observations from students and teachers who were absent for more than fifty days because I do not believe it makes sense to compare their influence and achievement to those present for more of the 182 day school year. I drop teacher-subject-grade-year groups (the closest I can get to looking at classes) with fewer than ten students to avoid extreme lack of precision in VA estimation. I also drop teacher-subject-grade-year groups with over 200 students, which I consider to be an error in the data-entry process. The excessively small and large classes accounted for 178,019 (4.4%) observations. I am left with 3,906,799 observations, or 66% of my original dataset. Table 4 compares selected summary statistics from the raw and clean datasets.

Table 4 reveals some interesting trends. As time goes on, the average teacher is getting significantly more experienced and better-educated. The increase in experience may be due to the lingering effects of hiring restrictions, although it also increases from 2007 to 2009, before the freeze went into effect.

Test scores are also rising generally over this period, both before and after the policy shift. I examine trends in test scores more closely later in this paper.

As expected, the number of matched teachers is lower in 2012 than in other years because of the duplication error in the linkage file. From 2007 to 2011, the ratio of
teachers in the raw data to teachers in the clean data is fairly constant at around 9:1. In 2006, the ratio is far closer to one. This is expected because there is no separate teacher data in that year (recall that the file that contains all NYC DOE employees only covers the period 2007-2013), so the raw number should only include teachers in grades 3 through 8, which appear only as identification numbers.

Unfortunately, the sample in the cleaned data is different in a number of ways from the initial population. On average, the students do better on both ELA and Math exams. They are slightly less likely to be eligible for free or reduced lunch and to enter the school system as an English language learner. Although it is not shown, a slightly higher percentage of them are white. The teacher groups are not obviously systematically different after cleaning the data.

To the extent that the third of the data that I dropped should have actually been in my sample, and to the extent that the policy shift affected the missing members of the population more than the ones that remain, I will misestimate the effects of the policy. Given that all mechanisms through which this policy might have a positive effect on public school education in NYC run through the teachers, and the teachers in the cleaned data have, on average, roughly the same characteristics as the teachers in the raw data, I am not too concerned about misestimating the effects because of a misrepresentative sample.

A likely reason that the students in the cleaned sample seem to be slightly more advantaged is that schools with fewer resources are less likely to accurately relay data and also have more disadvantaged students. If worse teachers sort into these schools based on unobserved characteristics, then perhaps the students that do not appear in the data have the most to gain from a policy that attempts to select better teachers. In that case, I would underestimate the impact of the program. Still, the average student characteristics do not vary wildly between the raw and cleaned datasets, so I would probably not underestimate the effects by much.

4.3 Limitations and Concerns

A greater obstacle to my identification strategy than the amount of data I lose because I cannot match students to teachers is that I lack much of the information on teachers that I would like to have in order to assess the effects of the policy change. Specifically, I cannot discern the outcomes of tenure decisions or when the teacher first went up for tenure. Furthermore, I cannot distinguish between endogenous turnover, exogenous turnover and missing data. In other words, I never know why teachers disappear
from the data. It may be because they left of their own accord, were induced to leave because of the policy, or simply were not entered into the system that year because of some error.

One variable I do have to get at the outcomes of tenure decisions is the date that the teacher’s probationary period ended or is projected to end. Unfortunately, this variable is missing for 18.8% of the teachers in the cleaned data, and the extant values are incongruous with the experience variable about 15% of the time. It is impossible to say whether the experience variable or the probation completion date variable is more accurate. Table 5 shows the breakdown of years of experience at the time of tenure decision.

The shortest possible path to tenure is two years in the State of New York, which can only occur if the teacher already earned tenure in another school district and certain other conditions are met. The usual time to earn tenure is three years of teaching, but probationary status can theoretically be extended indefinitely. Because of the way that NYC keeps its records, the probation completion date is updated retroactively so that it is identical in each year that a teacher appears in the data, meaning I cannot know if a teacher’s probationary period has ever been extended, and this might be a reason that experience at tenure is often high. Anecdotally, however, a probationary period is rarely extended more than once. Empirically, it was not until 2013 that probation re-extension could be processed electronically. In short, five years of experience at tenure is rare; six is a true anomaly; and more than six defies belief, yet this is apparently the case for 7% of teachers. Still, this method of determining the time of tenure decision seems appropriate, and if the experience variable is taken as true, probationary period completion date is an accurate measure of tenure completion about 87% of the time for about 80% of the data.

The real concern comes from the fact that the number of tenure decisions in my data is not consistent with the official numbers, found in press releases published by the Mayor’s Office as evidence of the success of their tenure evaluation program. Each year the NYC DOE releases the number of teachers granted and denied tenure, as well as the number whose probationary periods have been extended. Unfortunately, these numbers are only presented for all teachers, and there is no precise way to estimate how many should appear in the data. Simply using the attrition in the number of teachers from the raw data to the current data underestimates the number of decisions significantly, because the raw data included non-teachers and because of missing data. I estimate the number of decisions that should be in the data using Equation 4, which is simply the officially reported number of decisions, multiplied
by the number of grades that are covered in the data over the number of grades in
NYC public schools, multiplied by roughly the number of observations in the data
that had values for probation completion date, multiplied by the non-extension rate,
the percentage of teachers who were granted or denied tenure. (If a teacher had her
probationary period extended, the completion date in the data would automatically
update itself for all years, and would show up in the data as a decision made the
following year.)

\[
\text{Estimate} = \text{Official Count} \times \frac{7}{13} \text{Grades} \times \frac{8}{10} \text{Non-missing} \times \text{Non-extension rate} \quad (3)
\]

While this rough estimate will obviously not be exactly correct, the approach is
reasonable and no less consistent with the data than other reasonable approaches.
The results of this estimation, along with the official number of decisions and the
number of decisions appearing in the data appear in Table 6. The average of the
magnitude of the percent error between the estimate and the actual number in the
data is 25%. The below-average numbers of decisions in 2011 and, particularly, 2012
are likely a result of the hiring freeze begun in May 2009.

Although the probation completion date variable does not seem too inaccurate
when compared internally to the experience variable, it does seem troublingly unreliably when compared to an external data source. If the variable for determining
the time of tenure decision is as imprecise as it seems, this would likely cause me
to underestimate the impacts of the policy change by mixing treatment and control
groups. I determine who is affected by the change based on the time they received
tenure.

Because the experience variable seems reasonably well-aligned with the unbelievable probation completion date variable, and the experience variable is important on
its own when calculating value-added and assessing effort, I check the credibility of
the experience variable by comparing it to state data. At the same time, I am able
to assess my ability to look at deselection through turnover.

Looking at turnover is important because, for my analysis, I am forced to assume
that if a teacher disappears from the raw data and never returns, that teacher has
left. Admittedly, this assumption is better in the earlier years of the sample than
at the end, but only about 17% of the teachers in the raw data leave and return
between 2006 and 2012. In the clean data, this number plummets to 2%. For this
definition, it is particularly important that I include the 2013 data, since 2012 has
so many teachers missing and because it gives me another cohort of teachers, those whose tenure decisions were made in 2012, with which to work.

Of course, I am much more interested in those who were induced to leave because of the policy than those who leave for other reasons. To try and get less noisy measurements of those who were induced to leave because of the policy, I also look at those who left in the year that their probationary period ends, and, for a somewhat larger sample size, those who left during their probationary periods.

As with the number of tenure decisions, I cannot replicate the turnover rates that are reported in another dataset. This time, I check my data against the publicly available data from the New York State School Report Cards. Every year since 2005, the NYSED has published school-level data summaries to help New Yorkers assess the quality of their schools, and they report a number of variables that are relevant to this analysis: the turnover rate (analogous to my definition of leaving), the turnover rate for teachers with less than 3 years of experience (analogous to leaving during probation), and the percentage of teachers that have less than five years of experience (as a basic check of the accuracy of the experience variable in the NYC DOE data). Turnover is defined in the NYSED data as a teacher leaving the school, not leaving the NYC DOE, as it is in the NYC DOE data, but the NYSED data still gives me a reasonable test of the validity of two ways of getting at induced departure and the experience variable in the teacher data set.

Figure 1 through Figure 3 show the comparison in measured trends between the NYC DOE and NYSED data. In order to make these graphs more comparable, I use only NYC schools in my calculations with the NYSED data, and only those 86% of schools in NYC for which I am able to find Report Card data. I calculate the statistics from the NYC DOE data the same way that they were calculated with the NYSED data, averaging first by school and then over school. I count teachers leaving the data and those switching schools towards turnover. I do not do calculations involving experience level in the NYC DOE data for 2006 because I cannot impute experience for all teachers.

