Three Essays on the Economics of Education

Tong Geng

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2018

ABSTRACT

Three Essays on the Economics of Education

Tong Geng

This dissertation consists of three essays studying the impact of school organization, incentives, and complementarity on education production. The identification strategy relies on exogenous variation generated from several education policies in New York City, the largest school district in the United States, and the key outcomes include students' standardized test scores and subjective evaluation of their educational experiences.

The first chapter examines the complementarity of incentives in education production. Many production activities require cooperation between agents in an organization, and incentive alignment may take advantage of complementarities in such activities. This paper investigates such a possibility by examining two education policies that were implemented in New York City: a grade retention policy that incentivizes students and an accountability scheme that incentivizes schools. I employ double- and triple-difference strategies to estimate the individual and combined effects of these policies. The policies alone appear to have generated either modest or insignificant improvements in student outcomes. Combined, however, the retention and accountability policies led to a substantial increase in math test scores and reductions in student absences and suspension rates; the effect on English test scores is positive but not robust. These results underscore the value of using incentive alignment to realize complementarities in organizations.

The second chapter, co-authored with Jonah Rockoff, looks at the effect of repeating a grade on students' test scores and subjective evaluation of their educational experiences. When a student's academic knowledge or preparation is well below that of his or her age group, a common policy response is to have that student repeat a grade level and join the following, younger cohort. Evaluating the impacts of grade retention is made complicated by the potential incomparability of (1) retained students to promoted peers and (2) outcomes measured differently across grade levels. In this paper, we use novel data from New York City to ask whether parents' and students' self-reported educational experiences are significantly affected by grade retention. We take advantage of surveys that ask the same questions regardless of a student's grade level, and implement a regression discontinuity approach, identifying causal effects on students retained due to missed cutoffs on math and English exams. We find that parental satisfaction with the quality of their child's education and students' sense of personal safety both improve significantly over the three years we observe from the time of retention. Our findings suggest that the stringent and somewhat controversial test-based retention policies enacted in New York had positive effects on the educational experience of these marginal students.

The third chapter reviews and reassesses the overall impact of Children First, which consists of a series of educational policies during Bloomberg's administration in New York City. To expand our understanding of Children First, I first outline the key components of this education reform and review the literature on Children First and its associated policies. I also reassess the overall impact of Children First through the synthetic control method and find weak effects of this reform on student

performance. Lastly, I provide an economic analysis to understand the advantages and weaknesses of Children First.

Contents

List of Figures

List of Tables

Acknowledgments

I am grateful to my advisors for their generous support throughout my study at Columbia University. Miguel Urquiola provided critical guidance during my early years in the program, and has been supportive of me throughout the whole process. Bentley MacLeod has always been available to inspire me and greatly broadened my thinking as an economist. Jonah Rockoff provided me with tremendous support and encouragement. My experience of working with him shaped my development as an empirical economist and greatly improved my dissertation.

I thank my family and Yi for their unconditional support and love throughout the process. I could not have done this without them. My family has always been there to encourage me during difficult times throughout my study at Columbia. I thank my girlfriend, Yi, for doing everything she could to help me. Her encouragement and support made all the difference.

I am grateful to Peter Bergman, Michael Best, Alex Eble, Michael Gilraine, Wojciech Kopczuk, Randall Reback, Evan Riehl, Miikka Rokkanen, Mandy Shen, and participants of the Columbia Applied Microeconomics Colloquium for helpful discussions. I am grateful to Jonah Rockoff and New York City Department of Education for making the New York City data available for my dissertation.

Dedication

To my family and Yi.

Chapter 1

The Complementarity of Incentive Policies in Education: Evidence from New York City

1.1 Introduction

Organizations frequently adopt incentive policies to motivate agents to reach certain goals. Attaining these goals often requires coordination between multiple agents, which can potentially lead to complementarities between different incentive policies (Holmstrom and Milgrom, 1994). Such complementarities are often overlooked but can be important for efficient production. In education, where instruction typically requires collaboration between staff and students, combining incentive policies might take advantage of a potential complementarity in human capital production.

This paper investigates such a possibility by examining two types of commonly enacted incentive policies in education: an accountability scheme focused on school-side incentives and a grade retention policy emphasizing student-side incentives.[1] These two types of incentive policies may produce complementary effects if there is a complementarity between school effort and student effort in human capital production.

[1]The No Child Left Behind Act of 2002 required each state to bring students to a certain proficiency level. As a result, many states adopted accountability schemes to increase schools' efforts to improve students' test scores. In addition, sixteen states implemented grade retention policies (Rose, 2012), which may motivate students to exert more effort to avoid being retained.

This complementarity would appear if, for example, better prepared instructors are more effective at improving more attentive students' test scores. To my knowledge, no previous study has examined this complementarity, despite the great number of studies evaluating each type of policy in isolation. The lack of evidence may reflect that the identification of such a complementarity is challenging: It requires a suitable overlap of the two arguably exogenous policies, so that their individual and combined effects can both be estimated (Almond and Mazumder, 2013; Athey and Stern, 1998).

In the current paper, I take advantage of the staggered implementation of two policy reforms in New York City (NYC), which allows for estimation of their individual and combined effects. In 2004, NYC started implementing a grade retention policy on a subset of students in several grade levels, which required them to demonstrate a minimum proficiency level on standardized tests in both math and English Language Arts (English) to advance to the next grade. In 2007, NYC initiated an accountability scheme for all schools that placed additional weight on the performance of certain low-achieving students within each grade and school. Schools that were rated poorly under this system faced risk of closure. I employ double- and triple-difference strategies to estimate the individual and combined effects of these policies.[2]

My empirical analysis begins with the retention policy alone (prior to the introduction of the accountability scheme). The control group combines students who were exempt from the policy (special education students and English language learn-

[2] NYC is the largest school district in the United States, and this paper uses administrative data that include several key variables: (1) standardized math and English test scores; (2) students' days of absence and suspension, which can be used to approximate student effort; and (3) their assigned teachers' experience levels and days of absence, which can be used to capture an important part of school effort/resources.

ers) and students with high prior test scores, who faced little risk of failing the test. Using a synthetic control method and a difference-in-difference strategy, I find an improvement in at-risk students' math test scores (10% of a standard deviation) but no significant effects on other outcomes.

The analysis then turns to the effects of the accountability scheme alone by focusing on grade levels that were not subject to the retention policy.[3] Although the scheme awarded points for improvements in all students' test scores, NYC assigned more weight to improvements in the test scores of students who scored in the lowest third in each subject, grade, and school, and the city provided schools with a list of such students. This "lowest third" element of the accountability scheme allows me to investigate the effects of additional incentives on schools' allocation of effort by comparing lowest-third students with top-two-thirds students within each subject, grade, and school. The results show a relative drop in math-lowest-third students' math test scores (10% of a standard deviation) and a relative increase in English-lowest-third students' English test scores (4% of a standard deviation).[4] The effects on other outcomes are small and mostly insignificant.

Last but not least, the analysis examines the complementarity of the two policies, focusing on lowest-third students who were also subject to the retention policy using a triple-difference model. I find that math-lowest-third students who were subject to the retention policy exhibit a large improvement in math test scores (34% of a

[3]Since the policy retained more low-achieving students, changes in student composition are a potential concern for the analysis. However, these changes do not seem to be influencing the results.

[4]Throughout the paper, I refer to students in the lowest third in math as math-lowest-third students and those in the lowest third in English as English-lowest-third students. These groups are correlated but different.

standard deviation) and a decrease in both absences (0.48 days) and suspension rates (0.68 percentage points) when both the retention and accountability policies were in place.[5] Distributional analyses suggest that part of the estimated effect on lowest-third students comes at the expense of higher-achieving students. The analysis of English-lowest-third students suggests a positive and smaller effect on English test scores (8% of a standard deviation), but it is not robust. This finding is consistent with the overall small and insignificant effects of each policy in isolation on English.

Alignment of student and teacher effort may explain the complementarity of these policies. A decrease in math-lowest-third students' absences and suspension rates suggests their increased effort. The distributional effects also suggest that teachers may have allocated more effort/attention to math-lowest-third students. Additionally, there is no evidence that lowest-third students were assigned to teachers with more experience or fewer absences, or to smaller classes, under these policies. All of these findings support the interpretation that student and teacher behaviors are driving the results.

This paper depicts incentive alignment as a potential instrument for taking advantage of organizational complementarities, and it contributes to a small but growing literature on organizational practices and complementarities in schools (Bloom et al., 2015; Jacob and Rockoff, 2012; Mbiti et al., 2016) and a larger literature on organizational complementarities in other settings (Brynjolfsson and Milgrom, 2013;

[5] The complementary effect on math test scores seems quite large compared with effects found in several related studies on school accountability schemes, which represented roughly 10% to 15% of a standard deviation (e.g., Neal and Schanzenbach, 2010; Rockoff and Turner, 2010). Unlike these studies, which estimate the overall effect of a scheme, this paper examines a relative change induced by the lowest-third element and does not distinguish a likely shift of effort across students.

Milgrom and Roberts, 1995). The results support the importance of complementarity between student effort and school/teacher effort in human capital production and underscore the importance of jointly considering all agents' incentives in designing effective education policies.[6]

The rest of the paper proceeds as follows. Section 2 describes the retention policy and the accountability scheme in greater detail. Section 3 describes the data. Sections 4, 5, and 6 present the empirical strategy and results for the retention policy, the accountability scheme, and their interaction, respectively. Section 7 concludes the paper with a discussion of the findings and their implications.

1.2 Background

This section presents background information on each policy and how it motivates the empirical strategy used to identify the policies' individual and combined effects. In 2004, NYC started implementing a grade retention policy that required a subset of students in some grade levels to attain a minimal proficiency level in both mathematics and English to advance to the next grade. In 2007, NYC initiated a school accountability scheme in all public schools that associated rewards and punishments with students' test scores; one element of the scheme assigned additional weight to the performance of lowest-third students in each subject, grade, and school.

[6] A notable study with related findings was conducted by Behrman et al. (2015), who found a greater impact from providing both individual and group monetary incentives to students, teachers, and school administrators than from providing only individual incentives to students and teachers through a social experiment in 88 Mexican high schools. Another study that shares the spirit of this paper is by Johnson and Jackson (2017), who found that Head Start and school financing reforms are complementary in human capital production.

Retention Policy

NYC implemented a grade retention policy for all general education students in 3rd grade in 2004, 5th grade in 2005, 7th grade (English only) in 2006, 7th grade (English and math) in 2007, and 8th grade in 2009 (McCombs, Kirby, and Mariano, 2009).[7] The retention policy required students to achieve a proficiency level of 2 out of 4 on both math and English tests, which all students between 3rd and 8th grade in NYC public schools are required to take in spring.[8] Students in English language learner (ELL) programs, special education programs, and charter schools were exempt from this policy.

NYC also provided all students in high-stakes grades the opportunity to attend Saturday schools, a program specifically focused on test preparation, regardless of their exemption status and prior test scores. In practice, 16% of students attended this program, and they attended 40% of sessions (ibid.). Among attendees, one third of the students were actually at risk of failing the tests (with prior test scores below 3), another third had test scores above 3, and the final third were students exempt from the retention policy.

Students who failed to achieve the minimal proficiency level on the spring tests were required to attend summer school and pass the tests in August in order to be promoted to the next grade. Students who failed the tests in spring or August could

[7] All years refer to the year of the spring semester.

[8] Students in 8th grade were also required to pass tests in science and social studies, which only 4th and 8th graders take. In addition, 8th graders who are overage or who have been previously retained in middle school may be promoted on appeal in August if they demonstrate effort toward meeting the promotion standards.

also be promoted if they were able to demonstrate sufficient proficiency through their portfolios and coursework to their teachers and principals, who made all retention decisions. An appeal process was available for these students and their parents.

This paper hereafter limits the analysis to students in 4th, 5th, and 6th grades, with 5th grade being a high-stakes grade for the retention policy. Third grade did not count toward a major component in the accountability scheme and is thus excluded (see next section for more detail). The accountability scheme and another major change confound the analysis of the retention policy in isolation on 7th and 8th grades. Since NYC implemented the scheme in 2007, evaluating the retention policy for 7th grade (in math) and 8th grade is confounded. In 2006, two major policy changes occurred, which confounds the analysis of the retention policy for 7th grade (English). First, NYC stopped using the city tests for 3rd, 5th, 6th, and 7th grade and adopted the New York State Tests; Students in 4th and 8th grade had been taking the state tests since 1999. Second, New York state accountability measure (as part of No Child Left Behind) was extended to 3rd, 5th, 6th, and 7th grade; the other two grades were subject to the state accountability measure since 2004.[9]

To demonstrate the effects of the policy change, I show that (1) retention risks conditional on failing the tests increased after the policy in a regression-discontinuity design, and (2) the increase occurred exactly at the time when the policy was implemented in a time-series analysis.

[9]The state accountability measure was based on an index that counts twice the number of students who had a test score above 3 and counts twice the number of students who had a test score above 2. The state also required schools with low indexes to take certain actions. However, this state-level policy does not seem to have any large empirical impact on my analysis.

Figure 1.1 shows that if students subject to the policy failed the test, their probability of retention increased after the retention policy. The x-axis is an index that measures the distance between a student's spring test score and the cutoff score for passing the test: $index_{ist} = score_{ist} - \text{cutoff}_{gst}$, where $score_{ist}$ is student i's April test score in subject s in year t and cutoff_{gst} is the cutoff score in subject s for passing the test in year t and grade g. Failing a test is equivalent to $index_{ist} < 0$ and is indicated by the gray vertical line in the figure. Prior to the grade retention policy, retention risks were overall low and loosely connected to failing the tests; the policy increased the probability of retention at the cutoff by 20% in math and 10% in English, which indicates that a typical student saw passing math as more binding than passing English in the promotion standard.

Figure 1.2 converts Figure 1.1 into a time series and shows that the increase in retention risks occurred in 2005, when the policy was implemented. Each point restricts the observations to the students in Figure 1.1 and represents the probability of retention conditional on failing the test in each subject-grade-year cell — that is, $Prob(Retention|Fail) - Prob(Retention|Pass)$.[10] The is a clear jump for 5th grade but not for other grades when the policy took effect.[11]

The more demanding promotion criteria may have motivated students, especially those at risk of failing the tests, to exert additional effort to avoid attending summer school and being retained, since repeating a grade is associated with stigma and pres-

[10]The restriction deals with the change in the distribution of students who failed the test. It is also possible to estimate the discontinuity at the cutoff in Figure 1.1, but the results are noisier due to changes in the cutoff score in some years.

[11]In contrast, Appendix Figure A1 and A2 show that the policy did not affect the exempt students.

sure from peers and teachers (Andrew, 2014; Byrnes, 1989). The retention policy's incentives for school staff, however, were small for three reasons. First, there were no direct consequences associated with retaining students for teachers or principals. Second, public schools are fully funded by NYC, and retaining students does not impose additional financial burdens on schools. Third, retention rates were not public information, and there were few concerns regarding the impact of retaining students on school reputation.

The analysis of the incentive effect is related to Koppensteiner (2014), which found removing a retention policy in Brazil produced a disincentive effect, and is in contrast to most other studies on grade retention policies, which have evaluated the effects of repeating a grade (Eren, Depew, and Barnes, 2017; Geng and Rockoff, 2016; Jacob and Lefgren, 2004a; Ozek, 2015).

Accountability Scheme

In 2007, NYC implemented an accountability scheme for all public schools except those that only serve special education students. The scheme linked accountability ratings (letter grades ranging from A to F) with rewards and punishments.[12] High-performing (A and B) schools were awarded additional funding, while low-performing (D and F) schools faced substantial consequences, such as potential loss of students through a special transfer program, removal of the principal, and even closure.

The letter grades were based on three components: school environment (15% of

[12]The scheme experienced a major reform in 2010 and was removed in 2013.

the overall score), student performance (25%), and student progress (60%).[13] School environment scores were based on student attendance and survey responses from students, parents, and teachers; student performance scores were based on students' test scores; student progress scores were based on improvements in students' test scores. The calculation of student progress scores requires two years of test score data, the second of which is for a higher grade level. As a result, students in 3rd grade or repeating a grade are not counted in the student progress component.[14]

Schools' scores on all three components were first compared with scores of a set of similar schools within each school type ("peer schools") and then with scores of all schools citywide, then converted into an overall score, and finally assigned a letter grade.[15] The use of peer schools was intended to incentivize schools of all achievement levels.[16]

To examine the allocation of school effort within each school, I take advantage of one specific element in the student progress score: improvements in the school-wide lowest-third students' test scores, which counted for 15% more points than improvements in other students' test scores in the overall score. School-wide lowest-

[13]Appendix Figure A3 presents each component of the accountability grade rubric and its weight in calculating the overall score. Full documentation can be found at http://schools.nyc.gov/Accountability/tools/report/ProgressReport_2007-2013.htm.