I do not, however, expect the numbers from the NYC DOE and the NYSED data to line up perfectly for a number of reasons. The NYSED data includes a slightly larger sample of teachers because those who work at matched schools, but not in grades 3 through 8, are also included in the average. To the extent that turnover rates for these teachers differ from turnover rates for 3 through 8, the turnover rates will differ across datasets. I expect to underestimate the turnover rates in the NYC DOE data because I do not count teachers that go missing from the data, but reappear.
in the same school before the end of the observation window towards turnover. It
may be that some of these teachers really do leave and come back, which means they
are counted towards turnover in the NYSED data. If this happened for all teachers
in the NYC DOE data that go missing and then reappear, I would underestimate
turnover by around 1.3%.

Figure 1, however, shows that a fairly consistent discrepancy of around 10% be-
tween measured overall turnover rates in the NYC DOE and NYSED data. Figure 2,
which shows trends in turnover rates for teachers with fewer than 3 years of expe-
rience, shows a similar discrepancy. The measured turnover rates in the NYC DOE
data are roughly 10% lower than those in the NYSED data, but the trends in turnover
rates are relatively flat in both cases.

I have discussed already how the NYC DOE retroactively updates some its vari-
ables, such as probation completion date, so that the values of that variable match
over all years. If measured turnover is lower in the NYC DOE data than in the
NYSED data because some teachers who have left are retroactively removed from
the data, and which teachers are retroactively removed are random, then the low
measured turnover in the NYC DOE data should only hurt the precision with which
I measure the impact of deselection. If they are not removed randomly, then this may
cause some bias. Teachers that are removed may be slightly worse on average simply
because only teachers that leave can be retroactively deleted, and teachers that leave
may be less talented or less dedicated teachers than those who stay.

The trends in the percent of teachers that have fewer than 5 years of experience,
shown in Figure 3 are also inconsistent across the two data sets and consistent with a
retroactive deletion explanation. According to the NYC DOE data, about 40% of the
teachers at an average school had fewer than 5 years of experience in 2006, a statistic
that is unbelievably high. Of course, the number of teachers with fewer than 5 years
of experience in 2011, according to the NYSED data, is suspiciously under 5%. I
cannot say which, if either, of the two data sets is accurate, but it is obvious that
both cannot be correct. The only encouraging result in Figure 3 is that the trend in
both datasets seems to show the effects of the May 2009 hiring freeze.

If this teacher-experience measure only looks wrong on aggregate because of
retroactive deletion, then this should not have any negative impact on my results
because it is measured accurately at an individual level. If the experience variable is
indeed incorrect at the individual-level, then it is difficult to say how it will affect my
results. I attempt to control for experience when calculating VA, and I consider the
returns to experience when examining effort, but how the error will affect my results
depends entirely on the extent, direction and consistency of the mismeasurement.

Overall, it seems that my difficulties in determining the outcome of tenure decisions, and the fact that I seem to underestimate the number of tenure decisions made and teacher turnover, are likely to cause me to underestimate the new policy’s impact of teacher deselection. While not entirely safe, my estimations of the impact on student effort and on teacher supply are less likely to be affected by these issues, assuming retroactively deleted teachers are not much worse, on average, than the teachers that remain in the data. If the reasons for this are not clear already, I hope they will become clear as the method of ascertaining these effects is discussed in greater detail.

5 Methodological Overview

The ultimate goal of this study is to determine what effect, if any, NYC’s tenure approval policy from 2011 to 2013 had on student outcomes compared to the previous policy regime, but the most direct impacts of the policy change would be on teachers. Thus, I begin by looking briefly at city-wide trends in tenure-decision outcomes and student achievement on tests. I then turn to investigating a number of pathways through which this policy might impact student outcomes. First, I investigate whether or not the new tenure policy was an effective deselection program, that is, whether it induced more or worse teachers to leave the classroom than the previous policy. I estimate VA scores for each teacher and use a logit model to determine whether VA scores and other teacher characteristics were more predictive of teacher departure after the new policy was implemented. Second, I investigate whether the new evaluation system leads to increased teacher effort. I use the number of absences per year and the change in the returns to experience as proxies for teacher effort. Third, I look at the mean VA score of new hires as a crude way of investigating any labor-supply effects due to applicants’ knowledge of the new evaluation system.¹

¹If I could have shown that the policy did select for or make better teachers than the previous policy, I would have proceeded to investigate what effects this had on student outcomes. From the current data, it seems like there is not a significant difference in achievement between students of teachers selected for tenure under the new policy and students whose teachers came earlier.

Because the effect of deselection is better measured in terms of the number of tenure decisions made than in the number of school years that have passed with a deselection policy in place, I thought that I might measure the effects of policy better by comparing high- and low-turnover schools. Consistently high-turnover schools are for many reasons likely to be worse schools than low turnover schools, but perhaps the achievement gap would have narrowed because they are more heavily “treated.” With rare exception, however, trends in turnover within schools were not consistent enough to separate schools into high and low categories.
6 Specific Methods and their Results

6.1 City-wide Trends

Perhaps the primary mechanism through which the new policy might be expected to have an effect would be through identifying ineffective teachers and denying them tenure. Although, as discussed earlier, I lack individual-level data on tenure decisions, the city-wide numbers for tenure decisions are released to the press each year by the NYC DOE. I plot the trends in decision-making in Table 4.

The number of teachers denied tenure outright made no discernable jump after the policy change, but the percentage of teachers whose probationary periods were extended increased immediately by nearly 30%, with a corresponding decrease in tenure approvals. This is important, not only as evidence of a behavior change on the part of principals and superintendents, but also because regression discontinuity analysis of the IMPACT program in Washington D.C. schools, which features similar implications of dismissal for teachers with negative evaluations, shows that teachers with poor performance reviews would be more likely to leave on their own (Dee and Wyckoff 2013). Unfortunately, there is no way to see if this is also the case in the data NYC shared with me for this paper.

The ultimate goal of the policy shift, however, is to improve student outcomes. Figures 4 and 5 show the trends in student achievement for 4th and 8th grade students in NYC from 2003 to 2013. Scores on the NYS exams are shown alongside a linear transform of the city-wide average score on the National Assessment of Educational Progress (NAEP), a low-stakes test administered by the U.S. Department of Education to students nationwide every other year. As part of the Trial Urban District Assessment the NAEP has been given to a representative sample of NYC city students roughly every other year since 2003. Because the test has no consequences for students or teachers, and is designed to assess educational progress over time, it is widely considered a truer measure of student learning than high-stakes state exams.

The NAEP scores show some minor improvement over the entire period for students in both grades and in both subjects, but there is no particular acceleration in improvement after the policy change. This makes it seem unlikely that the rather large gains made by students on NYS exams represent a true increase in comprehension. The fact that scores level out between 2010 and 2011 is consistent with gains being caused by teaching to the test, since tests were altered in that year. There is nothing in Figures ?? and ?? that proves that the measured gains on NYS exams are not “real,” but because it is safer not to assume that a certain score corresponds to
the same relative level of comprehension across all grades and years, I transform the scores such that, for each grade and year, the mean score is zero and the standard deviation is one. Note that even if the new policy were very effective at inducing ineffective probationary teachers to leave, Figures ?? and ?? would be unlikely to show any substantial increase in test scores simply because the number of probationary teachers, and therefore the number of students that they teach, is small compared to the total population. In addition, with only three years of testing data after the policy shift, and only two years after the first group of teachers were selected for tenure after the new law, the effects of having consistently better teaching quality would not have had time to accumulate (if the effects were to be cumulative). Because absolute test scores may not be comparable from year to year, and the period after the policy change is short, I look at the policy’s effects on teachers rather than examining its direct impact on students. The best way available to me to assess teacher quality and impact is value-added modelling.

6.2 Value-added Model Estimation

Because there is no commonly agreed-upon best model for estimating teacher value-added, I will present the results for two models, a “mixed-effects” model and a fixed-effects model.

I estimate a mixed effects model that is standard in the literature (see, for example Kane and Staiger (2008)), using a method called Empirical Bayes estimation. This technique is like an OLS fixed-effects model, except that it pulls VA scores estimated with low precision back towards the mean. If it did not, one could expect the teachers with the fewest number of students to end up on the extremes of the VA-score distribution. Assuming random assignment and a normal distribution of teacher ability, it is the best linear unbiased predictor of the impact that a teacher has on average student achievement (Kane and Staiger 2008). The basic model is shown in Equation 5:

\[ A_{ijgt} = \beta_0 + \beta_1 A_{ijg,t-1} + \beta_2 \text{Experience}_{jt} + \beta_3 \text{Experience}^2_{jt} + \delta X_{it} + v_{ijgt} \]

where \[ v_{ijgt} = \mu_j + \theta_{jgt} + \varepsilon_{ijgt} \] (4)

The outcome variable \( A_{ijgt} \) is the test score, in either math or ELA, for student \( i \) under teacher \( j \) in grade \( g \) and year \( t \). It is predicted by the student’s test score in the previous year and a vector of student characteristics \( X_{it} \), which includes sex and
ethnicity, dummy variables for the receipt of Free or Reduced Lunch (FRL) (universal regardless of income in some NYC schools), English language learner (ELL) status, and whether the student has a disability that affects learning. It is common for some models to control for the classroom averages for these same characteristics, so as to account for peer effects, but since I cannot observe in my data which classrooms students were in, I do not.

I choose to control for teacher experience and its square in each year because studies of teachers’ VA consistently show diminishing returns to experience (see Taylor and Tyler 2012). Because of the trend of increasing experience over the period of study, as shown in Figure 2, failing to control for experience results in a trend in VA scores for new teachers that matches the one for percentage of teachers with fewer than three years of experience. Since VA is a comparative measure, entering a more experienced pool as a young teacher will result in a lower measured teacher effectiveness.

The error term is decomposed into persistent teacher effects ($\mu_j$), teacher-grade-year effects ($\theta_{jtg}$) and idiosyncratic student effects ($\varepsilon_{ijgt}$).

The Empirical Bayes part of this model comes from progressively estimating higher levels of organization in the error term using weighted averages of the lower levels. First I take a weighted average of the teacher-grade-year component of the error ($\bar{v}_j$) to estimate the persistent teacher effect $\mu_j$. The average is weighted according to the precision $h_{jgt}$, thus classes with more students will be given more weight. The number of students is represented by $n_{jgt}$, and the $\hat{\sigma}$ are the estimated variances of the different components of error:

$$\bar{v}_j = \sum_t \sum_g w_{jgt} \bar{v}_{jgt} \text{ where } w_{jgt} = \frac{h_{jgt}}{\sum_t \sum_g h_{jgt}}$$

$$h_{jgt} = \left(\frac{\hat{\sigma}_\theta^2 + \hat{\sigma}_\varepsilon^2}{n_{jgt}}\right)^{-1}$$

Finally, I estimate the VA for teacher $j$, which is the estimate for persistent teacher effects, multiplied by a “shrinkage” factor, which is just a weighted average of the classroom residuals.

$$VA_j = \bar{v}_j \left(\frac{\hat{\sigma}_\mu^2}{\hat{\sigma}_\mu^2 + \left(\sum_t \sum_g h_{jgt}\right)^{-1}}\right)$$

Note that the shrinkage factor approaches one as the estimate gets more and
more precise. The less precise it is, the closer it gets to zero, or the VA of the average teacher. This could be considered desirable from a deselection-policy perspective. If there is a cut-off below which principals fire teachers, the inclusion of a shrinkage factor effectively amounts to reserving judgment on teachers with less performance data.

A fixed effects model might be desirable for its simplicity, and, when there are sufficient observations, it yields similar results to the mixed model. My fixed effects model is also very standard. It looks similar to the model used by Taylor and Tyler (2012), and is shown in Equation 8:

\[
A_{ijgt} = \beta_0 + \beta_1 A_{ijg,t-1} + \beta_2 \text{Experience}_{jt} + \beta_3 \text{Experience}^2 + \delta X_{it} + \nu_{ijgt} + \mu_j + \theta_{jgt} + \varepsilon_{ijgt} \tag{8}
\]

The notation is essentially the same. Here a teacher’s VA score is simply taken to be \(\mu_j\). I cluster standard errors on teachers.

In most of my models, I am concerned with persistent teacher effects, but for the section on effort, I consider teacher-year effects. The change in both models is the same. I re-estimate both models with \(\mu_{ij}\) in place of \(\mu_j\) and no \(\theta\) term. For the mixed effects model, this will be the teacher-year portion of the error weighted by its precision. For the fixed effects model it is immediately estimated in the regression. It is mostly in the teacher-year effects, when there are fewer student test score observations for each teacher, that the difference between these two models becomes evident.

Figures 7 and 8 show the VA distributions in Math and ELA for the mixed effects and fixed effects VA scores. As expected, each figure shows the heterogeneity in persistent teacher effects. Because the mixed effects models pull imprecisely measured VA scores back towards the mean, they have shorter tails than the fixed effects models. In each case, the left tail, of below-average teachers, is longer than the right tail. If the new evaluation policy succeeds in identifying those teachers on the leftmost end of the distribution, VA is a reasonable method of assessing teacher quality, and those teachers are denied tenure more often than under the previous policy, then this new policy might be deemed successful.

6.3 Deselection

Reading the various press releases and reports on the policy shift in NYC, it is clear that the mechanism through which it was intended to work was deselection. Appendix
A includes one of the slides shown to principals who attended training for the new tenure-approval system. They were tasked with identifying the least effective teachers, who they believed to be beyond saving, and deny them tenure or suggest that they leave.

For this reason, I start by examining the effects of tenure on the mean VA of teachers that went up for tenure before and after the policy. Figures 6 and 7 show the mean VA of two teacher cohorts in the early years of their careers. I use persistent teacher effectiveness for these graphs, so changes in mean VA are caused by teacher departure. The new policy group is teachers whose probationary periods end in 2011. Teachers whose probationary periods end in 2011 worked in their first two years under the old regime, but were the first to face the new tenure evaluations. The old policy cohort in these graphs is teachers who went up for tenure in 2008, 2009 and 2010, because these are the oldest groups whose entire careers are visible in the data.

These graphs show fairly visible jumps in average VA after the year of the tenure decision for the 2011 cohort in each case and never for the older groups, which is suggestive that the new tenure policy is, in fact, a more effective screening mechanism for ineffective teachers. Given that the standard deviation for VA scores is about 0.2 for mixed effects and 0.3 for fixed effects, the magnitude of some of the jumps is also quite large. As would be expected from their higher standard deviations and slightly lower means, the fixed effects models have slightly more variable and slightly lower means over time. The mean VA for ELA in both the mixed effects and fixed effects model is variable enough that the increase in average VA after tenure under the new policy regime might simply be noise. To better tease out what is happening, I use the following logit model to estimate whether teachers who left under the new policy were any less effective, as determined by VA score, than the teachers who were not.

\[
\ln \left( \frac{P(\text{leave}_j)}{1 - P(\text{leave}_j)} \right) = \beta_0 + \beta_1 \text{NewPolicy} \times VA_j + \beta_2 \text{NewPolicy} + \beta_3 VA_j \quad (9)
\]

Equation 9 assesses what impact the VA score for teacher \( j \) had on whether or not he or she left NYC schools, and via the interaction of VA and the dummy for the new policy, whether VA was a more significant determinant of exit under the new regime. If the new tenure decision-making policy were effective at selecting better teachers, then one would expect \( \beta_1 \) to be negative. If the new policy also increased the quantity of teachers that were denied tenure, then \( \beta_2 \) would be positive.

To examine the general trends in departure, I start by using the odds that a
teacher leaves the NYC DOE as the outcome for Equation 9. I expect, however, that many of the departures are completely exogenous to tenure policy, so I next use the odds a teacher left in the year his or her tenure decision was made as the outcome. It seems reasonable to expect that teachers who were induced to leave by this policy were either forced to by being denied tenure or convinced to by having their probationary period extended. Because I know that tenure decisions are captured poorly in this data, I then use the odds a teacher left during his or her probationary period as my outcome. The results using the mixed effects VA scores are shown for Math in Table 7 and ELA in Tables 8. The results using fixed effects models are not shown because they are similar in sign, magnitude and significance.

Tables 7 and 8 do not show evidence that the new policy more effective than the previous policy at selecting better teachers. The coefficient $\beta_1$ is completely insignificant in all specifications, except for estimating whether math teachers were more likely to leave in the year of their tenure decisions. In this case, it is only marginally significant ($p < 0.1$) and is positive, suggesting that, if anything, math teachers with high VA scores were more likely to be induced to leave under the new policy.

The coefficient on the new policy dummy is significant and negative for both Math and ELA teachers when looking at all leavers, indicating that departure rates decreased after the new policy. The analogous coefficient on the same new policy dummy, however, is positive and significant when used to estimate the odds that a Math teacher left in his or her probationary period. This is consistent with the new policy inducing more probationary teachers to leave in general, but the insignificant point estimate on $\beta_3$ in this regression is positive, which would support the idea that the policy actually induced better teachers to leave.

The only consistently significant result is that teachers with low VA scores in both Math and English were more likely to leave within all time periods.