[14]Since retained students do not count toward this component and schools had some discretion on which students to retain, retention patterns may have changed after the accountability scheme was implemented. However, the overall low retention rate (2%) makes this potential change unlikely to be driving the main results. This change may be itself an interesting phenomenon, and there is a separate analysis on this topic in the appendix.

[15]School types include elementary schools, K-8 schools, middle schools, and high schools.

[16]Schools could also earn extra credit for substantially improving test scores among several student subgroups: ELL students, special education students, and students scoring in the city's lowest third the previous year.

third students are defined as those who scored in the lowest third in each subject, grade, and school in the previous year.

This element brings two more advantages to the identification strategy. First, since it varies at the grade level, the analysis may identify the effect of the accountability scheme on lowest-third students separately across grades. Moreover, lowest-third students are defined within each school and cover a wide range of student characteristics and achievement levels. As a result, it is unlikely that other concurrent policies are driving the effects on lowest-third students.[17]

NYC actively encouraged schools to focus on lowest-third students. For example, NYC sent out an annual list of lowest-third students to assist each school in identifying these students and providing additional assistance to them.[18] Other elements in the accountability scheme are symmetric, giving equal weight to all students. Therefore, the lowest-third element may direct additional instructional focus and attention toward lowest-third students in each school.[19]

One limitation is that this element only allows me to identify the relative change between lowest-third and top-two-thirds students. However, one thing to note is that identifying the effect of the whole accountability scheme in NYC is almost impossible, with virtually all schools being held accountable and compared with a set of similar schools. In addition, understanding how schools allocate effort is of great

[17]Other policies may include proficiency counting at the state level as part of No Child Left Behind and student performance scores in this accountability scheme.

[18]The list is not available to the author and is thus manually generated from the data.

[19]One potential concern is that the student performance component may interfere with the additional incentives on lowest third students. I test such possibilities in the empirical analysis and find no evidence.

importance for educational equity and many studies (e.g., Deming et al., 2016; Ladd and Lauen, 2010; Neal and Schanzenbach, 2010) have examined the distributional effects of accountability schemes. Lastly, the scheme in NYC mimics a common situation in education generally: Agents face multiple tasks (Dixit, 2002) and overlapping incentives (Fryer Jr, 2013).

The overall design of the accountability scheme in NYC also differs from several accountability policies in other settings, which provides an opportunity to examine a different incentive system. Accountability systems typically implement two models: a status model emphasizes the number of students attaining a certain proficiency level; a growth model emphasizes improvements in students' test scores.[20] Many studies focus on the distributional effect of a status model (Macartney, McMillan, and Petronijevic, 2015; Neal and Schanzenbach, 2010; Reback, 2008) and varying accountability pressure on students' test scores (Deming et al., 2016; Reback, Rockoff, and Schwartz, 2014), and they find evidence of teachers' targeted effort on "bubble students", who have the greatest potential in contributing to reaching the accountability requirement.[21] In contrast, the NYC system includes both models and provides additional incentives to lowest-third students.

[20]See Figlio and Loeb (2011) for a more thorough discussion of these two models.

[21]A few studies evaluated the effects of receiving different letter grades from the accountability scheme on students' test scores and survey responses (Chiang, 2009; Rockoff and Turner, 2010; Rouse et al., 2013).

1.3 Data

The data include individual-level administrative records of all students with linked teacher characteristics from grade 3 to grade 8 in NYC public schools from 1999 to 2009. These records contain each student's demographic characteristics, school and class identifiers, scale scores in math and English, days absent from school, and suspensions, as well as teachers' demographic characteristics, experience levels, and absence records.

The empirical analysis focuses on 4th, 5th, and 6th grades because other grades either did not count toward the accountability scheme or did not allow me to cleanly identify the retention policy in isolation.[22] In order to analyze the interaction of the two policies, the main analysis focuses on students who are subject to the retention policy and include the exempt students in certain estimations.

Certain observations are dropped from the analysis. Student records with missing current test scores in either math or English (6% of the data) are dropped to minimize the potential issue of selection into testing. Since prior covariates are used throughout the analysis, the first year of data (1999) and student with missing prior records (6% of the data).

Panels A, B, C, and D in Appendix Figure A4 present the percentage of exempt and eligible students who took the tests in each year, separately for math and English. Panels A and B show that the overall test-taking rate for eligible students was high (around 95%) and increased smoothly over the analysis period, with a small jump of

[22] A separate analysis of these grades is available upon request.

2% in 2003, possibly due to the passing of No Child Left Behind. Panels C and D indicate that many more exempt students started taking the math tests (20% more) in 2003 and the English tests (30% more) in 2007. Because of the data restriction, the composition change in the test-taking exempt students is not a concern until 2008 (see Panel E), two years after a subset of exempt students started taking tests in both subjects. Panel E shows the percentage of exempt students in each year after imposing the data restriction. There are two noteworthy patterns. First, some more (1.5% to 2%) students became exempt in 2002 and 2007. Since nonexempt students consist of more than 90% of the sample, this change might mostly complicate the analysis of exempt students. In the later analysis, this change does not seem to be empirically important. Second, many exempt students appeared in the dataset after 2008 because they started taking both tests in 2007, and the data restriction may only exclude them in 2007.

The analysis includes three types of outcomes. The first type directly measures academic achievement and includes math and English test scores. The second type measures students' behaviors, including days of absence and suspensions. Although teachers and principals have some discretion in the notice of suspension, the discipline code in NYC requires documentary evidence and witness testimony for suspension and provides a comprehensive list of relevant infractions, limiting flexibility in suspending students. Therefore, suspensions still partially account for student behaviors. The last type of "outcome" concerns teacher characteristics, including teachers' experience levels and absences.

Test scores across grades and years use different scales and are converted into

proficiency ratings according to the rule set by the accountability scheme. The rule converts each scale score to a measure from 1 to 4.5, with a continuous distribution of scores within each proficiency level. Specifically, the rule is defined as follows:

$$RescaledTS_{ist} = [\frac{RawTS_{ist} - Min(RawTS_{glst})}{Max(RawTS_{glst}) - Min(RawTS_{glst})}] - 0.01 \times \mathbb{1}(l < 4) + Level_{igst}$$

in which $RescaledTS_{igst}$ represents the rescaled test score of student i in subject s and year t, $RawTS_{ist}$ is the raw test score of the student, $Min(RawTS_{glst})$ and $Max(RawTS_{glst})$ are the minimum and maximum scores at student i's proficiency level l, and $Level_{ist}$ is student i's proficiency level. In the case of $Level_{igst} = 4$, the expression in brackets is divided by 2. This conversion rule allows me to preserve the variation in the means and standard deviations of the scale scores across years and grades.

Absence and suspension records are censored to minimize the influence of extreme values. Both absences and suspension records are censored at the 99th percentile to have a maximum of 70 days of absences and an indicator of ever being suspended during each academic year.

Table 1.1 presents summary statistics on the eligible students for the whole sample, school-wide lowest-third students in either subject, and school-wide top-two-thirds students in both subjects. Although lowest-third students are on average lower-achieving in all dimensions, the differences are not huge.[23] Appendix Figure A5

[23]In this table, teacher characteristics are the average of two subjects for simplicity. The difference in teacher experience seems to be driven by tracking within each school. For example, some schools have classes that contained no lowest-third students.

further demonstrates this argument by showing the kernel density of test scores for lowest-third students and top-two-thirds students: There is a large overlapping in the test scores of these two types of students.

1.4 The Effects of the Retention Policy Alone

This section uses a difference-in-difference (DID) strategy with both a simple control group and a synthetic control method to estimate the incentive effects of the retention policy in isolation.

Identification Strategy

To identify the incentive effects of the retention policy, I focus on students who are subject to the policy and at risk of failing the test in the grade subject to the policy (5th grade). According to the definition used by the Department of Education at NYC, students who had a prior test score below 3 are at risk of failing the test. Data validate this argument: Appendix Figure A6 plots the empirical probability of failing the test against prior test scores in 5th grade, and students with prior test scores above 3 have a close-to-zero probability of failing the test. Therefore, the identification strategy follows this definition.

The choice of an appropriate control group is difficult. A reasonable control group should come from the same grade to account for the availability of Saturday schools and different tests across grades — that is, students who are either exempt from the policy or have no risk of failing the test. However, both of these groups have no

overlap with at-risk students and might fail to satisfy the parallel trend assumption.

DID results with a control group containing both types of students show that the pre-treatment trend on test scores is not satisfactory. The empirical specification follows a DID model for 5th-grade students with year and group fixed effects prior to the accountability scheme:

$$A_{ist} = \beta_0 + \gamma' X_{it} + \delta_t + \beta_1 Risk_{ist} + \beta_2 Risk_{ist} * RetPol_{it} + \epsilon_{ist} \tag{1.1}$$

In this equation, A_{ist} is an outcome of interest in year t; X_{it} includes ethnicity, free lunch status, gender, and an indicator of repeating a grade; δ_t are year fixed effects; $Risk_{ist}$ is an indicator of at-risk students in subject s; $Risk_{ist} * RetPol_t$ is an interaction term between $Risk_{ist}$ and a dummy of implementing the policy.[24] Standard errors are clustered at the school-year level to account for idiosyncratic shocks within each school-year cell. β_2 estimates the incentive effect of the retention policy.

To address this challenge, I also adopt a synthetic control method (Abadie, Diamond, and Hainmueller, 2010) to select a subset of students from the control group in the DID specification. The "donor pool" is formed by splitting the control group into bins of prior outcomes. Prior math and English test scores are each divided into 35 groups with 0.1 points per group to estimate the effect on scores; prior absences are divided into 35 groups with 2 days per group to estimate the effect on absences and suspensions.

[24]When I estimate the effect on absences and suspension rates, a student at risk in either math or English is considered at risk.

The matching covariates include pre-treatment average of current and prior outcomes, along with percentage of students who are white, black, Hispanic, Asian, female, receiving free lunch, and repeating a grade. The matching algorithm uses a Stata package developed by Abadie, Diamond, and Hainmueller (2014), which minimizes the pre-treatment mean square prediction error (MSPE). However, the matching for teacher characteristics is unsatisfactory, and the graphical evidence looks messy. Therefore, estimation for these outcomes also includes the DID strategy with a simple control.

Inference is based on assigning a treatment status to each member of the donor pool and comparing the treatment effects on the actual treated group with the placebo treatment effects on the members of the donor pool. Such information is summarized in a ratio test that follows Abadie, Diamond, and Hainmueller (2015): $P(\frac{\text{Post-RMSPE}}{\text{Pre-RMSPE}}_{treat} < \frac{\text{Post-RMSPE}}{\text{Pre-RMSPE}}_{control})$, where Post- and Pre-RMSPE are post- and pre-treatment root mean square prediction error. Intuitively, a large $\frac{\text{Post-RMSPE}}{\text{Pre-RMSPE}}$ stands for a large treatment effect, which should be larger for the treatment group than for the control group. Therefore, the effect is more unlikely to occur if this probability is lower.[25] Loosely speaking, this ratio resembles the p-value in hypothesis testing.

The analysis focuses on the years between 2002 and 2006 to isolate the effects of the retention policy. Excluding pre-2002 years accounts for the compositional change shown in Appendix Figure A4; excluding post-2006 years avoids the interaction with the accountability scheme. There are potentially two issues associated with the year

[25]Members with lowest/highest prior outcomes are dropped due to inability to match them with a synthetic control group with similar prior outcomes.

2006. First, the policy retained more low-achieving 5th graders in 2005, so 5th graders in 2006 were more negatively selected, and 6th graders in 2006 were more positively selected. Second, the adoption of the state tests may differentially affect the treatment group and the synthetic control group. These two factors may confound the results in 2006.

To corroborate the results, I also show a placebo test that uses the same technique on grades that were not subject to the policy and a distributional effect that compares the eligible students (both at-risk and not-at-risk ones) with the exempt students.

Graphical Evidence and Inference

Appendix Figure A7 plots coefficients with 95% confidence intervals from an event-study version of Equation 1.1, which calculates β_2 for each year. Panels A and B present the results for math and English test scores and show a clear difference in the pre-treatment trends for the treatment and control groups, which prevents conclusions from being drawn the figure. Panels C and D seem to have a satisfactory pre-trend and show no effects on absences and suspension rates.

Figure 1.3 plots the difference between treatment group and the synthetic control (red line) and the difference between each member of the donor pool and its synthetic control as inference (gray lines). Panels A and B present the results for math and English test scores, and the red line shows a fairly flat pre-treatment trend for the treatment group. Post-treatment differences suggest an increase in both math and English test scores for at-risk students. Inference suggests that the improvement in

math is possibly "significant" but the one in English is likely not — several members in the donor pool show larger effects. Consistent with the graphical evidence, the ratio test for math test scores is 0% and that for English test scores is 14%. Panels C and D show no discernible effects on absences and suspension rates; the ratios are 38% and 61%, respectively.

Appendix Figure A8 presents a placebo test focusing on the grades not subject to the policy (4th and 6th grades) and shows no clear change in the year when the policy was implemented. All ratios are above 10%.

One concern is whether the policy only affected at-risk students, since teachers' efforts and Saturday schools may have benefited other students. To explore this possibility, I examine the distributional effects on eligible students, using exempt students as a control group.[26] Since these two groups of students might be incomparable, such evidence is suggestive.[27] Appendix Figure A9 presents the distributional effects in a change-in-change graph during 2003 and 2005. The x-axis represents prior test scores, which are divided into bins of 0.2 points each. Each point represents a difference-in-difference estimate of the retention policy for each bin of students. Above the horizontal line stands for improvements in the outcome. To the right of the black line are students who faced little risk of failure. The pattern that there is little evidence of improvement in the test scores of students who are not at risk of failing the test (those with prior test scores above 3) reassures us that the policy did not seem to

[26]It is also possible to examine the effects on exempt students in 5th grade, using exempt students in 4th and 6th grades as a control. However, because students take different tests in different grades, it is difficult to draw any firm conclusions from this estimation.

[27]A DID strategy would suggest the pre-treatment trends of these two groups are not paralleled.

have an overall improvement in all students' test scores.

Appendix Figure A10 assesses the role of teachers by presenting the evidence on teachers' experience levels and absences. Panels A and B show no discontinuity for teachers' average experience levels in the year when the policy was implemented; Panels C and D suggest a small increase in teachers' absences.[28] The gray dashed lines suggest that the inference test does not support any of these effects, although the gray lines' messiness weakens the test. Appendix Figure A11 uses the DID strategy to complement the analysis on teachers, and it shows no effects either.[29] These results suggest that being assigned to more experienced teachers or having teachers with fewer absences cannot explain the (lack of) effects of the retention policy.

In conclusion, the retention policy alone did not significantly improve students' academic achievement overall, apart from some evidence suggesting a positive effect on math test scores and English test scores (statistically insignificant) concentrated among at-risk students. Examining teachers' characteristics shows no effects. Placebo tests using grades not subject to the policy show no effects either.

1.5 The Accountability Scheme Alone

This section focuses on grades not subject to the retention policy and uses a DID strategy to estimate the effects of the accountability scheme in isolation on lowest-

[28]Schools may have assigned teachers based on a cutoff of three years' experience, since the probationary period for a nontenured teacher in NYC was three years, and Rivkin, Hanushek, and Kain (2005) showed that teacher effectiveness improves the most in the first three years. Using an indicator of three or more years of experience also shows no effects.

[29]Appendix Table A1 shows that all coefficients are small and statistically insignificant, consistent with the graphical evidence.

third students.

Identification Strategy

The identification strategy examines students in grade levels not subject to the retention policy to estimate the effects of the accountability scheme on the school-wide lowest-third students in isolation. The analysis adopts a DID strategy: The treatment group is lowest-third students, and the control group is top-two-thirds students. Because the retention policy only applied to general education students and the policy interaction will focus on these students, the following analysis separately examines general education students and special education/ELL students.[30]

The empirical specification follows a DID model with grade-year fixed effects and a control function:

$$A_{ist} = \beta_0 + \phi F_{gr}(A_{it'}) + \gamma' X_{it} + \theta_{gt} + \beta_1 Low_{ist'} + \beta_2 Low_{ist'} * Act_{it} + \epsilon_{ist} \qquad (1.2)$$

where $F_{gr}(A_{it'})$ includes grade-specific cubic polynomials of prior test scores in math and English, absences, and suspensions, which interact with an indicator of repeating a grade; θ_{gt} represents year-grade fixed effects; X_{it} is a vector of student characteristics; $Low_{ist'}$ indicates the status of being a school-wide lowest-third student in subject s and year t; and $Low_{ist'} * Act_{it}$ is an interaction term between $Low_{ist'}$ and an indicator of the post-accountability years, Act_{it}. β_2 estimates the effect of the accountability

[30]The latter students could potentially earn extra credit for schools in the accountability scheme, and thus might have received additional assistance.

scheme on lowest-third students. Standard errors are clustered at the school-year level.

The control function deals with a concern that arises from the fact that the distribution of test scores changed over time and the change differed across grades. Appendix Figure A12 shows that the average prior test scores for each grade (displayed separately for lowest-third and top-two-thirds students) increased in a non-monotonic manner.[31] This pattern may have induced different mean reversion patterns during the same period, which would confound the estimation of the effect when directly comparing lowest-third students with other students.

Since such trends are not monotonic, including a linear time trend may not address the issue. Moreover, mean reversion depends on not only the average of prior test scores but also the distribution of prior test scores. Appendix Figure A13 presents the relationships between current and prior outcomes for each grade for years prior to the implementation of either policy. Clearly, these relationships are non-linear and vary across grades, especially for math test scores.

Including grade-specific cubic polynomials of lagged outcomes may address this issue by controlling for differences in the distribution of prior test scores across grades and years. Allowing the coefficients to vary by repeating a grade deals with the issue that the percentage of retained students changed during this period. The coefficients might change over years due to other concurrent shocks. Examining the pre-treatment trend may check this issue, and a flat pre-trend alleviates such a concern.

[31]The non-monotonicity is partially due to the policies implemented on certain subgroups of students in different years and grades, such as the retention policy.

The graphical analysis also shows the distributional effects of the policy, plotting the means of residuals against students' prior ranks in each subject. The residuals are obtained by regressing the outcomes according to the following specification:

$$A_{it} = \beta_0 + \phi F_{gr}(A_{it'}) + \gamma' X_{it} + \theta_{gt} + \underbrace{\epsilon_{it}}_{\widetilde{A_{it}}} \tag{1.3}$$

in which $F_{gr}(A_{it'})$ is the control function, X_{it} is a vector of demographic characteristics, and θ_{gt} contains grade-year fixed effects. Residuals $\widetilde{A_{i,t}}$ are obtained for graphical analysis.

Graphical Evidence

Figure 1.4 presents an event-study version of Equation 1.2, which plots the coefficient β_2 and its 95% confidence interval for each year, focusing on general education students. The left panels (A, C, and E) examine math-lowest-third students. Panel A shows a flat pre-treatment trend and a small drop in math test scores in the year when the accountability scheme was implemented. Panel C shows that students' absences are flat prior to the policy except for a small jump right before the policy was enacted; there is another jump in the year when the policy was implemented; Students' suspension rates (Panel E) rise steadily before the policy and seem to increase slightly when the policy was implemented.[32] Panels B, D, and F present the results for English-lowest-third students. There is an upward trend in the pre-treatment

[32]The change in 2005 might reflect the effect of adopting the state tests/accountability. However, the main conclusion seems robust to accounting for this change.

period but no clear jump in the year of the policy.

Appendix Figure A14 presents the distributional effects and supports Figure 1.4. Prior ranks are divided into 33 quantiles at the subject-grade-school level. There are three lines in each panel: The lighter dashed line plots the means of the residuals in the years 2003 and 2004, the darker one plots the years 2005 and 2006, and triangles represent the post-accountability era. The left panels (A, C, and E) use prior math ranks as the x-axis and show that math-lowest-third students are driving the effects. The right panels (B, D, and F) use prior English ranks as the x-axis and show overall negligible effects on English-lowest-third students.

The lowest-third students do not seem to have received different teachers. Appendix Figure A15 uses the same specification and shows no discernible discontinuity for teachers' experience levels and absences in the year of the accountability scheme.

Appendix Figure A16 presents the results for special education/ELL students. Because of the smaller sample size, the overall movement is more jumpy, and the confidence intervals are larger than in Figure 1.4. However, it appears that the accountability scheme induced little improvement in these lowest-third students' test scores, absences, and suspension rates.

Regression Results

Table 1.2 presents the point estimates for general education students. The results are generated by Equation 1.2, with a time trend for lowest-third students to accommodate the pre-treatment trend. Consistent with the graphical evidence, Panel A shows

that math-lowest-third students experienced a decline of 0.075 points in math test scores (10% of a standard deviation); absences increased by 0.23 days (marginally significant), and suspension rates increased by 0.004 percentage points (insignificant). Panel B shows that English-lowest-third students experienced a negligible change in their English test scores (0.031 points, or 4% of a standard deviation), absences (0.048 days), and suspension rates (-0.27 percentage points). The effects on teachers' experience levels and absences (Appendix Table A2) are all small and statistically insignificant.[33]

A potential concern is that adopting the state tests may have changed the distribution of students across achievement levels and confounded the effects. Appendix Figure A17 plots the percentage of free lunch recipients (a proxy for socioeconomic status) across students' ranks in 2005 and 2006 (the year when the state tests were adopted) and shows no evidence of such a change.

In summation, the accountability scheme alone did not substantially improve English-lowest-third students' academic achievements and may have slightly harmed (in a relative sense) math-lowest-third students' academic achievements. Further, there is no evidence that more experienced or less absent teachers were assigned to lowest-third students.

[33]Replacing teachers' experience levels with an indicator of having three or more years of experience also shows no effects.

1.6 Policy Interaction

This section uses a triple-difference model to estimate the interactive effects of the retention policy and the accountability scheme on lowest-third students.

Identification Strategy

The identification strategy focuses on students subject to the retention policy and uses a triple-difference model to estimate the interactive effects of the two policies on school-wide lowest-third students. The model essentially subtracts the sum of the individual effects of the retention policy and the accountability scheme from their combined effects on lowest-third students. The empirical specification is as follows:

$$A_{ist} = \beta_0 + \phi F_{gr}(A_{it'}) + \gamma' X_{it} + \theta_{gt} + \beta_1 Low_{ist'} + \beta_2 Low_{ist'} * G5_{it}$$
$$+ \beta_3 Low_{ist'} * RetPol_{it} + \beta_4 Low_{ist'} * Act_{it} + \beta_5 Low_{ist'} * RetPol_{it} * Act_{it} + \epsilon_{ist} \quad (1.4)$$

in which $Low_{ist'}$ indicates being a school-wide lowest-third student in subject s; $Low_{ist'} * RetPol_{igt}$, $Low_{ist'} * Act_{it}$, and $Low_{ist'} * G5_{it}$ stand for three interactive terms between $Low_{ist'}$ and indicators of the accountability scheme, the retention policy, and being in 5th grade, respectively; $Low_{ist'} * RetPol_{igt} * Act_{it}$ indicates the triple interaction between lowest-third students, the accountability scheme, and the retention policy. β_5 provides the interactive effect between these two policies. Standard errors are clustered at the school-year level.

Since the estimation compares students across grade levels, a potential concern is

that the results might be confounded by the use of different tests. Since the lowest-third element also applies to students who are exempt from the retention policy, the analysis applies the same specification to these students as a placebo test.

Graphical Evidence

Figure 1.5 presents the graphical evidence on the interactive effects by plotting the effect of being a lowest-third student in 5th grade in each year. There are three periods: Years 2003 and 2004 capture the pre-policy differences; years 2005 and 2006 show the effects of the retention policy on lowest-third students; years 2007, 2008, and 2009 reflect the interactive effect of the two policies.

The left panels (A, C, and E) present the evidence on math-lowest-third students.It is evident that the retention policy did not differentially affect math-lowest-third students, possibly because the retention policy concerns absolute test scores while lowest-third students are defined by their relative test scores. When the accountability scheme was implemented two years later, there is a clear and substantial jump in math test scores and a drop in students' absences and suspension rates.[34] Panels B, D, and F present the results for English-lowest-third students and show overall negligible effects, except for a modest increase in English test scores. The small effect in English is consistent with the overall insignificant effects of each policy in isolation on English test scores.

Figure 1.6 presents the placebo test, using students exempt from the retention

[34]The jump in 2006 might reflect the impact of the state test/accountability, but the magnitude looks fairly small.

policy. Because of the smaller sample size, the overall patterns are jumpier and noisier. Panels A, C, and E present the results for math-lowest-third exempt students and show no evidence of any effects in the year when the accountability scheme was implemented. Panels B, D, and F also show little improvement among English-lowest-third exempt students, with some suggestive evidence of increased suspension rates.[35] There seems to be an upward trend after the policy was implemented, which is perhaps driven by the compositional change depicted in Appendix Figure A4 (as discussed in the data section).

Appendix Figure A18 presents the distributional effects and supports the main results. The x-axis is a student's prior rank in each subject, and each point reflects the difference between students with a particular prior rank in 5th grade and those with the same prior rank in the control grades. Years in Figure 1.5 are divided into three periods: years prior to both policies, years with only the retention policy, and years with both policies.

The left panels (A, C, and E) present the effects on math-lowest-third students. Panel A shows little change in math test scores when the retention policy took effect and a substantial improvement in math test scores for all lowest-third students when both policies were in effect.[36] Panels C and E exhibit a relatively uniform decline in math-lowest-third students' absences and suspension rates with both policies in place. The right panels (B, D, and F) display the results for English-lowest-third

[35]The increase in suspension rates is driven by both a sharp increase in 5th grade and a drop in 6th grade, but the exact cause is unclear.

[36]This figure also suggests that the median score component in the accountability scheme does not seem to have affected median students' test scores.

students. There is an improvement in English test scores but negligible changes on absences and suspension rates when both policies were in effect.

The distributional effects suggest a reallocation of school effort from higher-achieving students to lowest-third students in math but not in English. Because students are compared with one another, each outcome is zero-sum in a given year, and additional gains among all students are absent from this figure. If lowest-third students received additional school effort while the others received a similar amount of effort after the accountability scheme, the negative effects on higher-achieving students should be flat as opposed to oblique. There is a clear downward-sloping curve in the figure for higher achievers in math and a uniform change for those in English. This difference might be because the input for learning math is more incompatible across student achievement levels than the input for learning English, and accommodating lower-achieving math students might necessarily harm high achievers in the class.[37]

However, Appendix Figure A19 shows that teacher experience levels and absences do not seem to explain such a reallocation. All panels show little evidence of change when the accountability scheme was implemented.

Regression Results

Table 1.3 presents the point estimates based on Equation 1.4 for students subject to the retention policy. Panel A shows that math-lowest-third students experienced

[37]Some evidence supports such an explanation: Data show that within-class variance in math (0.5) is larger than that in English (0.35).

an improvement of 0.26 points in math test scores (34% of a standard deviation), a reduction of 0.5 days in absences, and a decline of 0.68% in suspension rates, all of which are statistically significant at the .1% level. Panel B presents the results for English-lowest-third students: English test scores increased by 0.053 points (8% of a standard deviation), absences declined by 0.2 days, and suspension rates decreased by 0.45 percentage points. The latter two estimates are marginally significant.[38][39]

Appendix Table A3 presents the results for students exempt from the retention policy and restricts the estimation to the years between 2003 and 2007 to account for the compositional change in 2008. The point estimates show no significant impact of the policy interaction.

Appendix Table A4 shows small and statistically insignificant effects on all outcomes, which are consistent with the graphical evidence.[40] Since teachers are possibly the most important resource that schools may allocate across classes to improve students' test scores, these results suggest that the reallocation of school effort might be within rather than across classes. Data also show little evidence that lowest-third students were assigned to smaller classes or were more likely to be clustered with other lowest-third students. This evidence also supports the argument in favor of within-class reallocation of effort, which is most probably from teachers.

[38]Clustering the errors at the school level has a negligible effect on the standard errors; controlling for prior exempt/nonexempt status has a negligible effect on the point estimates

[39]I also explore the possibility that schools receiving D and F may have exerted more effort and induced a larger complementary effect. The point estimates support this possibility but they are not statistically significant.

[40]Replacing the dependent variable with an indicator of being assigned to teachers with three or more years of experience produces similar results.

Robustness and Placebo Tests

This section presents additional evidence that the interactive effects of the retention policy and the accountability scheme are unlikely to be driven by other confounding factors.

The first exercise performs a robustness check to deal with potential confounding factors due to other elements in the accountability scheme. This concern is likely small, since lowest-third students cover a variety of student characteristics and achievement levels. Such elements include citywide lowest-third students, students in certain ethnic groups, and the percentage of students achieving proficiency levels 3 and 4 on the standardized tests. The check formally tests these elements by including year-specific covariates of being a citywide lowest-third student, categorical dummies of ethnicity groups, and having prior test scores between 2.5 and 3.5.[41] Appendix Table A5 shows the point estimates, which are quite similar to the main results.

The second exercise conducts a placebo test focusing on schools where most students had no risks of failing the test. Since the passing threshold in the retention policy was in absolute terms and the lowest-third element in the accountability scheme concerns low-achieving students in relative terms, a placebo test may examine those who are not at risk of being retained under the retention policy but are defined as lowest-third students in their schools. The estimation follows the same specification as the main regression but focuses on schools with average test scores above the 75th percentile among all schools.[42] Appendix Table A6 presents the point estimates and

[41] These students have a higher marginal probability of reaching proficiency level 3.

[42] Because there are students at-risk of failing the test even in the very high achieving schools,

shows no effects on all outcomes.[43]

A Possible Mechanism

The main results suggest that the complementary effects of the two policies may be due to complementarity of teacher and student effort. Formally connecting the policy interaction and the complementarity in the production function is more challenging. The appendix presents a conceptual framework that illustrates this connection under certain assumptions. The following paragraphs describe a potential mechanism for the results for math-lowest-third students in the empirical analysis.

The accountability scheme in NYC aimed at improving the test scores of students across achievement levels, with an additional emphasis on students scoring in the lowest third. As a result, teachers needed to perform multiple tasks, from tailoring the coursework toward skills covered in the standardized tests to identifying and working on "bubble students" whose test scores were most likely to be improved by teachers' efforts. The question is then which students were seen as "bubble students" when the accountability scheme took effect.

When the retention policy was in effect, students' incentives to improve test scores were low, especially among low-achieving students, and these barely motivated students might have disliked and resisted the test-preparation atmosphere at the school. Although the accountability scheme assigned greater weight to lowest-third students,

this placebo test is somewhat impure but still can provide important evidence of the interactive effects in schools which had much fewer at-risk students.

[43]Separately estimating the effect on general education and exempt students generated positive effects of similar magnitudes.

these students might not have been seen as "bubble students" if teachers found it difficult to teach them. As a result, teachers may have shifted their focus to other students.

However, the presence of the retention policy increased lowest-third students' incentives to improve their test scores, and they may have paid more attention in class. As a result, these students became "bubble students," and therefore the lowest-third element in the accountability scheme incentivized teachers to shift effort toward them, which complemented student effort.

1.7 Conclusion

The collaborative nature of school instruction gives rise to the possibility of using incentive alignment to realize organizational complementarities in human capital production. This paper investigates this possibility by examining the interaction between a grade retention policy (a student-side incentive) and an accountability scheme (a school-side incentive) in NYC. Although grade retention and accountability policies have each been implemented in many settings and evaluated in many studies, the current study is the first to evaluate their interactive effects.

The empirical analysis shows that the retention policy alone improved at-risk students' math scores modestly (by 10% of a standard deviation) but not their English scores, absences, or suspensions. The accountability scheme aimed at increasing lowest-third students' test scores, but it alone did not greatly improve these students' test scores relative to top-two-thirds students: English-lowest-third students

comparatively experienced 4% of a standard deviation increase in English test scores, math-lowest-third experienced 10% of a standard deviation decrease in math test scores, and both experienced little effect on absences or suspensions.