Although there is no strong evidence thus far to support the claim that the new policy makes tenure decisions a better tool for selecting high-quality teachers, this may be due to looking at noisy departure measures for evaluating the effects of the policy.

### 6.4 Effort

Aside from deselection effects, the new policy may have an impact on teacher VA by encouraging greater effort. Because all teachers face evaluation and the threat of dis-
missal, it may improve the performance of all teachers through increased effort. Since this threat is more credible for probationary teachers, it may lead to a proportionally larger increase in their effort levels. It could also be, however, that the perceived increase in stakes of evaluations under the new regime has been relatively larger for tenured teachers.

In any case, I set out to measure effort in two ways: first, by examining teacher absence data, and second by examining the change in the gradient of returns to experience. Using the absence data, I allow myself the possibility of observing whether or not the new policy has had a larger impact, if any, on the effort of tenured or untenured teachers. Hansen (2010) also uses absence data as a measure of teacher effort in his analysis of North Carolina school teachers. He finds that teachers use fewer sick days in their first years at a school and in the years immediately following the entry of a new principal, which seems to indicate that absences decrease in response to career concerns. Although there are many reasons to be absent that have nothing to do with day-to-day effort levels, Hansen also finds that the trend in the number of sick days closely matches the trend in the self-reported number of hours that a teacher works outside of the classroom.

NYC teachers, regardless of tenure or experience, receive up to ten paid absences each year, although there are some exceptions made for long-term illness. Teachers may also “donate” their allotted sick days to colleagues who have exhausted their own. Although the ten days are only supposed to be used in cases of illness, it is considered acceptable to use up to three days to attend to personal business. Principals have the right to investigate the cause of absences. As is shown in Table 4, the typical NYC teacher takes between six and seven days off each year.

To investigate the effect of the policy change, I regress the number of absences taken annually on a now-familiar set of explanatory variables as shown in Equation 10. Recall that I have eliminated teachers with over fifty absences from the data.

\[
Absences_j = \beta_0 + \beta_1 Tenured_j \times New_j + \beta_2 Tenured_j + \beta_3 New_j + \beta_4 X_j \quad (10)
\]

The results of this regression are given in Table 9, with standard errors cluster on teachers. Column (1) shows the results of a basic OLS regression with some controls. These results seem to suggest that the new policy motivated fewer absences and that the job protection of tenure allows teachers to be absent more. In addition, the new policy seems to have put a wider gulf between tenured and probationary teachers in
terms of effort, although $\beta_1$ is only marginally significant ($p < 0.1$).

When I control for teacher fixed effects in Column (2), however, the results are not consistent with this explanation of events. The sign on $\beta_3$ flips to become positive, indicating that teachers took more days off after the new policy than before, even after controlling for year fixed effects. One might guess that if this is a consequence of heightened fear over job security, the primary consequence seems to be not increased effort, but increased stress, making work less desirable. This, however, is mere speculation. Suffice it to say that the new policy is associated with a slight increase in absences.

Because most of the teachers in the regression shown in Column (2) had tenure over the entire period of observation, I re-estimate the regression using only teachers whose tenure status changed over the course of the observation window (from 2007 to 2012). As is shown in Column (3), the point estimate for $\beta_1$ is actually more positive than in the previous regression, although the two coefficients are not statistically distinguishable. If they were statistically different, it would suggest that less senior teachers, who should arguably have less job security, actually took more additional time off than their more experienced colleagues after the policy. None of this is consistent with the new policy inducing effort.

In both Columns (2) and (3), the sign on $\beta_1$ is negative. This result is highly significant ($p < 0.01$) in Column (2), but becomes insignificant in Column (3) perhaps due to the decrease in sample size. While it is hard to explain this result using just effort or job security, this could be taken as evidence of the new policy denying tenure to teachers that put in less effort into their work. It may also mean that teachers who put little effort into teaching are more likely to leave before receiving tenure.

My previous findings provide little support for this story, but the prospect is certainly worth investigating. To try and eliminate the possibility that selection is driving the results, I rerun the regression in Column (3), this time restricting the population to only teachers who appear in the data in each year from 2009 to 2012. I select this period so that there will be an ample number of teachers who receive tenure before and after the policy.

The results of this regression are displayed in Column (4). The point estimate for $\beta_1$ becomes less negative, but it remains insignificant. It is also not statistically different from the estimate for $\beta_1$ in the unbalanced panel in Column (3). I will speculate no further about insignificant coefficients. The results in Table ?? provide little suggestion that the new policy selected teachers that worked harder, and no evidence that it induced the average teacher to exert more effort, at least so far as
effort is reflected in the number of times a teacher is absent.

Another way to measure effort is in the change in the gradient of returns to experience. The idea here is that, while all teachers can be expected to improve in terms of VA over at least the early years of their careers, teachers who exert more effort in the classroom may improve faster. To investigate this possibility, I use time-variant teacher effects from year to year. Figures 8 and 9 are a preliminary look at the improvement in VA over time for cohorts of teachers that began teaching in NYC in different years. If the mean time-variant VA score increases 2011 cohort more rapidly than it did for earlier cohorts, this would be consistent with the new policy encouraging greater effort from teachers.

Figures 8 and 9 do not show this. Although time-variant VA scores for the 2011 cohort increase in 2012 and the magnitude of this increase is not particularly larger than that of other groups. As expected, the means for the mixed effects models are far more stable than for the fixed effects models. Indeed, the fixed effects models seem almost to show a negative serial correlation between VA scores. Interestingly, the fixed effect VA scores fall in 2011 for every cohort except Math teachers that began working in 2010. I do not know why this is, but it may be because older teachers were better at adjusting to the new testing standards that took effect in the 2011 school year. This trend, however, is not apparent with the mixed effects VA scores.

Figures 11 and 12 do not suggest that I will find an effect of the new policy on the change in the gradient of returns to experience, but in order to be sure, I specify the following model, which looks at the determinants of time-variant VA across years.

\[
VA_{jt} = \beta_0 + \beta_1 VA_{j,t-1} + \beta_2 New_{jt} + \beta_3 Experience_{jt} + X_{jt} \quad (11)
\]

where \(VA_{jt}\) is the time-variant VA score in year \(t\) for teacher \(j\). Because students, and therefore VA, are tested at the end of each year, it is appropriate for all explanatory variables, with the exception of the VA score in the previous year, to be measured in year \(t\). I cluster standard errors on teachers.

I do not want my results to be confounded by the difference in returns to experience at different absolute levels of experience, so I control for experience in my regressions, having already controlled for experience and experience squared in my initial VA estimation. Because I do not expect the trend in gains to experience to be linear,
I run the regression first on every teacher, then on only those teachers that have 15 years of experience or fewer, and finally on those that have 5 or fewer years of experience. Ideally, the estimated regression coefficients will not change much across different experience cohorts, and the coefficient on experience will be consistently insignificant, suggesting that experience is not driving my results. Tables 10 through 13 show the results.

Once again, the results do not suggest that the new policy induced more effort in teachers. Unsurprisingly, the time-variant VA score in the previous year is an important predictor of VA in the reference year. It is a much more important predictor for Math teachers than ELA teachers. Fortunately, VA score in the previous year has a similarly-sized impact in each experience cohort, except perhaps for ELA teacher as measured by the mixed effect model in the ‘fewer than five years of experience’ cohort. For that group, the effect of the prior year’s test score is about a third of the size that it was for the other two ELA mixed effects groups.

Strangely, time-variant VA decreases in years after the policy change for both Math and ELA teachers and in both VAM specifications. This is the exactly opposite of the effect I would expect if the new policy induced greater effort. Furthermore, the coefficients on the tenure dummy are usually insignificant. If teachers increased their efforts in response to a dismissal threat, then the coefficients on the tenure dummies should be negative. Instead, they are insignificant or positive.

After examining the policy’s effect on both the number of days that a teacher takes off, and on the change in returns to experience, I cannot say that the new policy increased the efforts of any teacher.

6.5 Teacher Supply

Because it often comes up in debates about tenure and VAMs, it is worth examining the trends in quality of teachers that NYC has hired, which I take to be indicative of the applicant pool because I have no data on those who applied and were not hired. Some people argue that more rigorous evaluations scare away bad applicants, while others argue that such unfair measures of teacher quality scare most applicants away and reduce competition for employment. Still others argue that teachers need strong job protection to compensate for low pay, and weakening that job protection will restrict supply.

I do not have a strong identification strategy for determining what is causing the decline in VA among newly hired teachers, but Figures 13 and 14 both show the same
trend. Figure 13 shows the mean time-variant VA in the first year of teaching on the vertical axis and the first year of teaching on the horizontal axis. Figure 14 does the analogous thing for persistent teacher effects. Although the VA scores from the fixed effects model bounce around a bit, the trend is clearly downward. While this trend is intriguing, it does not begin nor accelerate after the policy change, so these graphs do not suggest a causal relationship.