Combining the retention policy and the accountability scheme showed substantial complementarity among lowest-third students in the grade subject to the retention policy, improving math-lowest-third students' math scores by 33% of a standard deviation and English-lowest-third students' English scores by 10% of a standard deviation. Math-lowest-third students also experienced a decline in absences and suspension rates. Robustness checks support the results for math but not for English.

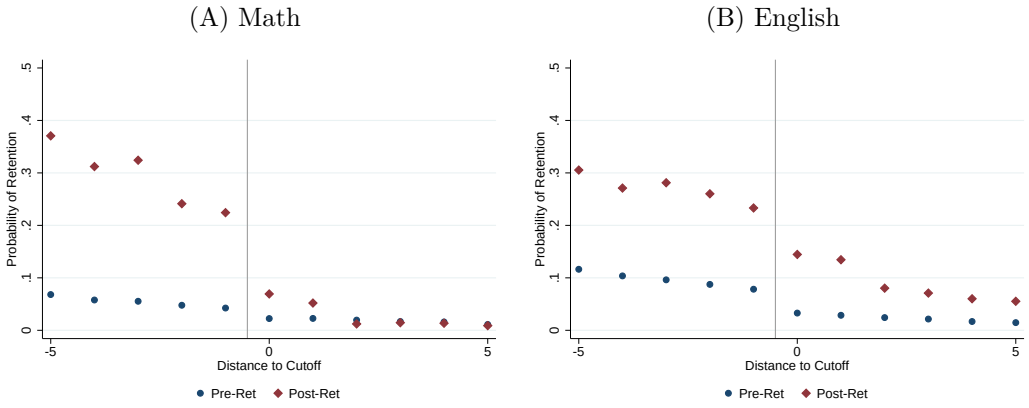
Evidence suggests that the complementary effects are likely driven by complementarity of student and teacher effort rather than by more experienced teachers, smaller class sizes, or assignment of lowest-third students to the same class. These results suggest that there are additional benefits obtained by aligning teacher and student incentives, and that cooperation between teachers and students is essential in education production.

The complementarity of these two incentive-based policies in education provides further evidence that incentive alignment is an important source of organizational complementarities and suggests that school/teacher effort and student effort may be complements in human capital production. Such complementarities provide an explanation of why improving performance at low-achieving schools is very difficult: The marginal benefit of one specific practice is small without other complementary organizational practices.

The substantial interaction between the two policies in the current study underscores the importance of considering incentive policies in combination with each other. Policy design is more efficient when it involves a joint consideration of all possible interventions and their combined impact — and particularly when it takes into account agents' potential behavioral responses (Malamud, Pop-Eleches, and Urquiola, 2016; Todd and Wolpin, 2003). The prevalence of various incentive programs and the interactive nature of production in education and other areas makes this consideration highly relevant and important. The prevalence of various incentive programs and the interactive nature of production in education and other areas makes policies' interactive effects an important concern for policy-makers.

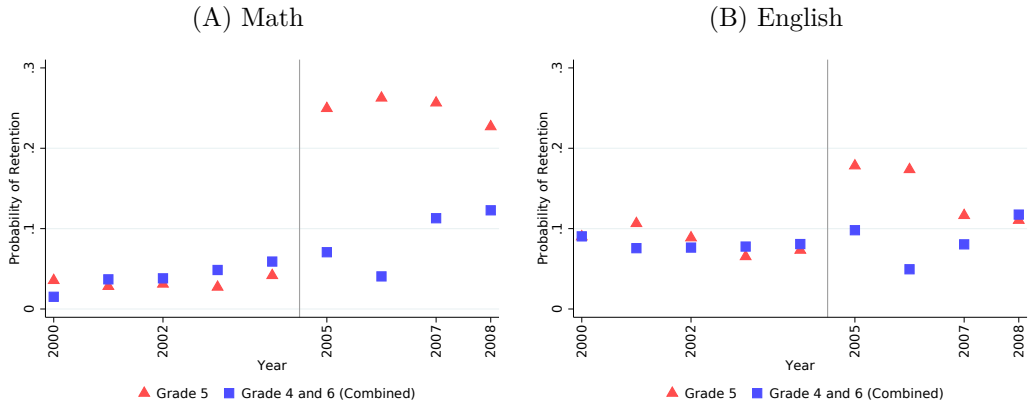
1.8 Figures

Figure 1.1: The Probability of Retention for Eligible Students



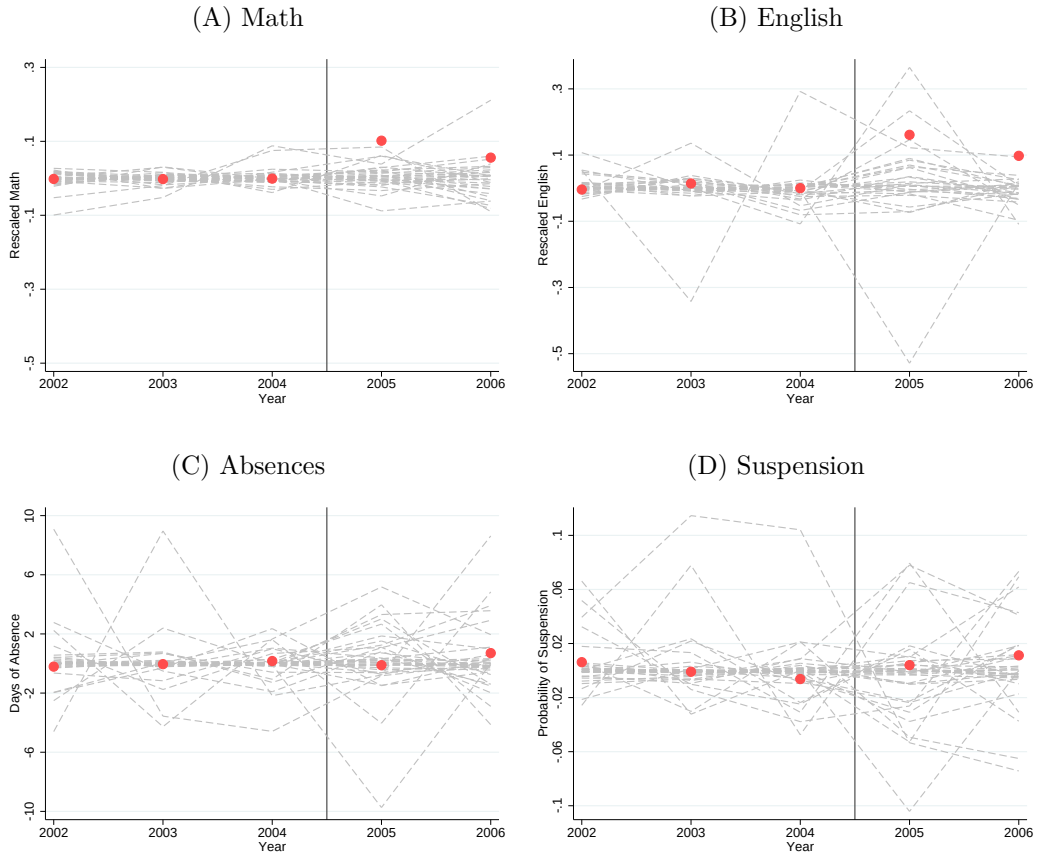
Notes: Both panels are restricted to the years prior to the accountability scheme (prior to 2007). Each point represents the probability of being retained at each value of the index. The index is defined as the difference between a student's spring test score and the cutoff in each subject. Students on the left of the gray vertical line failed the test. Pre-Ret combines the grades/years not subject to the retention policy, and Post-Ret combines the grades/years subject to the retention policy.

Figure 1.2: The Probability of Retention for Eligible Students: Time Series



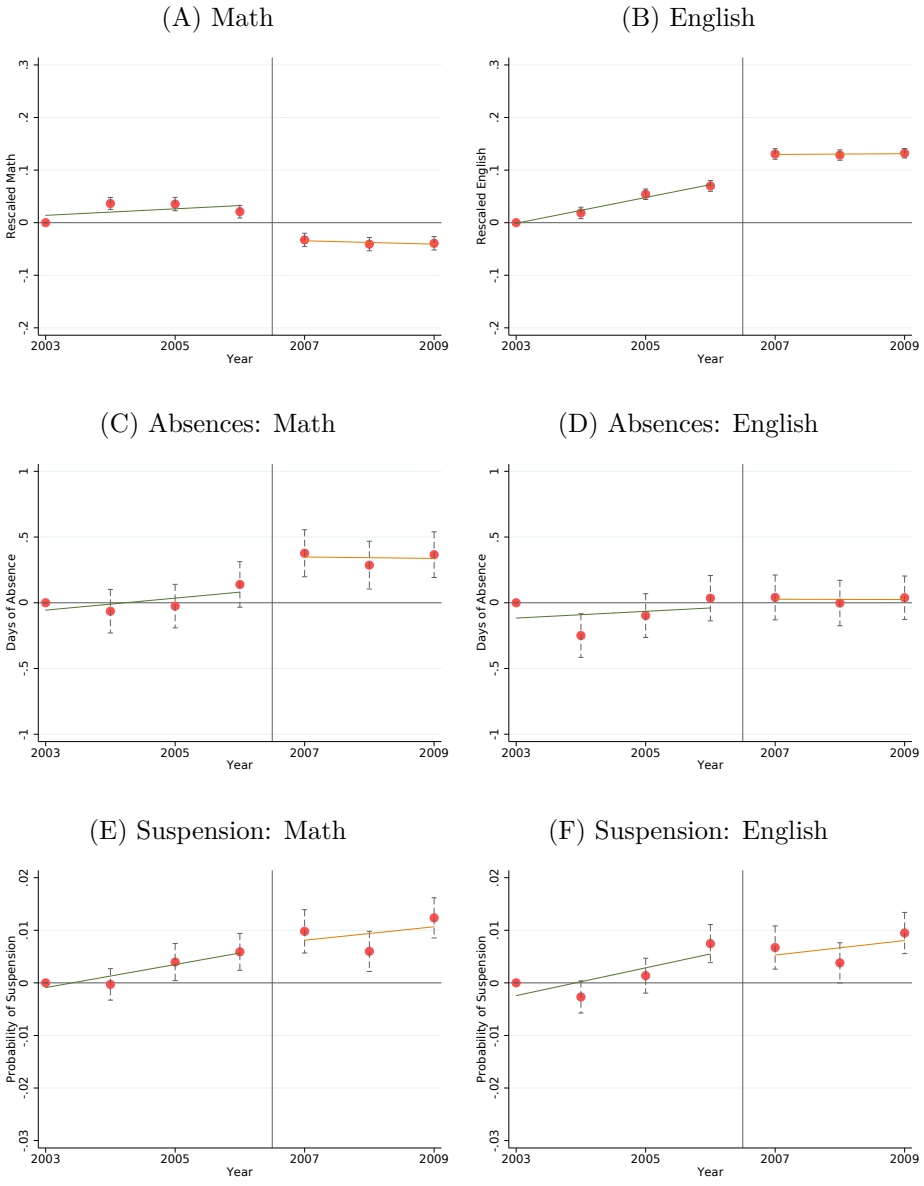
Notes: Both panels focus on students subject to the retention policy. Each point restricts the observations to the students in Figure 1.1 and represents the probability of retention conditional on failing the test in each subject-grade-year cell — that is, $Prob(Retention|Fail) - Prob(Retention|Pass)$. Blue triangles present the probability of retention for 5th grade; Gray squares present the probability of retention for 4th and 6th grades. To the right of the black line are years after the retention policy was implemented.

Figure 1.3: Effects of the Retention Policy (Synthetic Control)



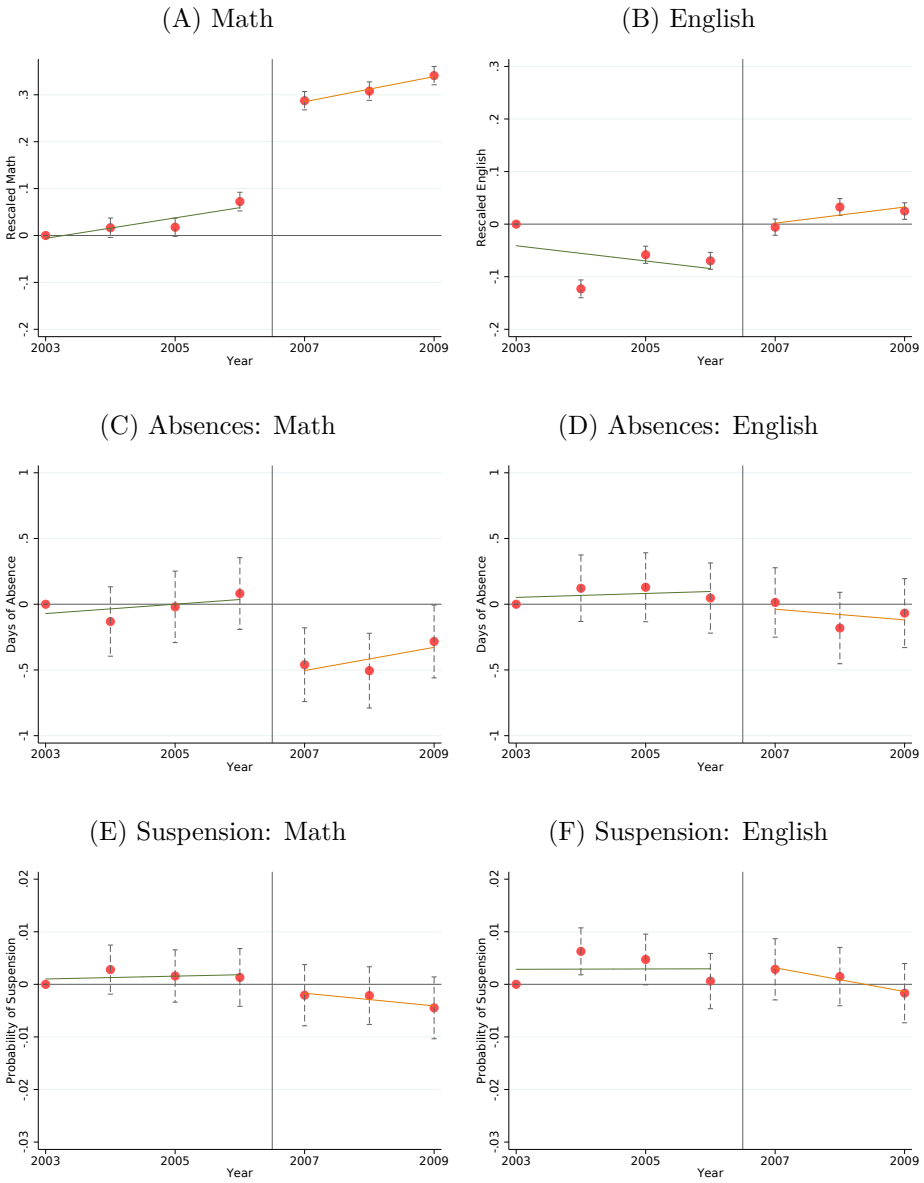
Notes: All panels are based on data from 2002 to 2006 and use the grade subject to the retention policy. The red line plots the difference between the treatment group and the synthetic control group in each year; the gray lines plot the difference between each member in the donor pool and its synthetic control group in each year. The dependent variables in Panels A and B are test scores in math and English; the dependent variables in Panels C and D are the number of days absent from school and an indicator of ever being suspended from school. To the right of the black line are years after the retention policy was implemented.

Figure 1.4: Effects of the Accountability Scheme: General Education



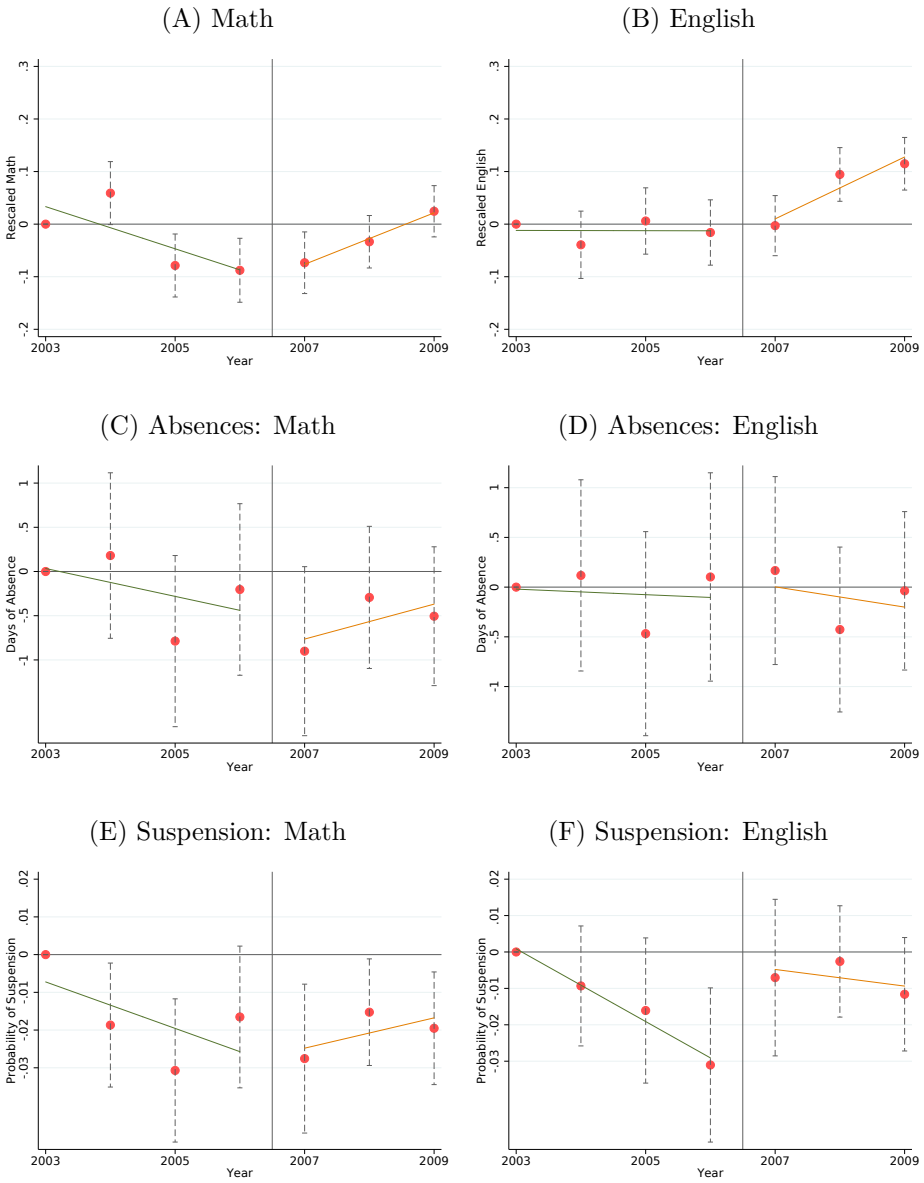
Notes: All panels are based on data from 2003 to 2009 and focus on general education students in 4th and 6th grades. This figure plots coefficients β_2 for each year from an event-study version of Equation 1.2. The dependent variables in Panels A and B are test scores in each subject; the dependent variables in Panels C and D are days absent from school; the dependent variables in Panels E and F are probability of suspension. To the right of the black line are years after the accountability scheme was implemented.

Figure 1.5: Effects of the Policy Interaction on Lowest-Third Students



Notes: All panels use data from 2003 to 2009, focus on students subject to the retention policy, and plot a time series of DID estimates that use the residuals generated from regression 1.3 to measure the effects of being a lowest-third student in the high-stakes grade in terms of the retention policy. The left panels focus on lowest-third students in math, and the right panels examine lowest-third students in English. The dependent variables in Panels A and B are test scores in math and English; the dependent variable in Panels C and D is the number of days absent from school; the dependent variable in Panels E and F is an indicator of ever being suspended from school. To the right of the black line are years after the accountability scheme was implemented.

Figure 1.6: Effects of the Policy Interaction Among Exempt Students (Placebo)



Notes: All panels use data from 2003 to 2009, focus on students exempt from the retention policy, and plot a time series of DID estimates that use the residuals generated from regression 1.3 to measure the effect of being a lowest-third student in the high-stakes grade in terms of the retention policy. The left panels focus on lowest-third students in math, and the right panels examine lowest-third students in English. The dependent variables in Panels A and B are test scores in math and English; the dependent variable in Panels C and D is the number of days absent from school; the dependent variable in Panels E and F is an indicator of ever being suspended from school. To the right of the black line are years after the accountability scheme was implemented.

1.9 Tables

Table 1.1: Summary Statistics

	Full Sample	Lowest-Third	Top-Two-Thirds
Retention	0.02	0.04	0
Free Lunch	0.83	0.85	0.81
Rescaled Math	3.20 (0.76)	2.66	3.52
Rescaled English	3.09 (0.64)	2.61	3.37
Absences	11.44 (10.73)	13.37	10.32
Suspension	0.02	0.03	0.02
Teacher Experience	6.65	6.22	6.90
Teacher Absences	8.15	8.21	8.11
Observations	1,703,423	629,611	1,073,812

Notes: Table shows summary statistics on the eligible students for the whole sample, school-wide lowest-third students in either subject, and school-wide top-two-thirds students in both subjects. Standard deviations are in parentheses.

Table 1.2: Effects of the Accountability Scheme

	Test Scores	Absences	Suspension
Panel A: Math-lowest-third			
Low*Act	-0.075***	0.23*	0.00041
	(0.0068)	(0.098)	(0.0022)
Panel B: English-lowest-third			
Low*Act	0.031***	0.048	-0.0027
	(0.0056)	(0.094)	(0.0022)
Observations	764,941	764,941	764,941

Notes: All regressions restrict observations to grades not subject to the retention policy, implement specification 1.2, and display the coefficient of $Low_{ist'} * Act_{it}$, the interaction term. In Panels A and B, the interaction term is a dummy for the interaction of being in the post-accountability era and being a lowest-third student in math and English, respectively. Standard errors are clustered at school-year level in parentheses. * $p < .05$, ** $p < .01$, *** $p < .001$.

Table 1.3: Interactive Effects on Students

	Test Scores	Absences	Suspension
Panel A: Math-lowest-third			
Low*Ret*Act	0.26***	-0.48***	-0.0068***
	(0.0060)	(0.086)	(0.0018)
Panel B: English-lowest-third			
Low*Ret*Act	0.053***	-0.20*	-0.0045*
	(0.0050)	(0.082)	(0.0018)
Observations	1,155,107	1,155,107	1,155,107

Notes: All regressions implement specification 1.4. The coefficient of the triple-interaction term $Low_{ist'} * Act_{it} * RetPol_{igt}$ is displayed. The triple-interaction term is a dummy for the triple interaction of being a lowest-third student in math or English, being in the post-accountability era, and being subject to the retention policy. Standard errors are clustered at school-year level in parentheses.
* $p < .05$, ** $p < .01$, *** $p < .001$.

Chapter 2

Does Repeating a Grade Make Students (and Parents) Happier? Regression Discontinuity Evidence from New York City

(with Jonah Rockoff)

2.1 Introduction

Schools across the globe routinely organize students by grade levels, where individuals of a similar age are taught together. Children typically enter school with members of their cohort, as determined by a date-of-birth cutoff, and advance one grade level per year. Undoubtedly, this practice arises from the notion that some form of tracking, i.e. grouping together students with relatively similar levels of knowledge and maturity, is the most efficient way to provide instruction. However, the primary use of age to determine grade levels inevitably leads to the following problem: what should public school systems do when a student's level of knowledge or preparation is well below that of his/her age group?[1]

[1]Of course, public school systems may also have to deal with students whose knowledge or preparation far exceeds that of their age cohort. To the best of our knowledge, there is little, if any, research in economics on promoting students ahead of their cohort. Some research on "Gifted and Talented" programs finds that the marginal students admitted to the program did not see improvements in achievement (Bui, Craig, and Imberman (2014)). The practice of "redshirting," i.e., holding children out of school for an additional year before they start kindergarten (see Deming and Dynarski (2008)), is also similar in many ways to retention, but is beyond the scope of this paper. Similarly, we do not address the larger literature on the effects of tracking students by age

One policy used to address this problem is retention, whereby a student repeats the same grade level with the following (younger) cohort of students, and is expected to remain with this younger cohort for the remaining years of public instruction. The use of grade retention is common in the U.S., where Eide and Showalter (2001) estimate that 2 percent of all students in public schools are retained every year. Retention is typically part of a broader set of interventions, such as summer school or course remediation, which are designed to help students improve when they lag behind their grade level. Retention decisions can be based on various measures of academic performance, and the use of high-stakes tests to determine grade retention has grown in the U.S. since the adoption of test-based accountability programs in the last two decades.

Grade retention is highly controversial, with critics arguing that it imposes negative academic and psychological effects on low performing students (Anderson, Whipple, and Jimerson (2005)) and advocates contending that the policy can be academically helpful and psychologically encouraging (Wu, West, and Hughes (2010)).[2] Addressing this controversy with empirical research is also difficult, as it necessitates understanding the likely counterfactual experience of retained students who almost certainly are experiencing severe difficulties in school. For this reason, a number of researchers have turned to the use of regression discontinuity, comparing the outcomes of students who just fail or just pass high-stakes academic assessments that determine

or ability, e.g., Duflo, Dupas, and Kremer (2011).

[2]Note that if retention is undesirable for students (or parents), such policies may also have positive incentive effects on students who are in danger of failing, and thus exert greater effort to pass. See Koppensteiner (2014) for evidence of incentive effects from a change in retention policy in Brazil. Our approach does not capture these broader effects of retention.

grade retention (e.g., Jacob and Lefgren (2004b, 2009), Manacorda (2012), Mariano and Martorell (2013), Ozek (2015), and Schwerdt, West, and Winters (2015)). These studies conclude that retention leads to increased academic achievement in the short-run, particularly for students held back in elementary school, but also find evidence of short-run increases in disciplinary incidents and long-run decreases in educational attainment, particularly for students held back at later grade levels.[3]

An additional hurdle in evaluating the effects of grade retention is that many outcomes are not easily comparable between retained and promoted students. For instance, students who are retained typically take different exams than those who are promoted, making it difficult to compare their relative academic performance.[4] Examinations of longer-run outcomes (e.g., school completion or wages) avoid this type of measurement problem, but face other issues related to interpretation.[5]

[3]Other studies of retention use different empirical approaches for identification and also paint a mixed picture. For example, Eide and Showalter (2001) use age as an instrument for retention and find positive effects on wages, and Wu, West, and Hughes (2010) find retention to be associated with lower teacher-rated hyperactivity, peer-rated sadness, and higher academic competence based on propensity score matching. In contrast, Garcia-Pereza, Hidalgo-Hidalgoa, and Robles-Zurita (2014), in a study based on quarter of birth as an instrument, present evidence that retained students in Spain score lower on PISA examinations.

[4]Researchers have addressed this measurement problem with methods based on somewhat strong econometric or psychometric assumptions on the vertical scaling of scores for tests developed for different grade levels which cover different material (Mariano and Martorell (2013) and Schwerdt, West, and Winters (2015)). Nevertheless, in a deeper sense, measuring the effect of retention on short-run academic performance is always a complicated question when achievement measures are not curriculum-free. To illustrate, suppose some fraction of a school's 7th graders were randomly assigned to repeat 7th grade math, with the remaining promoted to 8th grade math, and all of these students take the exact same math test the following year. If that math test is based purely on the 8th grade curricula, it would not be surprising if the retained students did worse (since they have never seen this material), just as it would not be surprising if the retained students did better if the test was based purely on the material taught in 7th grade (which they have seen twice). More generally, the conclusions from any test will depend on the relationship of the tested material to the material that students have been taught. Without a curriculum-free manner of assessing academic knowledge, the exercise is somewhat meaningless.

[5]For example, suppose that students (randomly) retained had higher years of schooling but lower grades completed than those who were (randomly) promoted. Whether this represents a positive or negative net effect on human capital is unclear. Similarly, if one found that students (randomly)

In this paper, we examine the impact of retention in New York City public schools. Our contribution to the retention literature stems from our use of an unusual source of data: annual surveys of students (in grades 6 through 12) and parents (of students in all grades), which are administered late in the school year (but prior to retention decisions) and include many questions about the quality of students' educational experience. Since the survey questions are the same regardless of a student's grade level, we avoid the measurement problems associated with short-term outcomes like test scores. Because our data contain four waves of surveys, we can also address a number of issues related to interpretation, such as the separation of grade effects from retention effects or whether short run effects of retention fade out quickly over time.[6] As in previous studies, we rely on a discontinuity in the relationship between the probability of retention and scores on mathematics and English Language Arts (ELA) exams taken by students in grades three through eight. Failing these exams was always a factor in the determination of retention in New York City, but it became a much stronger determinant of retention after reforms which were phased in between 2004 and 2008.

We find robust evidence that overall parental satisfaction with school quality rises

retained had slightly lower earnings as those that were (randomly) promoted early in their work careers, does this mean retention (which entails greater educational costs) is not a cost-effective policy? If returns to experience and job-tenure are concave (e.g., Topel 1991), then the early-career and lifetime earnings gaps may have opposite signs, given that promoted students, who finished school a year earlier, are likely to have one more year of early career labor market experience.

[6]Because we have many cohorts and multiple waves of surveys, we can separate grade and year effects from effects of retention. One issue of interpretation we cannot resolve is the fact that retention policies can be bundled with other services, such as attending summer school or receiving more attention from teachers if they are retained. Although we discuss the effects we document as stemming from retention, it is possible that repeating grades without offering additional services would lead to different outcomes.

significantly in the three years after students are retained. We do not see similar impacts for students' overall satisfaction with school quality, but we do find significant positive effects on students' feelings of personal safety in school in the years following retention. We also examine more conventional outcomes such as test scores, student absences, student suspension, and receiving special education. As in prior research, we find retention has large positive effects on students' test scores relative to their (younger) same-grade peers. We see little impacts on student absences and student suspension but we discern a positive effect on the likelihood of receiving special education. Overall, our results indicate that grade retention has positive impacts on the educational experience of students who comply with the test-based policies in place in New York City, as indicated by their parents' opinions. Whether parental opinion is a good barometer of educational quality is beyond the scope of this paper. Yet we would note that society relies on parents to make myriad decisions related to their children's education, and these views are therefore important to examine.

The remainder of the paper is organized as follows. Section 2 describes our data and retention policies in New York City. Section 3 provides a brief overview of our identification strategy. Section 4 presents our main findings, and Section 5 presents an extension of our analysis aimed at identifying effects of retention away from the cutoff. Section 6 concludes.

2.2 Data Description and Policy Background

We link two databases in order to conduct our analyses. The first is administrative records from the New York City Department of Education (NYCDOE) with basic information of all third to eighth graders who were enrolled in NYC public schools. The NYCDOE is the largest district in the nation, with roughly 80,000 students per grade. These data include each student's enrollment in a school and grade level, mathematics and ELA test scores, gender, ethnicity, English language learner status, special education status, free lunch status, total absences, and total suspensions. We use students' grade information between adjacent years to determine retention. We drop a small number of observations with vary rare test scores (i.e., 25 students or less), as these likely come from make-up tests that use a different scale than the normally scheduled exam. We normalize test scores by grade and year to have mean zero and standard deviation one. The standard deviation of scores in New York City on the National Assessment of Educational Progress (NAEP) is comparable with the standard deviation nationwide.

The second database includes responses of parents and students to survey questions collected by the NYCDOE between 2007 and 2010.[7] Starting in the spring of 2007, the NYCDOE has distributed annual surveys to all students from grade 6 to 12 and all parents in public schools. Survey results count for 10-15 percent of a school's score in its annual Progress Report, the main school accountability tool used by NYCDOE (see Rockoff and Turner 2010 for details). The surveys have roughly 20

[7]Copies of these surveys, as well as more recent versions, can be found at http://schools.nyc.gov/Accountability/tools/survey/default.htm

questions, are translated into nine languages and, of great importance for our study, ask the same questions of all parents and students regardless of grade level.[8] In our sample around 80 percent of students and 50 percent of parents responded. Survey questions differed slightly between years but the vast majority of questions remained the same throughout and response rates were relatively high compared to other school surveys (Nathanson, McCormick, and Kemple, 2013).