The one argument that might be made for a teacher supply effect from these figures is that there was an information lag, such that the policy began to impact teacher supply in 2012 rather than 2011. The increase in mean VA scores between 2011 and 2012, however, is no greater than the oscillations in other years, and it only increasing above the 2010 mean for time-variant scores. I do not pretend that plotting the trends in mean VA scores for new hires would capture the manifold possible ways that the new policy may have affected teacher supply, but, prima facie, the new policy had no obvious impact.

7 Conclusions

Just before the 2011 school year, NYC changed its tenure decision-making system in the hopes of making it harder for ineffective teachers to earn tenure. In so doing, it joined the now-popular trend in education of weakening tenure protection and attempting to make earning it conditional on improving student test scores.

Over the course of the following three years, far fewer teachers were approved for tenure the first time, but there was no corresponding increase in tenure denials. During this same period, the scores of NYC students rose considerably, but no faster than they had been rising for the four years before the policy change. In spite of rising scores on state exams, the scores of NYC students on national exams rose only slightly.

Although the policy was designed to force less effective teachers out, teacher turnover did not increase in the wake of the policy change, nor do I find conclusive evidence that the teachers that did leave were on average less effective than the ones that left before. Problems with the NYC DOE data make it difficult to determine the outcome of tenure decisions, and these issues may have caused me to underestimate the effects of the program.

I also fail to find increases in effort for either new or tenured teachers as measured by the frequency of absences or the change in returns to teaching ability from experience. In fact, the frequency of absences rose slightly after the new policy, perhaps
reflecting a less-enjoyable work environment.

I do note a trend in the decreasing relative effectiveness of teachers hired later by NYC, even though the VA measure of effectiveness employed accounted for experience. This trend, however, begins under the old tenure policy regime, and seems unaffected by the new one.

A number of things could explain why the program was ineffective. There could be a cultural norm in NYC of granting tenure to almost anybody, and the new legislation may have had little impact on that practice. Principals or VAMs may be bad at determining who is effective. Deselected teachers may have been replaced by equally effective ones. Truly though, data issues prevent me from making any solid claims about the real impact of the program. I know that NYC keeps richer and better records, including tenure decisions and ratings under the new evaluation system. Similar datasets to my own have been used successfully in papers such as Rockoff et al. (2012), Rockoff and Speroni (2010), and Atteberry et al. (2013). If researchers can gain access to this data, it will be worthwhile to replicate and extend my analysis.

I was, however, only able to investigate a milder implementation than the one that was originally mandated in New York State in 2010. The full program was implemented this school year, and will likely exacerbate any effects that the program may have had, making them easier for future research to detect.

A number of other states and cities, such as Florida, Delaware, Washington D.C. and Denver, implemented teacher evaluation systems similar to the one in New York State enacted in 2010 at around the same time (Steele, Hamilton and Stecher 2010). Studying these policies and the difference between them should provide a much deeper understanding of how these programs work or do not work.

These programs are an innovative and, for some, attractive way of improving student outcomes and avoiding what seems to be a leniency bias in traditional teacher evaluation systems. If these policies succeed in identifying ineffective teacher and removing them from the classroom, the potential gains are considerable, multiplied over many children over many years of teaching. The costs of firing and replacing a teacher, in contrast, seem trivial when compared to (say) hiring an additional ten teachers to offer smaller classes.

If, however, as it seems in NYC, the effects of these sorts of programs are found to be quite small, a more careful cost-benefit analysis is in order. Currently, New York State and NYC spend millions of dollars on standardized testing and maintaining data systems. NYC spends $3.5 million a year simply on training school administrators.
to use these data systems, and when it estimated VA scores, it spent $200,000 only
the calculations alone. The data systems involved in these programs are huge. In
addition to the monetary costs, there is the opportunity cost of using school time to
take and prepare for these tests. If New York and other states want to use VAMs to
judge the effectiveness of teachers who do not teach Math and ELA in the 3rd to 8th
grades, then administrators will be have to develop additional tests for those teachers,
at the cost of additional time and money. Unlike in-classroom observations, VAMs
currently offer little insight into how teachers might improve their performance. If
administrators value the opportunity to provide specific feedback, the two evaluation
systems are likely to be run in parallel, so schools are unlikely to save money by
eliminating older programs. If policies that tie VA scores to evaluations and continued
employment prove to be effective, then all this may be well worth it, but my results
suggest that the benefit is not substantial.

Even if these programs do not benefit students or teachers, the massive amounts
of data that they require, if made available to researchers, should make the process of
developing better programs easier. The best way to measure teacher effectiveness and
the best thing to do with this information remain difficult and pressing questions. I
do not believe, and my results do not show, that the policy enacted by NYC in 2010
gave the best answers.
Table 1: Summary of APPR Rubric (2010)

<table>
<thead>
<tr>
<th>Rating</th>
<th>State Student Performance Measure</th>
<th>Local Student Performance Measure</th>
<th>Other Local Measure</th>
<th>Composite Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Highly Effective</td>
<td>18-20</td>
<td>18-20</td>
<td></td>
<td>91-100</td>
</tr>
<tr>
<td>Effective</td>
<td>9-17</td>
<td>9-17</td>
<td>0-60</td>
<td>75-90</td>
</tr>
<tr>
<td>Developing</td>
<td>3-8</td>
<td>3-8</td>
<td></td>
<td>65-74</td>
</tr>
<tr>
<td>Ineffective</td>
<td>0-2</td>
<td>0-2</td>
<td></td>
<td>0-64</td>
</tr>
</tbody>
</table>
### Table 2: Data Structure and Sources

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Years</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student-level</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Student Data</td>
<td>2001-2012</td>
<td>Contains individual student test scores and demographic characteristics. It can be requested at <a href="http://schools.nyc.gov/Accountability/data/DataRequests">http://schools.nyc.gov/Accountability/data/DataRequests</a>.</td>
</tr>
<tr>
<td>Linkage Data</td>
<td>2006-2012</td>
<td>For each student and subject, contains his or her teacher’s identification code. It can be requested at <a href="http://schools.nyc.gov/Accountability/data/DataRequests">http://schools.nyc.gov/Accountability/data/DataRequests</a>.</td>
</tr>
<tr>
<td>Teacher-level</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teacher Data</td>
<td>2007-2013</td>
<td>Contains individual teacher demographic characteristics as well as data on salary, experience and attendance. It can be requested at <a href="http://schools.nyc.gov/Accountability/data/DataRequests">http://schools.nyc.gov/Accountability/data/DataRequests</a>.</td>
</tr>
<tr>
<td>School-level</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Report Card Data</td>
<td>2005-2012</td>
<td>Contains a variety of school-performance indicators, such as teacher turnover and demographic information. It is accessible at <a href="https://reportcards.nysed.gov/databasedownload.php">https://reportcards.nysed.gov/databasedownload.php</a>.</td>
</tr>
<tr>
<td>State-City Link Data</td>
<td></td>
<td>Contains both the state-assigned school identification code and city-assigned identification code for each school in NYC. It is available at <a href="http://www.nyc.gov/html/doed/downloads/datasets/DOE_LocationMasterData_001.xls">http://www.nyc.gov/html/doed/downloads/datasets/DOE_LocationMasterData_001.xls</a></td>
</tr>
<tr>
<td>City-level</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Decision Data</td>
<td>2006-2012</td>
<td>Contains the number of tenure decisions made and their outcomes. It has been released in multiple locations by the NYC DOE, including <a href="http://cfn207.org/s/TeacherTenureOverviewforPrincipalsFINAL.pdf">http://cfn207.org/s/TeacherTenureOverviewforPrincipalsFINAL.pdf</a></td>
</tr>
</tbody>
</table>
Table 3: Mock-up of Selected Entries and Variables in Main Dataset

<table>
<thead>
<tr>
<th>Year</th>
<th>Teacher</th>
<th>Sex (static)</th>
<th>Experience (mutable)</th>
<th>Grade</th>
<th>Subject</th>
<th>Student</th>
<th>ELL (static)</th>
<th>Test Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>2005</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>3</td>
<td>Math</td>
<td>Tejesh</td>
<td>No</td>
<td>0.5</td>
</tr>
<tr>
<td>2006</td>
<td>Bakija</td>
<td>Male</td>
<td>2</td>
<td>4</td>
<td>Math</td>
<td>Tejesh</td>
<td>No</td>
<td>0.9</td>
</tr>
<tr>
<td>2006</td>
<td>Bakija</td>
<td>Male</td>
<td>2</td>
<td>4</td>
<td>Math</td>
<td>Marika</td>
<td>No</td>
<td>1</td>
</tr>
<tr>
<td>2007</td>
<td>Bakija</td>
<td>Male</td>
<td>3</td>
<td>4</td>
<td>Math</td>
<td>Carson</td>
<td>Yes</td>
<td>-0.5</td>
</tr>
<tr>
<td>2007</td>
<td>Bakija</td>
<td>Male</td>
<td>3</td>
<td>4</td>
<td>Math</td>
<td>Ivan</td>
<td>No</td>
<td>1</td>
</tr>
<tr>
<td>2008-2011</td>
<td>[Omitted] [Omitted] [Omitted] [Omitted] [Omitted] [Omitted] [Omitted] [Omitted] [Omitted]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2012</td>
<td>Zimmerman</td>
<td>Male</td>
<td>5</td>
<td>8</td>
<td>ELA</td>
<td>Carson</td>
<td>Yes</td>
<td>1.2</td>
</tr>
<tr>
<td>2013</td>
<td>Zimmerman</td>
<td>Male</td>
<td>6</td>
<td>8</td>
<td>ELA</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
</tbody>
</table>

Source

Teacher Data (2007-2013)

Student Data (2005-2006)

Source

Linkage Data (2006-2012)

Linkage Data (2006-2012)

Note: ELL stands for English language learner. I fill in all static characteristics for student or teachers who appear in the data without those characteristics in the raw data in any given year, such as every teacher does in 2006 because I do not have a teacher data file for that year. When experience is missing, it is imputed by adding or subtracting one based on the previous or subsequent year’s experience. Boxes for each raw dataset appear below what columns they contributed to the final dataset.
Table 4: Selected Summary Statistics

<table>
<thead>
<tr>
<th>Year</th>
<th>Student-Subject Teachers N</th>
<th>M.Ed of Higher Ed</th>
<th>Experience</th>
<th>Absences</th>
<th>Students N</th>
<th>FRL</th>
<th>ELL Status</th>
<th>ELA Score</th>
<th>Math Score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>%</td>
<td>ave. years</td>
<td>ave. days</td>
<td></td>
<td></td>
<td></td>
<td>ave. score</td>
<td>ave. score</td>
</tr>
<tr>
<td>2006</td>
<td>772,893</td>
<td>14.972</td>
<td>.</td>
<td>6.6</td>
<td>469,172</td>
<td>59.7</td>
<td>23.4</td>
<td>649</td>
<td>655</td>
</tr>
<tr>
<td>2007</td>
<td>843,212</td>
<td>99,155</td>
<td>75.7%</td>
<td>6.9</td>
<td>461,610</td>
<td>69.0</td>
<td>24.4</td>
<td>651</td>
<td>665</td>
</tr>
<tr>
<td>2008</td>
<td>833,689</td>
<td>99,824</td>
<td>82.2%</td>
<td>7.0</td>
<td>451,425</td>
<td>79.8</td>
<td>21.2</td>
<td>656</td>
<td>672</td>
</tr>
<tr>
<td>2009</td>
<td>852,290</td>
<td>99,811</td>
<td>81.6%</td>
<td>7.3</td>
<td>447,692</td>
<td>64.8</td>
<td>21.1</td>
<td>662</td>
<td>680</td>
</tr>
<tr>
<td>2010</td>
<td>852,059</td>
<td>96,826</td>
<td>85.7%</td>
<td>8.2</td>
<td>446,985</td>
<td>77.1</td>
<td>13.7</td>
<td>662</td>
<td>673</td>
</tr>
<tr>
<td>2011</td>
<td>930,552</td>
<td>94,570</td>
<td>91.0%</td>
<td>8.9</td>
<td>464,995</td>
<td>86.0</td>
<td>21.7</td>
<td>660</td>
<td>680</td>
</tr>
<tr>
<td>2012</td>
<td>800,910</td>
<td>92,972</td>
<td>92.2%</td>
<td>9.0</td>
<td>448,145</td>
<td>84.3</td>
<td>20.0</td>
<td>664</td>
<td>684</td>
</tr>
</tbody>
</table>

Note: "Teachers" appears in quotes when referencing the raw data because the raw data contains all NYC DOE employees, including non-teachers. Averages are over all observations, thus experience, for instance, is not the average experience of teachers, but the average weighted by the number of student-subjects each teacher is linked to that year. FRL stands for Free or Reduced Lunch. ELL stands for English language learner. Statistics are calculated from NYC DOE data.
Table 5: Years Experience at time of Tenure Decision

<table>
<thead>
<tr>
<th>Years Experience</th>
<th>Frequency</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>11</td>
<td>0.15</td>
</tr>
<tr>
<td>2</td>
<td>180</td>
<td>2.52</td>
</tr>
<tr>
<td>3</td>
<td>5,600</td>
<td>78.49</td>
</tr>
<tr>
<td>4</td>
<td>383</td>
<td>5.37</td>
</tr>
<tr>
<td>5</td>
<td>246</td>
<td>3.45</td>
</tr>
<tr>
<td>6</td>
<td>195</td>
<td>2.73</td>
</tr>
<tr>
<td>More than 6</td>
<td>520</td>
<td>7.29</td>
</tr>
</tbody>
</table>

Note: A teacher is expected to have between 2 and 4 years at the time he or she receives tenure.

Table 6: The Number of Tenure Decisions Made Annually

<table>
<thead>
<tr>
<th>Year</th>
<th>All Decisions</th>
<th>Pop. Adjusted</th>
<th>In Data</th>
<th>% Error</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>N</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>2006</td>
<td>6,112</td>
<td>2,622</td>
<td>1,417</td>
<td>-46.0%</td>
</tr>
<tr>
<td>2007</td>
<td>6,250</td>
<td>2,638</td>
<td>1,840</td>
<td>-30.3%</td>
</tr>
<tr>
<td>2008</td>
<td>6,115</td>
<td>2,529</td>
<td>1,743</td>
<td>-31.1%</td>
</tr>
<tr>
<td>2009</td>
<td>5,859</td>
<td>2,423</td>
<td>2,078</td>
<td>-14.2%</td>
</tr>
<tr>
<td>2010</td>
<td>6,292</td>
<td>2,494</td>
<td>2,234</td>
<td>-10.4%</td>
</tr>
<tr>
<td>2011</td>
<td>5,209</td>
<td>1,369</td>
<td>1,957</td>
<td>43.0%</td>
</tr>
<tr>
<td>2012</td>
<td>3,954</td>
<td>988</td>
<td>981</td>
<td>-0.7%</td>
</tr>
</tbody>
</table>

Note: The numbers for all decisions come from press releases issued by the NYC DOE. The “Pop. Adjusted,” population adjusted, numbers account for the fact that my sample includes only grades 3-8 math and ELA teachers. The number of decisions in my data are calculated using a variable for probation completion date.
Table 7: Predicting Induced Turnover for Math Teachers

<table>
<thead>
<tr>
<th>VARIABLES</th>
<th>(1) \ P(Leave)</th>
<th>(2) P(Leave on Decision)</th>
<th>(3) P(Leave in Probation)</th>
</tr>
</thead>
<tbody>
<tr>
<td>New Policy * VA</td>
<td>-0.379</td>
<td>1.525*</td>
<td>0.0859</td>
</tr>
<tr>
<td></td>
<td>(0.290)</td>
<td>(0.908)</td>
<td>(0.581)</td>
</tr>
<tr>
<td>New Policy</td>
<td>-0.470***</td>
<td>-0.158</td>
<td>0.457***</td>
</tr>
<tr>
<td></td>
<td>(0.0793)</td>
<td>(0.244)</td>
<td>(0.154)</td>
</tr>
<tr>
<td>Math M.E. VA</td>
<td>-0.723***</td>
<td>-1.388***</td>
<td>-1.459***</td>
</tr>
<tr>
<td></td>
<td>(0.124)</td>
<td>(0.431)</td>
<td>(0.344)</td>
</tr>
<tr>
<td>Constant</td>
<td>-1.772***</td>
<td>-3.959***</td>
<td>-2.766***</td>
</tr>
<tr>
<td></td>
<td>(0.453)</td>
<td>(0.120)</td>
<td>(1.047)</td>
</tr>
<tr>
<td>New Policy * VA</td>
<td>1.525*</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.908)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Observations          | 16,431         | 8,130                    | 8,274                     |
Teacher Characteristics | Yes            | Yes                      | Yes                       |

Note: Robust standard errors are in parentheses (*** p<0.01, ** p<0.05, * p<0.1). P(Leave) is the probability that a teacher leaves the NYC DOE. P(Leave on Decision) is the probability that a teacher leaves in the year of his or her tenure decision. P(Leave in Probation) is the probability that a teacher leaves in any year of his or her probationary period. Controls for teacher characteristics include race and sex dummies.
Table 8: Predicting Induced Turnover for ELA Teachers