In 2003, the new NYCDOE administration under Mayor Michael Bloomberg began work to end the practice of "social promotion," where promotion was by default and retention was a rare occurrence, and replace it with a stricter test-based retention policy. This major policy shift, which received a lot of media attention (e.g., Campanile, 2004; Dobbs, 2004; Gootman, 2004; Herszenhorn, 2004) and was fairly controversial, meant that students in grades 3 through 8, with the important exceptions of English language learners and special education students, could be prevented from moving to the next grade if they failed to meet a cutoff score on either the mathematics or ELA tests. Importantly, this more intensive retention policy regime was phased in across grade levels: third grade starting in 2004, fifth grade starting in 2005, and seventh grade starting in 2006.[9]

Figure 2.1 describes the chronological order of testing, survey administration, and

[8]Two exceptions are that high school students are asked about college/career counseling and parents of high school students are asked about the presence of security staff at the school.

[9]Eighth grade was also subject to the new policy starting in 2009, but we do not examine these tests due to a data limitation. Our administrative records on student enrollment end in grade eight, we cannot observe promoted 8th graders who do not respond to surveys in 9th grade, whereas we observe all retained 8th graders regardless of survey response. We therefore do not examine retention in 8th grade in our main tables, although our results are robust to their inclusion (see Appendix Tables A7, A8, and A9).

the steps in the process leading to retention decisions. Tests and survey administration are completed by April, but test results are not typically reported until close to the end of the school year. If a student fails either test, the school principal and teachers review the student's academic portfolio and decide whether to promote the student or require the student to attend summer school. At the end of summer school, students are given another opportunity to pass the tests and, after another review, final retention/promotion decisions are made. We do not have any information on portfolio review, summer school, or make-up testing.

We focus on students tested in school years 2004-2005 through 2007-2008, as we can measure grade retention and link subsequent survey responses for these students. Descriptive statistics are shown in Table 2.1 for our sample, which excludes observations when students would not have been subject to the test-based retention policy due to receipt of special education services or classification as an English Language Learner (ELL). Before focusing on the set of students scoring close to the cutoffs, we present some descriptive statistics on the broader sample (Table 2.1 Column 1). This sample's average test scores are around 0.2 standard deviations higher than the district mean due to our dropping ELL and special education students exempt from test-based retention policies. Six percent of the students in the sample failed their English exam, eight percent failed their math exam, two percent were retained the year after taking the test, and 5.7% of the student sample was retained at least once.[10]

[10]This is similar to the retention rate of 5.1% we calculate for public school students (who were not receiving special education or ELL services) in the Early Childhood Longitudinal Study (ECLS-K). The ECLS-K follows a nationally representative cohort of students that (absent retention) would reach third grade in the fall of 2001, just a few years prior to the cohorts examined in our NYC sample.

Among our sample of students subject to the retention policy in a particular year, 5.5% are later exempt due to a future change in their ELL or special education status; in order to prevent potential bias due to endogenous selection into these categories, we examine outcomes in future years regardless of this future classification. Last, but not least, Table 2.1 shows that the vast majority of students in our NYC sample are poor, as indicated by receipt of free or reduced price lunch (86 percent), that these students are absent an average of 12 days during the year, and that three percent were suspended from school for misbehavior at least once during the year.

Our regression discontinuity design is based on student-year observations with English and/or math test scores close to the cutoff. We show summary statistics separately for students whose (lowest) score is just below the cutoff (Column 2) from those whose scores are exactly at or just above the cutoff (Column 3).[11] On average, 13 percent of students scoring just below a cutoff are retained, compared with around 0.01% for students scoring just above the cutoff. Thus, failing an exam seems to be a necessary, but clearly not sufficient, condition for retention. Compared to the entire sample, it is not surprising that students scoring near the cutoff are far more likely to be from poor households, from disadvantaged minority groups, have higher absences from school, and are more likely to have been suspended. There are also much smaller but still statistically significant differences in observables between those who score just above versus those who score just below the cutoff. Students just above the cutoff are less likely to receive free/reduced lunch (93% vs. 95%), had fewer absences (14.7 vs.

[11]This sample includes students with test scores within five scores of the cutoff, for a total of eleven possible scores on each test administration, which is the window we use for our main results. We examine robustness of our results to inclusion/exclusion of more scores in Section 4.4.

17) and were less likely to be suspended from school (4.0% vs. 5.4%). In Columns 4 and 5 of Table 2.1, we further split the students scoring just below a cutoff by whether or not they were actually retained. Again, there are small but significant differences, showing that retained students are clearly not a random subset among the students who barely fail these exams. Those who are eventually promoted have roughly 0.40 standard deviations higher average test scores, fewer absences (16 vs. 20), lower suspension rates (5.0% vs 7.0%), and are slightly less likely to receive free/reduced lunch (94% vs. 96%).

We performed factor analyses of responses on the parent and student surveys to generate a small number of outcome variables. The results of this analysis (available upon request) showed three underlying factors for students: overall satisfaction, sense of personal safety, and perception of the school environment. For parents, there were just two factors: overall satisfaction and perception of school safety.[12] Appendix A lists the question numbers (taken from the 2008 survey) for the items used to construct each of these variables. We code survey variable values to range from 0 to 100 for easier interpretation, where 0 means that the least favorable answers were always selected and 100 percent means that the most favorable answers were always selected.[13]

[12]Survey questions were originally designed to measure four dimensions of school quality for both parents and students: Academic Expectation, Communication, Engagement, and Safety & Respect. However, Rockoff and Speroni (2008) analyzes the reliability, consistency, and validity of the surveys and finds, as we do here, that responses do not line up along these four dimensions.

[13]This rescaling also deal with questions that do not have the same number of choices. For example, if a question had five possible answers, we gave 0 points for the least favorable, followed by 25, 50, 75, and 100 points, respectively, for answers leading up to the most positive. Likewise, a question with only four possible answers would be scaled using points of 0, 33.3, 66.6, and 100. If a subset of the answers are missing, we simply use the the answered questions.

The lower rows of Table 2.1 provide information on the values of our survey measures in the year that students were tested. Compared to the sample as a whole, students with scores near the cutoff have considerably worse survey outcomes and, even within the sample near the cutoff, students who passed both exams have better parental survey outcomes (about 0.1 standard deviations) and better sense of personal safety (0.07 standard deviations). We also see consistent differences if we compare students who were promoted vs. those who were retained among students who failed at least one of their exams. Thus, it is again evident that there is positive selection of students for promotion among those who fail the exams, reinforcing the need for a credible identification strategy that addresses potential selection on unobservables.

2.3 Empirical Strategy

Each student has two scores (ELA and mathematics) that affect his/her retention outcome. We define our running variable as an index for each student i at year t who is in grade g: $Index_{i,t} = min(ELA_{i,t} - Cutoff_{t,g,ELA}, Math - Cutoff_{t,g,Math})$. We define failure by $F_{i,t} = 1(Index_{i,t} < 0)$. A student whose index falls below zero must have failed at least one of the two tests and, as we show below, is therefore significantly more likely to be retained.

To illustrate our identification strategy, we plot the percentage of students who repeat a grade the following year against the test score index (Figure 2.2), dividing the sample by whether the test was taken in a grade-year cell before or after the

implementation of the more intensive test-based retention policy.[14] Students who failed at least one test by a wide margin (i.e. an index score at or below -10) had a probability of grade repetition of about 20 percent prior to the new regime and almost 60 percent after the more intensive policy took effect. In both pre- and post-policy testing, the probability of retention decreases steadily as the index improves, and students with an index value of -1 had grade repetition rates of around 5 percent and a little over 20 percent, respectively, pre- and post-policy change. There is a discontinuous drop in retention at an index value of 0, i.e. students who just reached the cutoff. Students with non-negative index scores within two points of the cutoff have rates of retention below 2 percent, and students with index values at 3 or above have practically zero chance of being retained. This discontinuous drop in retention across the zero index threshold is the basis for our identification of the impact of grade repetition. It is clear, however, that our statistical power is greatly amplified in the grades and years when the more stringent retention policy is in effect. We return to this issue below.

In our data we essentially have 25 quasi-experiments — five test years and five tested grades — and we combine them for our analyses. As noted before, some grades/years are affected by a more intensive retention policy regime, and the exact cutoff score for failing the tests varies by grade and year. To accommodate these factors, we allow each test grade in each year to have its own control function but impose the same retention jump at the cutoff within each policy regime and a single

[14] Appendix Figure A21 plots the discontinuity of retention at the cutoff in each grade-year cell. The post-policy retention rates are much larger than the pre-policy retention rates. This pattern supports our empirical strategy.

Local Average Treatment Effect (LATE). Later we will explore allowing the LATE to differ by policy regime.

We use two-stage least squares for estimation:

$$\begin{aligned} r_{i,t} &= \theta_1 * 1(Index_{i,t} < 0) + \theta_2 * policy_{t,g} * 1(Index_{i,t} < 0) \\ &\quad + G_{t,g}(index_{i,t}) + FE + \mu_{i,t} \end{aligned} \tag{2.1}$$

$$Y_{i,t,l} = \sigma * \widehat{r_{i,t}} + G_{t,g}(index_{i,t}) + FE + \eta_{i,t,l} \tag{2.2}$$

Each observation is represented by student i, test year t, and the number of years l ("lag") between when the test was taken and when the outcome $Y_{i,t,l}$ is measured. $r_{i,t}$ is an indicator of retention in year t ($\widehat{r_{i,t}}$ is the predicted value from equation 1) and does not vary between lags; $policy_{t,g}$ indicates whether individual i is enrolled at time t in a grade g affected by the new policy regime; and $G_{t,g}(index_{i,t})$ is a grade-year specific cubic function of index. We include grade $\times$ year fixed effects to account for different cutoffs between years and grades and also outcome grade fixed effects to separate out grade effects on survey responses. Our outcomes $Y_{i,t,l}$ include normalized test scores, absences, suspension from school, students' overall satisfaction, personal sense of safety, and perception of the environment, and parents' overall satisfaction,

and perception of school safety within three years after test year t, i.e., $l \in (0, 1, 2, 3)$.

We stack each observation in the administrative datasets up to four times to match with both current outcomes and future outcomes within three years following the initial retention decision. Appendix Table A10 provides an illustrative example. Given that we observe the same student multiple times in the data, we cluster standard errors at student level. [15] We choose the index range of [-5,5] as our main bandwidth and check other bandwidths for robustness. We also present an analysis that includes 8th grade tests and the spring 2009 tests as a robustness check.

Stacking the datasets allows us to evaluate the effects of retention on current and future outcomes in one regression by interacting Equation 1 with lag l. This pooled set-up provides two advantages. First, we simultaneously run placebo tests (l = 0) and observe how effects change over time (l = 1, 2, 3).[16] Second, we can control for outcome grade fixed effects. Since retained students will mechanically attend lower grade levels than their promoted peers, estimates that do not control for grade level effects might conflate any systematic effects of grade level with the effects of retention. Stacking the datasets and combining quasi-experiments allows us to identify outcome grade fixed effects by looking at multiple cohorts and multiple lags simultaneously.

To support the validity of a regression discontinuity design (RDD), it is important that scores are not manipulated around the cutoff. There is little reason to believe

[15] As a robustness check (see section 4.4), we also implement two-way clustering at both the student and index level (Lee and Card (2008)). Clustering at the index level has become somewhat standard practice in the RD literature, but we see little reason to believe that our survey outcomes are correlated at the index level due to common shocks. In the absence of these shocks, clustering at the index level can do more harm than good (see Koles r and Rothe (2016)).

[16] The outcomes when $l = 0$ were realized before any retention decisions, and we run placebo tests by examining the effects of retention on them.

such manipulation takes places, as the math and English tests are developed and graded externally to the school district, and Figure 2.3 shows that the density of observations at each index runs smoothly across the cutoff.[17] Further evidence that there is no manipulation is provided in Figure 2.4, which shows that the percentage of female students and students who receive free/reduced price lunch are also smooth through the cutoff. Appendix B provides additional continuity graphs and regression analyses of other covariates, including attrition rates and survey response rates.

Before proceeding to our results, it is worth nothing that we cannot pin down the mechanisms underlying any effects of retention on parents' and students' views on the quality of education being provided. There are obvious potential mechanisms such as seeing the same material twice and being moved to a younger peer group. There can also be various other mechanisms driven by the "labelling" of retained students at the start of the next school year, e.g., negative effects associated with stigmatization by classmates or positive effects of increased attention from teachers. Very much in line with previous studies of retention, we do not seek to separate out these potential channels but, rather, to provide greater insight into the (local) effects of a widespread policy.

[17] In contrast, Dee et al. (2016) show that exams taken by high school students and graded locally by teachers within a school show significant manipulation around the failing cutoff.

2.4 Main Results

First Stage Results

Formal estimates of the impact of test failure on retention are presented in Table 2.2. The first stage is strong and the coefficients are consistent regardless of whether we include all students (Column 1) or restrict the sample to students for whom we have any survey data, parental survey data, or student survey data (Columns 2 to 4, respectively). Consistent with Figure 2.2, failing at least one test increases the probability of being retained by around 3 percent under the less intensive policy regime and by around 25 percent under the more intensive policy regime.

The Effects of Retention on Non-Survey Outcomes

The main contribution of our paper is to examine how retention affects subjective measures such as parental satisfaction about educational quality. However, in order to provide comparisons with earlier literature and some context for interpreting the survey evidence, we first present effects of retention on test scores, school absences, suspensions, subsequent grade repetition, and subsequent receipt of special education services. Graphical evidence is shown in Figures 2.5 and 2.6, which plot residuals from a regression of each outcome on grade $\times$ year fixed effects against our index variable, while Table 2.3 presents point estimates from regressions based on Equations 1 and 2.

Figure 2.5 shows outcomes in the year of the test ($l = 0$) and is therefore akin to a placebo, since retention decisions are made after these measures are taken. Con-

sequently, there is no visual evidence of a significant jump at the cutoff, and the estimates in Row 1 of Table 2.3 confirm this conclusion. Figure 2.6 shows future outcomes, combining all data within three years after the test ($l \in (1, 2, 3)$) for simplicity. Figure 2.6a and 2.6b show that test scores relative to same-grade peers are dramatically higher for students who just fall below the cutoff on at least one exam; estimated effects of retention in Table 2.3 are 0.55 and 0.63 standard deviations for English and math, respectively.[18] These results are consistent with Jacob and Lefgren (2004b), Schwerdt, West, and Winters (2015), and Mariano and Martorell (2013). Of course, retained students take different tests and, as discussed in the introduction, we cannot interpret these results as an improvement in academic achievement without further assumptions. Any effect on absences and suspensions over the following three years (Figures 2.6c and 2.6d) is difficult to discern graphically, and regression estimates in Table 2.3 suggest being retained has little impact on aggregate absences and some (marginally significant) effect on suspension over the subsequent three years.[19] Figure 2.6e shows that students just below the cutoff are more likely to receive special education within three years after the test; Table 2.3 indicates an point estimate of

[18]An important issue to consider is that failing an exam can have a discontinuous effect on educational experience outside of retention, such as having to attend summer school. If, for example, summer school leads to improved achievement, regardless of retention outcomes, then our interpretation of the two-stage least squares estimates may be incorrect. To shed some light on this issue, we present some admittedly suggestive evidence in Appendix Figure A22, which takes average future test scores for students with index values below zero and plots them separately for retained and non-retained students. We can see that the future scores of non-retained students are quite continuous through the cutoff, while those of retained students are discontinuously higher. While retention within the set of students below the cutoff is obviously endogenous, we believe this graph is reassuring that our RD estimates are driven through the effects of retention, rather than other experiences related to having failed an exam.

[19]When we expand the bandwidth for analysis, the effect on suspension becomes small (about 2%) and statistically insignificant. We use two-stage least squares to estimate the effect of retention on suspension, a binary variable, but estimates from a probit model, not shown here, lead to the same conclusions.