<table>
<thead>
<tr>
<th>VARIABLES</th>
<th>(1) P(Leave)</th>
<th>(2) P(Leave on Decision)</th>
<th>(3) P(Leave in Probation)</th>
</tr>
</thead>
<tbody>
<tr>
<td>New Policy * VA</td>
<td>-0.300</td>
<td>1.331</td>
<td>-0.0471</td>
</tr>
<tr>
<td></td>
<td>(0.308)</td>
<td>(0.849)</td>
<td>(0.573)</td>
</tr>
<tr>
<td>New Policy</td>
<td>-0.557***</td>
<td>-0.134</td>
<td>0.234</td>
</tr>
<tr>
<td></td>
<td>(0.0785)</td>
<td>(0.232)</td>
<td>(0.161)</td>
</tr>
<tr>
<td>ELA M.E. VA</td>
<td>-0.922***</td>
<td>-1.611***</td>
<td>-1.542***</td>
</tr>
<tr>
<td></td>
<td>(0.135)</td>
<td>(0.367)</td>
<td>(0.292)</td>
</tr>
<tr>
<td>Constant</td>
<td>-2.321***</td>
<td>-3.903***</td>
<td>-3.240***</td>
</tr>
<tr>
<td></td>
<td>(0.535)</td>
<td>(0.112)</td>
<td>(0.984)</td>
</tr>
<tr>
<td>Observations</td>
<td>17,011</td>
<td>8,263</td>
<td>8,410</td>
</tr>
<tr>
<td>Teacher Characteristics</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Note: Robust standard errors are in parentheses (*** p<0.01, ** p<0.05, * p<0.1). P(Leave) is the probability that a teacher leaves the NYC DOE. P(Leave on Decision) is the probability that a teacher leaves in the year of his or her tenure decision. P(Leave in Probation) is the probability that a teacher leaves in any year of his or her probationary period. Controls for teacher characteristics include race and sex dummies.
Table 9: Teacher Absences

<table>
<thead>
<tr>
<th>VARIABLES</th>
<th>(1) Absences</th>
<th>(2) Absences</th>
<th>(3) Absences</th>
<th>(4) Absences</th>
</tr>
</thead>
<tbody>
<tr>
<td>New Policy * After Tenure</td>
<td>0.242*</td>
<td>-0.618***</td>
<td>-0.315</td>
<td>-0.0850</td>
</tr>
<tr>
<td></td>
<td>(0.136)</td>
<td>(0.176)</td>
<td>(0.296)</td>
<td>(0.422)</td>
</tr>
<tr>
<td>New Policy</td>
<td>-1.232***</td>
<td>0.995***</td>
<td>1.405***</td>
<td>0.770**</td>
</tr>
<tr>
<td></td>
<td>(0.141)</td>
<td>(0.177)</td>
<td>(0.270)</td>
<td>(0.384)</td>
</tr>
<tr>
<td>After Tenure</td>
<td>1.336***</td>
<td>0.571***</td>
<td>0.177</td>
<td>0.186</td>
</tr>
<tr>
<td></td>
<td>(0.0883)</td>
<td>(0.117)</td>
<td>(0.160)</td>
<td>(0.222)</td>
</tr>
<tr>
<td>Constant</td>
<td>17.31***</td>
<td>6.320***</td>
<td>6.051***</td>
<td>5.860***</td>
</tr>
<tr>
<td></td>
<td>(2.531)</td>
<td>(0.0826)</td>
<td>(0.0922)</td>
<td>(0.137)</td>
</tr>
<tr>
<td>Observations</td>
<td>60,977</td>
<td>61,006</td>
<td>26,557</td>
<td>7,940</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.015</td>
<td>0.005</td>
<td>0.005</td>
<td>0.004</td>
</tr>
<tr>
<td>Teacher Characteristics</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Student Characteristics</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Year FE</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Number of Teachers</td>
<td>19,833</td>
<td>8,038</td>
<td>1,985</td>
<td></td>
</tr>
</tbody>
</table>

Note: Standard errors, which are clustered on teachers, are reported in parentheses (*** p<0.01, ** p<0.05, * p<0.1). Controls for teacher characteristics include race, sex and ‘M.Ed or above’ dummies as well as log of income. Student characteristic controls include variables for the percentage of students with disabilities, FRL receipt and ELL status that a teacher taught in a given year. Column (1) is a basic OLS regression, Columns (2) through (4) include teacher fixed effects. Column (3) restricts the population of teachers to those whose tenure status has changed between 2007 and 2012. Column (4) restricts the population further to create a balanced panel from 2009 to 2012.
Table 10: Year-to-Year Change in Math M.E. VA

<table>
<thead>
<tr>
<th>VARIABLES</th>
<th>(1) Time-Variant M.E. VA</th>
<th>(2) Time-Variant M.E. VA</th>
<th>(3) Time-Variant M.E. VA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prior-year VA Score</td>
<td>0.161***</td>
<td>0.159***</td>
<td>0.155***</td>
</tr>
<tr>
<td></td>
<td>(0.00562)</td>
<td>(0.00584)</td>
<td>(0.00850)</td>
</tr>
<tr>
<td>New Policy</td>
<td>-0.0184***</td>
<td>-0.0194***</td>
<td>-0.0159**</td>
</tr>
<tr>
<td></td>
<td>(0.00368)</td>
<td>(0.00396)</td>
<td>(0.00624)</td>
</tr>
<tr>
<td>After Tenure</td>
<td>-0.000969</td>
<td>-0.00250</td>
<td>0.00417</td>
</tr>
<tr>
<td></td>
<td>(0.00279)</td>
<td>(0.00299)</td>
<td>(0.00371)</td>
</tr>
<tr>
<td>Experience (yrs)</td>
<td>-0.000535</td>
<td>0.000258</td>
<td>-0.00340**</td>
</tr>
<tr>
<td></td>
<td>(0.000398)</td>
<td>(0.000646)</td>
<td>(0.00145)</td>
</tr>
<tr>
<td>Constant</td>
<td>0.202</td>
<td>0.339**</td>
<td>0.584***</td>
</tr>
<tr>
<td></td>
<td>(0.154)</td>
<td>(0.170)</td>
<td>(0.211)</td>
</tr>
<tr>
<td>Observations</td>
<td>28,648</td>
<td>24,728</td>
<td>12,196</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.722</td>
<td>0.718</td>
<td>0.719</td>
</tr>
<tr>
<td>Year FE</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Teacher Characteristics</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Student Characteristics</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Max Experience</td>
<td>50</td>
<td>15</td>
<td>5</td>
</tr>
</tbody>
</table>

Note: Standard errors, which are clustered on teachers, are reported in parentheses (*** p<0.01, ** p<0.05, * p<0.1). Populations are progressively restricted to lower levels of experience. In Column (3), for instance, the maximum level of experience allowed in the population is 5 years. Controls for teacher characteristics include race, sex and 'M.Ed or above' dummies as well as log of income. Student characteristic controls include variables for the percentage of students with disabilities, FRL receipt and ELL status that a teacher taught in a given year.
<table>
<thead>
<tr>
<th>VARIABLES</th>
<th>(1) Time-Variant F.E. VA</th>
<th>(2) Time-Variant F.E. VA</th>
<th>(3) Time-Variant F.E. VA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prior-year VA Score</td>
<td>0.284***</td>
<td>0.277***</td>
<td>0.292***</td>
</tr>
<tr>
<td></td>
<td>(0.00590)</td>
<td>(0.00628)</td>
<td>(0.00872)</td>
</tr>
<tr>
<td>New Policy</td>
<td>-0.0623***</td>
<td>-0.0627***</td>
<td>-0.0563***</td>
</tr>
<tr>
<td></td>
<td>(0.00486)</td>
<td>(0.00525)</td>
<td>(0.00812)</td>
</tr>
<tr>
<td>After Tenure</td>
<td>0.0165***</td>
<td>0.00500</td>
<td>0.0104*</td>
</tr>
<tr>
<td></td>
<td>(0.00406)</td>
<td>(0.00433)</td>
<td>(0.00543)</td>
</tr>
<tr>
<td>Experience (yrs)</td>
<td>0.00148**</td>
<td>0.00684***</td>
<td>-0.00228</td>
</tr>
<tr>
<td></td>
<td>(0.000610)</td>
<td>(0.000960)</td>
<td>(0.00219)</td>
</tr>
<tr>
<td>Constant</td>
<td>1.228***</td>
<td>2.027***</td>
<td>2.545***</td>
</tr>
<tr>
<td></td>
<td>(0.232)</td>
<td>(0.254)</td>
<td>(0.312)</td>
</tr>
<tr>
<td>Observations</td>
<td>27,500</td>
<td>23,743</td>
<td>11,678</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.677</td>
<td>0.677</td>
<td>0.683</td>
</tr>
<tr>
<td>Year FE</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Teacher Characteristics</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Student Characteristics</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Max Experience</td>
<td>50</td>
<td>15</td>
<td>5</td>
</tr>
</tbody>
</table>

Note: Standard errors, which are clustered on teachers, are reported in parentheses (*** p<0.01, ** p<0.05, * p<0.1). Populations are progressively restricted to lower levels of experience. In Column (3), for instance, the maximum level of experience allowed in the population is 5 years. Controls for teacher characteristics include race, sex and ‘M.Ed or above’ dummies as well as log of income. Student characteristic controls include variables for the percentage of students with disabilities, FRL receipt and ELL status that a teacher taught in a given year.
Table 12: Year-to-Year Change in ELA M.E. VA

<table>
<thead>
<tr>
<th>VARIABLES</th>
<th>(1) Time-Variant M.E. VA</th>
<th>(2) Time-Variant M.E. VA</th>
<th>(3) Time-Variant M.E. VA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prior-year VA Score</td>
<td>0.0337***</td>
<td>0.0301***</td>
<td>0.0180***</td>
</tr>
<tr>
<td></td>
<td>(0.00457)</td>
<td>(0.00473)</td>
<td>(0.00651)</td>
</tr>
<tr>
<td>New Policy</td>
<td>-0.0763***</td>
<td>-0.0772***</td>
<td>-0.0756***</td>
</tr>
<tr>
<td></td>
<td>(0.00286)</td>
<td>(0.00304)</td>
<td>(0.00467)</td>
</tr>
<tr>
<td>After Tenure</td>
<td>0.00323</td>
<td>0.00289</td>
<td>0.00218</td>
</tr>
<tr>
<td></td>
<td>(0.00215)</td>
<td>(0.00229)</td>
<td>(0.00300)</td>
</tr>
<tr>
<td>Experience (yrs)</td>
<td>0.000458</td>
<td>0.000646</td>
<td>0.000829</td>
</tr>
<tr>
<td></td>
<td>(0.000327)</td>
<td>(0.000519)</td>
<td>(0.00122)</td>
</tr>
<tr>
<td>Constant</td>
<td>0.0838</td>
<td>0.126</td>
<td>0.0867</td>
</tr>
<tr>
<td></td>
<td>(0.124)</td>
<td>(0.136)</td>
<td>(0.181)</td>
</tr>
<tr>
<td>Observations</td>
<td>28,396</td>
<td>24,698</td>
<td>11,825</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.795</td>
<td>0.796</td>
<td>0.795</td>
</tr>
<tr>
<td>Year FE</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Teacher Characteristics</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Student Characteristics</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Max Experience</td>
<td>50</td>
<td>15</td>
<td>5</td>
</tr>
</tbody>
</table>

Note: Standard errors, which are clustered on teachers, are reported in parentheses (*** p<0.01, ** p<0.05, * p<0.1). Populations are progressively restricted to lower levels of experience. In Column (3), for instance, the maximum level of experience allowed in the population is 5 years. Controls for teacher characteristics include race, sex and ‘M.Ed or above’ dummies as well as log of income. Student characteristic controls include variables for the percentage of students with disabilities, FRL receipt and ELL status that a teacher taught in a given year.
Table 13: Year-to-Year Change in ELA F.E. VA

<table>
<thead>
<tr>
<th>VARIABLES</th>
<th>(1) Time-Variant F.E. VA</th>
<th>(2) Time-Variant F.E. VA</th>
<th>(3) Time-Variant F.E. VA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prior-year VA Score</td>
<td>0.572***</td>
<td>0.569***</td>
<td>0.532***</td>
</tr>
<tr>
<td></td>
<td>(0.0110)</td>
<td>(0.0116)</td>
<td>(0.0173)</td>
</tr>
<tr>
<td>New Policy</td>
<td>1.623***</td>
<td>1.612***</td>
<td>1.456***</td>
</tr>
<tr>
<td></td>
<td>(0.0391)</td>
<td>(0.0413)</td>
<td>(0.0630)</td>
</tr>
<tr>
<td>After Tenure</td>
<td>0.0278***</td>
<td>0.0198***</td>
<td>-0.00972</td>
</tr>
<tr>
<td></td>
<td>(0.00588)</td>
<td>(0.00640)</td>
<td>(0.00876)</td>
</tr>
<tr>
<td>Experience (yrs)</td>
<td>-0.000271</td>
<td>0.00201**</td>
<td>0.0141***</td>
</tr>
<tr>
<td></td>
<td>(0.000415)</td>
<td>(0.000794)</td>
<td>(0.00324)</td>
</tr>
<tr>
<td>Constant</td>
<td>-2.274***</td>
<td>-2.272***</td>
<td>-2.173***</td>
</tr>
<tr>
<td></td>
<td>(0.0357)</td>
<td>(0.0375)</td>
<td>(0.0549)</td>
</tr>
<tr>
<td>Observations</td>
<td>26,450</td>
<td>22,824</td>
<td>10,872</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.509</td>
<td>0.509</td>
<td>0.487</td>
</tr>
<tr>
<td>Year FE</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Teacher Characteristics</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Student Characteristics</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Max Experience</td>
<td>50</td>
<td>15</td>
<td>5</td>
</tr>
</tbody>
</table>

Note: Standard errors, which are clustered on teachers, are reported in parentheses (** * p<0.01, ** p<0.05, * p<0.1). Populations are progressively restricted to lower levels of experience. In Column (3), for instance, the maximum level of experience allowed in the population is 5 years. Controls for teacher characteristics include race, sex and ‘M.Ed or above’ dummies as well as log of income. Student characteristic controls include variables for the percentage of students with disabilities, FRL receipt and ELL status that a teacher taught in a given year.
Figure 1: Comparison of Overall Turnover Trends

Note: Data from NYS only includes NYC schools
Figure 2: Comparison of Turnover Trends for Teachers with < 3 Years of Experience

Figure 3: Percent of Teachers with < 5 Years of Experience

Note: Data from NYS only includes NYC schools
Figure 4: Tenure Decisions over Time

Note: Author's compilation of data from official NYC press releases
Figure 5: Average 4th and 8th Grade Math Scores on NAEP and NYS Exams

Figure 6: Average 4th and 8th Grade ELA Scores on NAEP and NYS Exams

Note: Average NAEP scores come from a representative sample of only NYC students.
Figure 7: Distributions of Math VA Scores

Top graph:
- Mean: .001
- Standard Deviation: 0.234
- Min: -1.029
- Max: 1.103

Bottom graph:
- Mean: -.162
- Standard Deviation: 0.311
- Min: -1.876
- Max: 1.109
Figure 8: Distributions of ELA VA Scores
Figure 9: Mean Math VA Score by Year in Career
Figure 10: Mean ELA VA Score by Year in Career
Figure 11: Early gains in VA for Math Teachers

Mean Math M.E. Time-Variant VA Score by Year in Career

Mean Math F.E. Time-Variant VA Score by Year in Career
Figure 12: Early gains in VA for ELA Teachers

Mean ELA M.E. Time-Variant VA Score by Year in Career

Mean ELA F.E. Time-Variant VA Score by Year in Career

First year 2008
First year 2009
First year 2010
First year 2011
Figure 13: Time-Variant VA Score in First Year
Figure 14: Persistent VA Score in First Year

Mean Math VA Score for New Hires by Year

Note: Author's calculations using NYC DOE data

Mean ELA VA Score for New Hires by Year

Note: Author's calculations using NYC DOE data
10 References


Slentz, K. and J. D’Agati (2013). Adoption of the Proposed Amendments Relating to the Definitions of Teacher or Principal Growth Percentile Score and Value-Added Growth Score.

Song, J. (2009, 3 May). Firing teachers can be a costly and tortuous task. *Los Angeles Times*.


Figure 15: The Principal’s Task

A priority for New York City is to ensure that all students graduate high school college and career ready. We can help to realize this priority by improving student outcomes and teacher practice.

1. Retain and leverage the most effective educators.
2. Boost effectiveness of all teachers.
3. Improve or counsel out persistently less effective educators.

Teacher Effectiveness (Student Outcomes, Instructional Practice)

Note: This is a slide taken from a NYC DOE principal training session on teacher tenure.
Note: This rubric is identical to the rubrics used in 2010-2011 and 2011-2012. “Growth Scores” are equivalent to VA scores.