5.7% (more than twice the sample average after the tests). Although classification of special education follows some absolute standard, parents may see retention as a signal and react by seeking additional assistance through special education. This reaction seems more natural since special education exempts students from the retention policy in NYC. [20]

Students who barely pass exams and avoid retention in a given year may have significantly higher probabilities of failing and/or being retained in the future. The tendency for students who initially act as the "control group" to be given the "treatment" of grade retention at a later date can dampen our estimated effects of retention at time t for lags greater than 1. In Column 1 of Table 2.4, we present estimates of the effect of retention on grade level at one, two, and three years after the exam ($l = 1, 2, 3$). The (mechanical) coefficient of 1.00 at $l = 1$ fades slightly to 0.96 at $l = 2$ and slides further to 0.90 at $l = 3$, suggesting that 10 percent of students who would have been retained had they not barely passed their exams are still retained at some point within three years. This "fade-out" of the first-stage effects of failure on retention in NYCDOE is somewhat smaller than what Schwerdt, West, and Winters (2015) document in the state of Florida, where approximately 17 percent of the "marginally promoted" students are retained within three years and 25 percent retained within five years. Not surprisingly then, the effects of retention on academic performance relative to same grade peers is largest at $l = 1$ (0.66 in English, 0.79 in

[20]We suspect that parents may react to failing exams instead of retention and plot the average probability of receiving special education separately by whether the students below the cutoff are actually retained or not in Appendix Figure A22. Although retention is endegenous, we are assured by this figure that our result is driven by actual retention.

math), and declines through $l = 3$ (.36 in English, 0.39 in math). This pattern is also unsurprising given the wider literature documenting "fade-out" of the impacts of academic interventions on standardized test scores (e.g., Cascio and Staiger (2012) and Chetty et al. (2011)), but it is notable that retention has substantial positive effects on test scores – relative to same-grade peers – several years later. [21]

The Effects of Retention on Survey Outcomes

In this section, we turn to our main outcomes of interest from parent and student surveys. As in the previous section, we provide graphical evidence first by regressing each survey outcome on grade × year fixed effects and outcome grade fixed effects and plot the average residual at each index around the cutoff. Before we present figures, it is worth emphasizing that only students above 6th grade respond to surveys and results on student surveys does not necessarily apply to students in elementary schools. Figure 2.7 shows results pooling surveys taken in the three years after the test.[22] Panel A of Figure 2.7 shows a clear jump in parental satisfaction at the cutoff; parents whose children barely passed the tests are less satisfied than parents whose children barely failed. It is also interesting to note that, while there is a weak positive relationship between satisfaction and index above the cutoff, the relationship below the cutoff is strongly negative, which mirrors the "first-stage" relationship of

[21] The results on absences and suspensions are mostly unaffected and not shown here. The results on special education is not shown here because the standard errors in this specification cannot be computed.

[22] Appendix Figure A23 shows there is no evidence of "placebo" effects for surveys taken in the year of the test; recall that the surveys are administered after the tests but prior to scores being known or retention decisions being made.

index with retention. We also see a smaller and slightly less clear jump at the cutoff in students' sense of safety, while students' overall satisfaction, students' views about the school environment, and parents' beliefs about school safety appear fairly continuous through the cutoff.

Before moving to our regression results, we provide two more pieces of graphical evidence, focusing on parental satisfaction and students' sense of safety. First, we plot results separately for tests in grade-year cells with and without the more stringent retention policy (Figure 2.8). For both outcomes, we see clear discontinuities in survey outcomes in the post-policy grade-year cells, with more positive survey responses among students who just failed one of their exams, but no noticeable change at the cutoff in the pre-policy years.[23] The fact that we see clearer patterns in the policy years may simply be due to the first stage being dramatically stronger when the policy was in place. However, given the large (and not uncontroversial) increase in retention brought about by the policy change, the evidence of positive effects on parental satisfaction is interesting. Second, we plot average outcomes for students below the cutoff separately by whether or not the student was actually retained (Figure 2.9). Retention is clearly endogenous, as we cannot separate students to the right of the cutoff by whether or not they would have been retained if the cutoff were higher. However, this plot is reassuring, albeit only suggestive, as it shows that the differences across the threshold seen in Figure 2.7 are driven by relatively high parental satisfaction and sense of safety among retained students; the outcomes for

[23]In Appendix Figure A24, we replicate these plots using outcomes in the same year as the test and show that, regardless of the policy in place, these outcomes are smooth through the cutoff.

non-retained students below the cutoff appear to match in a very continuous manner with outcomes for students above the cutoff.

Regression results in Table 2.5 are largely consistent with what we observe in the figures described above. Retained students' parents are estimated to be 5.4 points (or 0.3 standard deviations) happier than promoted students' parents in the three years following the retention decision. We also find that retained students feel 5.4 points (0.25 standard deviations) safer than promoted ones in the years following the retention decision, while the effects on parental views on school safety, students' views on school environment, and students' overall satisfaction are statistically insignificant.[24] The difference between students' personal sense of safety and their views on the school environment (which include measures of school safety) is instructive. Retained students feel that personally they are more safe, even though their general views of safety at the school level and other measures of school environmental quality are unchanged.[25]

Rather than pooling up to three years, it is interesting to ask whether the effects of retention grow or decay over time. Table 2.6 presents separate estimates of the effects of retention after one, two, and three years, focusing on parental satisfaction

[24]Tests for "placebo effects" on survey outcomes in the year of the test ($l = 0$) do not reveal any statistically significant coefficients. Though the placebo coefficient for students' sense of safety is somewhat large, it is of the opposite sign as the main effect of interest and suggests that, if anything, students just to the left of the cutoff felt somewhat less safe in the year prior to the retention decision.

[25]It is also interesting that parental satisfaction improves while students' overall satisfaction appears unaffected. The parental surveys include students tested below grade 6, and this difference in sample could conceivably make a difference. However, in results not reported here, we find that the effect on parental satisfaction is significant and quite similar in magnitude if we limit to parents whose students were tested in grades 6 and higher. While we lack data to explore this issue further, it is worth noting that Rockoff and Turner (2010) find that short-run improvements in student achievement caused by the NYCDOE accountability system also led parents, but not students, to be happier with the quality of education they received.

and students' personal safety. We find that retained students' parents are slightly less satisfied (although not statistically significant) with their child's education in the year after retention, but significantly happier (roughly 35 percent of a standard deviation) two and three years after retention. The fact that we see an immediate improvement in academic performance (at least relative to same-grade peers) but a delayed effect on parental satisfaction is interesting. On one hand, it may be that satisfaction from academic improvement is wiped out by negative aspects of retention (e.g., stigma). On the other hand, since surveys are administered before test results are known each year, it may be that parents do not know how much their child has improved (at least relative to his/her new same-grade peers) one year after retention.

One might naturally wonder whether improvements in test scores relative to same-grade peers might possibly explain the positive effects on parental satisfaction that we find over three years. This is an important issue; if parents simply value their child's performance rank relative to same-grade peers, then retention may simply re-order students so that parents of those retained are happier but parents of low-achieving students in the younger cohort are made less happy. [26] A purely ordinal interpretation would essentially mean that retention is a zero-sum game.[27] To investigate this question a bit further, we ran cross sectional regressions of parental satisfaction on on students' test scores and find that a one standard deviation increase in both mathematics and ELA test scores increases parental satisfaction by about 1.3 points.

[26]Lavy, Paserman, and Schlosser (2012) report that low-achieving students defined as those who repeat grades are more satisfied with their teachers at the expense of regular students.

[27]The importance of ordinal rank has been shown in the workplace (Card et al., 2012) as well as in educational contexts (Murphy and Weinhardt, 2014).

This does not represent a causal estimate of the impact of test score performance on satisfaction, and there are several reasons to think such coefficients might be biased upward. Nevertheless, if we apply this coefficient, test scores would explain less than 25 percent of the effect of retention on parental satisfaction. We cannot rule out that improvements in performance relative to a new, younger peer group explains some of our results, nor that parents do not care about ordinal rank, but it is unlikely that these factors are the main drivers of our findings.

The pattern of effects on students' personal sense of safety in Table 2.6 reveal a different pattern, with the largest effect in the first year after retention (about half a standard deviation) and positive but gradually declining effects over the next two years. One interpretation is that students feel much safer when enrolled alongside younger peers, and that this age advantage grows less important over time. We explore the explanatory power of a relative-age effect by running a cross-sectional regression of students' personal sense of safety on students' relative age for students who have never been retained; this yields a coefficient that implies that the oldest child in a class responds only about 0.75 points (0.04 standard deviations) higher on average than the youngest child to the questions about personal safety. Thus, the marginal students who are retained due to test failure would have to be much more sensitive than the typical student to their age position for this to explain the effects we find on personal safety.

As mentioned above, we are also interested in whether the effects of retention differ between the pre- and post-policy retention regimes. Estimated effects of retention on parental satisfaction and students' safety that are allowed to differ between policy

regimes (Table 2.7) show that our main findings are driven by the grade-year cells in the new policy regime. The point estimates of the old policy regime for parental satisfaction and students' safety are both negative but statistically insignificant and very imprecisely estimated. This is not terribly surprising since the first stage power under the old policy regime is rather weak. However, similar estimation with test scores as the outcome (Appendix Table A11) shows that the effects of retention on academic performance relative to same-grade peers is remarkably similar in the pre- and post-policy periods. If the effect of retention on parental satisfaction is simply due to academic improvement, we should see similar effects between the two policy regimes but we do not. [28]

Robustness Checks

In this subsection, we present four robustness checks that further support our main findings. First, we re-analyze the effects of retention on parental satisfaction and students' personal safety while widening our bandwidths in one point increments from [-4,4] through [-10,10] (see Appendix Table A12). Our point estimates for effects on parental satisfaction are quite insensitive to bandwidth. Indeed, the only noticeable change is that the coefficient for a "placebo effect" of retention on students' current (i.e., pre-retention) sense of personal safety goes closer towards zero as the bandwidth widens, while the estimated effect on students' future sense of personal safety remain quite stable. Thus, we do not find any evidence that the choice of bandwidth is

[28] We have also examined heterogeneous effects between male and female students and between younger and older students within a cohort but we fail to find any significant differences.

driving our results.

Second, we note that Lee and Card (2008) suggests RDDs should cluster errors at the running variable level to minimize specification errors and this practice has become somewhat standard. However, we see little reason to perform this clustering practice because our outcomes are not subject to any common shocks at the index level and Koles r and Rothe (2016) suggests that this practice may do more hard than good. Nevertheless, we implement a two-way clustering at both the individual and index level in order to make sure that this does not have a major impact on our statistical inference. Reassuringly, we find that the two-way clustered standard errors are quite similar to clustering at the student level (see Appendix Tables A13, A14, and A15).

Third, recall that we do not examine retention in grade 8 or in the school year 2009-2010 in our main results. We omit these observations because we are only able to observe the retention decision and future outcomes of students in 8th grade (and/or tested in school year 2009-2010) if they stayed in the New York City public school system in the next year and they (or their parents) responded to the surveys. In other words, we cannot distinguish between a student who was promoted to grade 9 and left NYC schools from a student who was promoted but did not respond to the NYC survey, nor can we measure retention for students who were tested in 2011 who did not respond to surveys. The addition of these observations to regressions of parental satisfaction and students' sense of personal safety (see Appendix Table A7, A8, and A9) do not significantly alter our main findings.

Last, but not least, retention may induce students to trasfer to another school

and, as we investigated before, to seek assistance through special education. We include the number of years a student has spent in a school, the type of the school, or an indicator of receiving special education as additional covariates in our estimation. These results are shown in Table A16 and similar to our previous estimates. [29]

2.5 Regression Discontinuity Extrapolation

Our regression discontinuity design identifies a local average treatment effect (LATE) for students near the cutoff, but we are also interested in the effect of retention on inframarginal students. We follow recent research in regression discontinuity techniques (Angrist and Rokkanen (2015)) to identify LATEs on students away from the cutoff.

In addition to standard RDD assumptions, this technique requires a Conditional Independence Assumption (CIA) and Common Support (CS). CIA requires the potential outcomes to be mean-independent of the running variable after conditioning on other pre-determined covariates; CS requires treatment status to vary conditional on these covariates. Following Angrist and Rokkanen (ibid.), we test the CIA assumption by regressing our survey outcomes on predetermined covariates (e.g., two-year prior test scores), and then examining the relationship between residuals of this regression and our running variable on each side of the cutoff.[30] We focus on the grades and years under the new policy regime to maximize the power of first stage and

[29]Note that the sample size is different from previous estimation because of missing data.

[30]Specifically, we use standardized mathematics and ELA test scores from one year before each student's current (i.e. using scores a student obtained in 2006 as conditioning covariates of his/her 2007 running variable) as well as gender, ethnicity, free lunch status, and grade × year fixed effects.

explore LATEs on students' personal sense of safety and parental satisfaction.

Results for tests of the CIA assumption (Appendix Table A17 and Appendix Figure A25) show that, conditional on our pre-determined covariates, the relationship between parental satisfaction and the running variable is no longer significant, but the relationship between students' personal safety and the running variable remains. Thus, we only have support for the CIA assumption with respect to the parental satisfaction outcome.

We indirectly test CS by checking the distribution of pre-determined covariates at each index score. Appendix Figure A26 shows a box plot of two-year prior standardized mathematics scores at each index score. The extensive coverage at each index score supports CS.

We calculate a linear reweighting estimator discussed in Kline (2011) to estimate LATEs of retention on parental satisfaction at each index score over the range -11 to 6, which is the largest range of our running variable in which the test of the CIA assumption holds. The estimator is equal to:

$$\frac{E(Y_{1i} - Y_{0i}|x_i, r_i)}{E(W_{1i} - W_{0i}|x_i, r_i)} = \frac{E(Y_{1i}|x_i, r_i) - E(Y_{0i}|x_i, r_i)}{E(W_{1i}|x_i, r_i) - E(W_{0i}|x_i, r_i)} \tag{2.3}$$

in which Y_{1i} and Y_{0i} denote the potential outcomes when treated and untreated, x_i are the conditioning covariates, r_i is the running variable, and W_{1i} and W_{0i} denote the potential treatment (retention) status. Kline's estimator assumes linear models

for conditional means:

$$E(y_i|x_i, r_i < 0) = x_i'\beta_1$$

$$E(y_i|x_i, r_i \geqq 0) = x_i'\beta_0$$

$$E(w_i|x_i, r_i < 0) = x_i'\delta_1$$

$$E(w_i|x_i, r_i \geqq 0) = x_i'\delta_0 \tag{2.4}$$

in which y_i is the realized outcome and w_i is the realized treatment. These linear models reduce the estimator to:

$$\frac{(\beta_1 - \beta_0)'E(x_i|r_i = c)}{(\delta_1 - \delta_0)'E(x_i|r_i = c)}. \tag{2.5}$$

Implicitly we assume the linear models for the conditional mean at each side of the cutoff are the same. In practice, we first use observations with $r_i < 0$ and regress y_i on x_i to estimate β_1 and, likewise, use observations with $r_i \geqq 0$ and regress y_i on x_i to estimate β_0. We apply an analogous procedure to estimate δ_0 and δ_1. Armed with these estimates and our predetermined covariates, we calculate the estimator based on Equation 5.[31] We compute the standard errors by bootstrapping non-parametrically with 500 replications. Our estimates, displayed in Figure 2.10, suggest that the impact of retention on parental satisfaction would be smallest (roughly 2 points) among students with scores well below the threshold, roughly constant (around 6

[31]To align with our estimates in section 4, we also include indicators of survey grade to estimate β_0 and β_1. Since we passed the CIA test without including them, controlling for them in the estimation does not bias our results.

points, equivalent to our RDD estimate) in the range of index scores between -5 and +2, and then slope upward until reaching 11 points for students with an index score of 6. Our confidence intervals become quite wide for estimates farther away from the cutoff, but the evidence clearly suggests larger positive treatment effects on parental satisfaction for students who passed the exams by at least 3-4 index points. Of course, these effects apply only to students who would have been retained after the process of portfolio review, summer school, and re-testing, and our estimates of δ_0 and δ_1 suggest that a relatively small fraction of these students (16%) would have been retained. Interestingly, these results are not consistent with our prior beliefs, which were that the positive effects of retention would have been greatest among students scoring well below the cutoff; these "inframarginal" students have far higher retention rates which suggested to us that school officials and parents are more likely to agree that retention would be a beneficial educational intervention for the child.

2.6 Conclusion

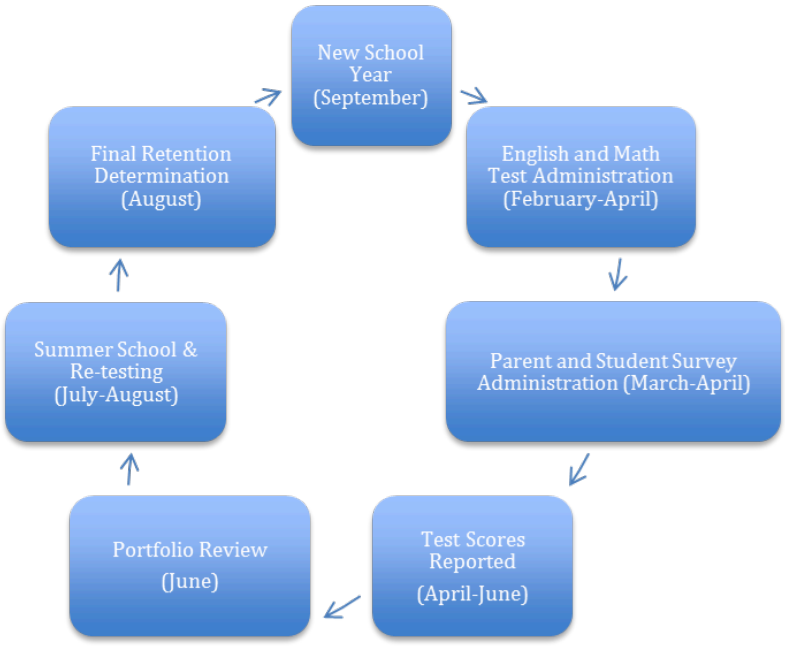
We examine variation in grade retention stemming from policies in New York City public schools which create discontinuities in the relationship between retention probability and test scores. Merging administrative data on student enrollment and testing with self-reports by students and parents about the quality of their educational experience, we contribute to the literature on the effects of retention by examining outcomes which, unlike test scores, can be easily compared across students in different grade levels. We find that students retained in NYC as a result of the district's

more stringent test-based retention policy saw significant improvements in parental satisfaction with the quality of their child's education and students' personal sense of safety. We provide evidence that suggests these effects are driven by factors beyond attending summer school, changes in age relative to classmates, or changes in performance on high stakes tests relative to same-grade peers. However, there are many additional ways in which retention can alter a student's school experience, and we lack the data to examine these other various channels.

Additionally, we use recently developed econometric methods to examine treatment effects of retention away from the cutoff and find suggestive evidence that the positive effects of this retention policy on parental satisfaction might be even greater for students scoring above the cutoff than those below. Our results thus provide an important and broadened look at the effects of grade retention. While the long-term academic and labor market outcomes of retained students are of ultimate interest, the opinions of parents and students about educational quality govern many of the educational investment decisions made in society. As such, they are an important short-term indicator on the benefits accruing to students affected by educational policy.

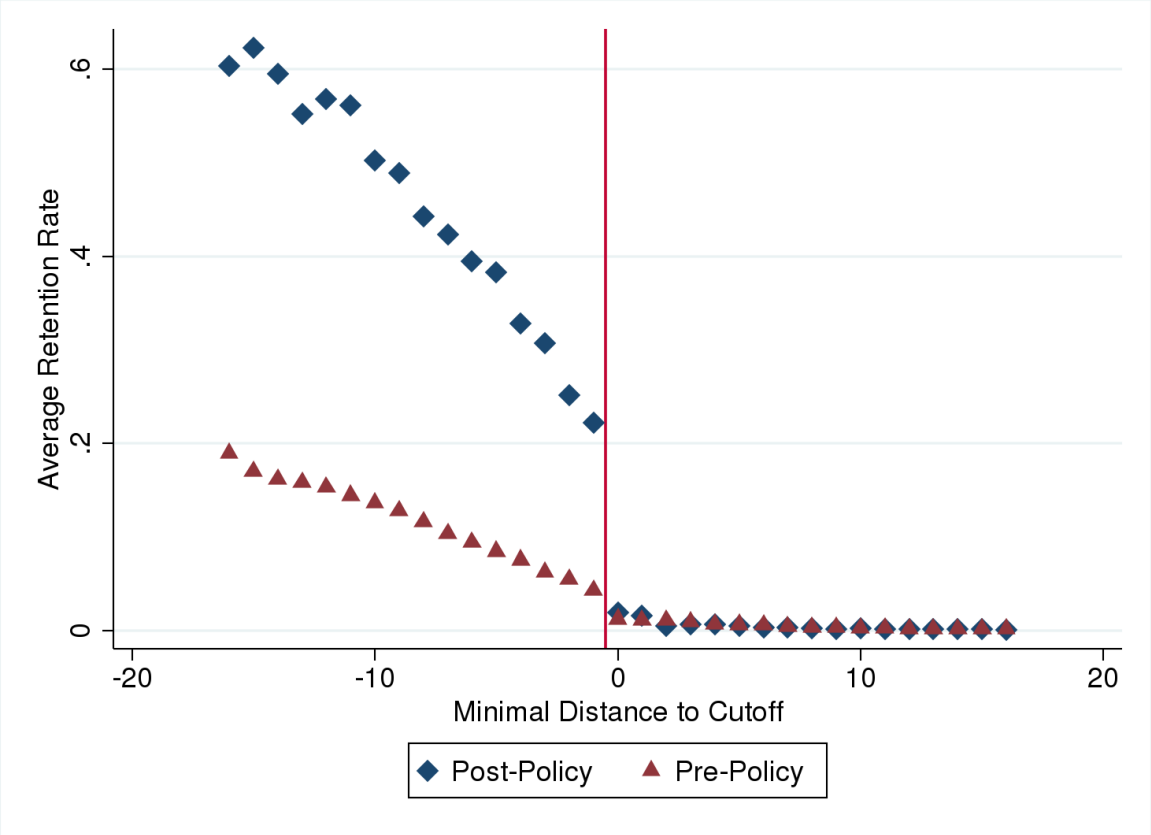
2.7 Figures

Figure 2.1: Timing and Process for Testing, Surveys, and Promotion Decisions



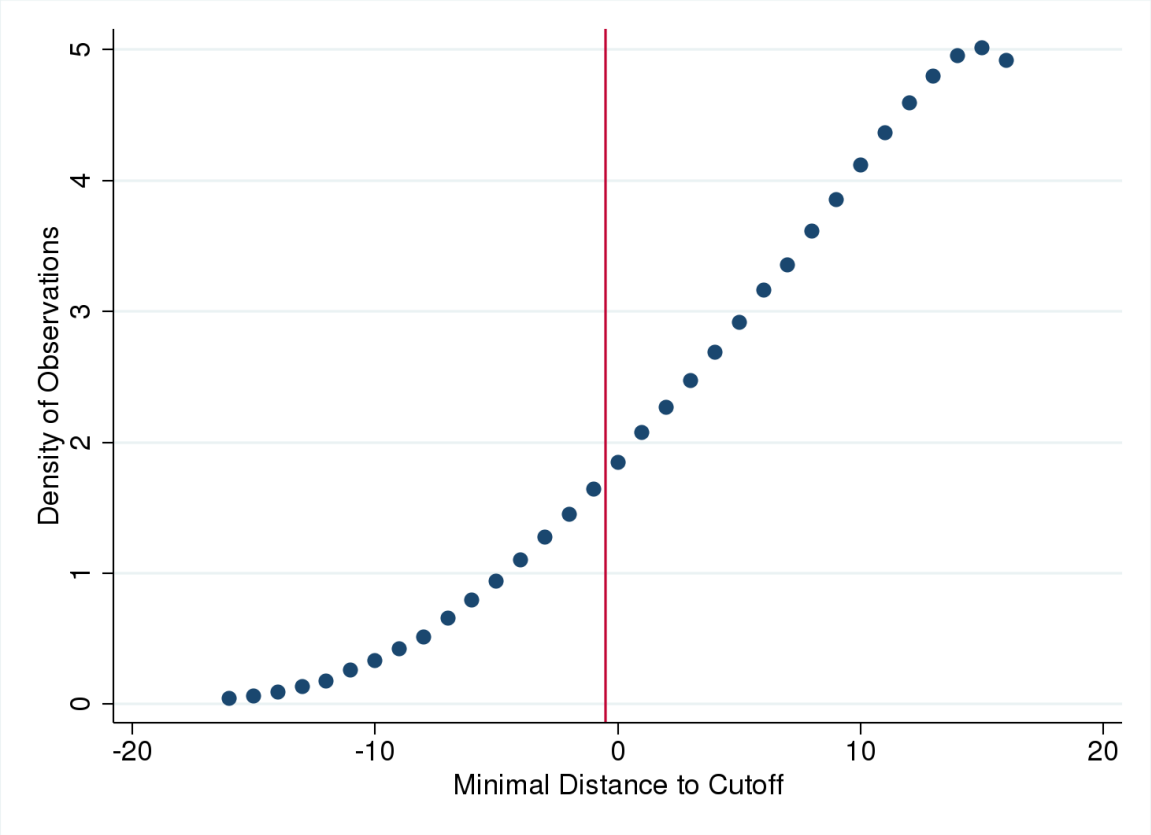
Note: Information on the timing and elements of promotion decisions is sourced largely from Crego et al. (2009).

Figure 2.2: Test-Score Based Retention Under Two Policy Regimes



Note: This figure plots the average percentage of retained students at each index score, where a score of zero is equal to the cutoff for passing both exams. Retention rates are plotted separately by policy regime, where "post-policy" designates grade-year cells that have implemented a more stringent test-based retention policy.

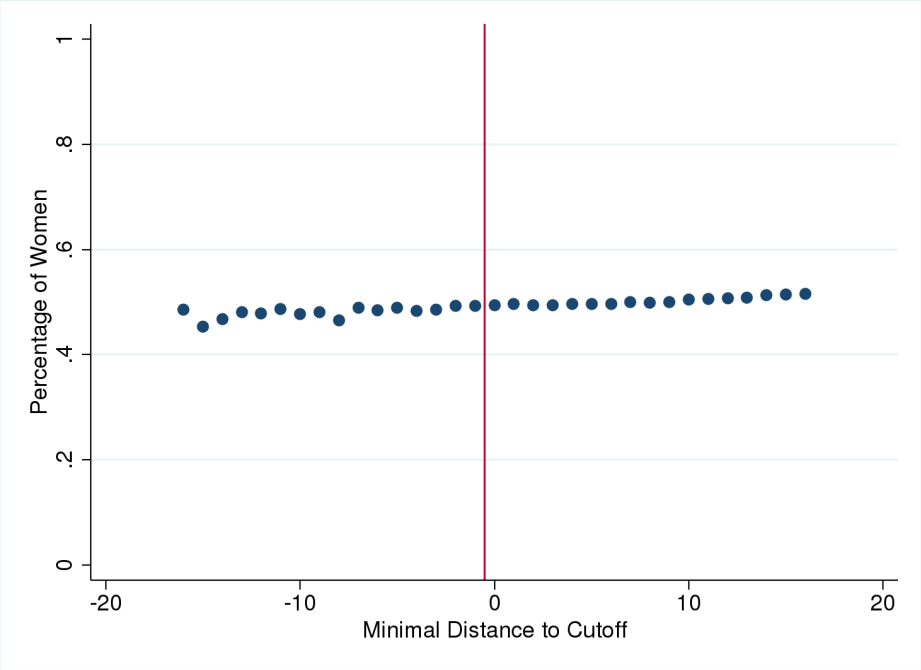
Figure 2.3: Density of Observations Across Cutoffs



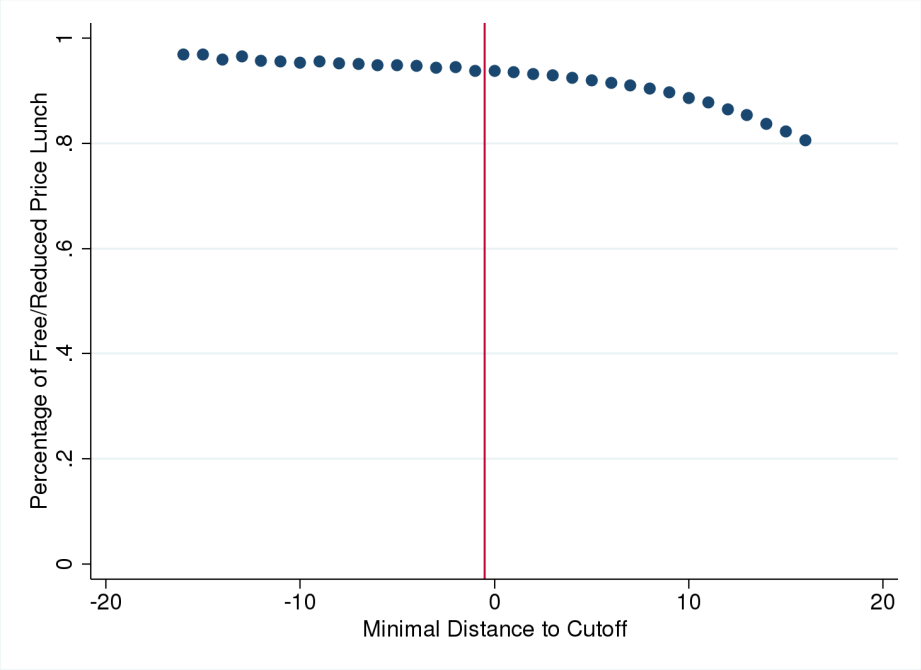
Note: Each point represents the density of student test scores at each index score in our sample.

Figure 2.4: Continuity of Covariates Across Cutoffs

(A) Female

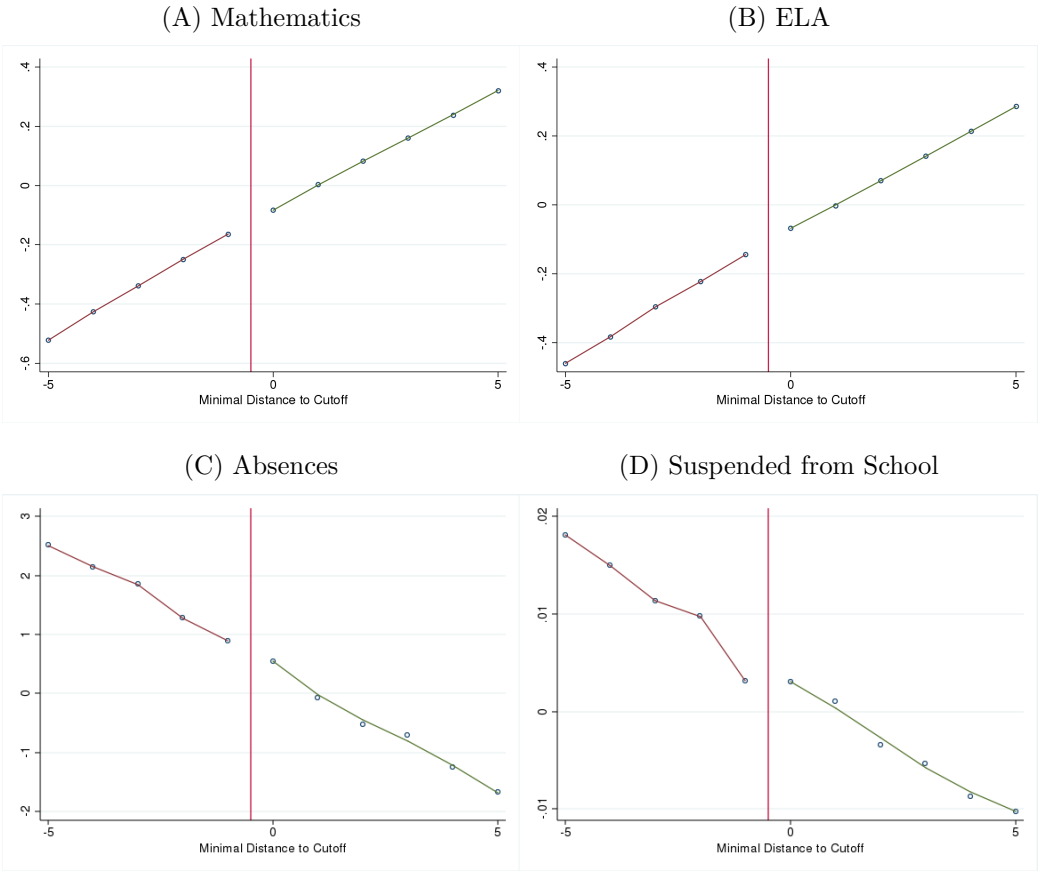


(B) Free/Reduced Price Lunch



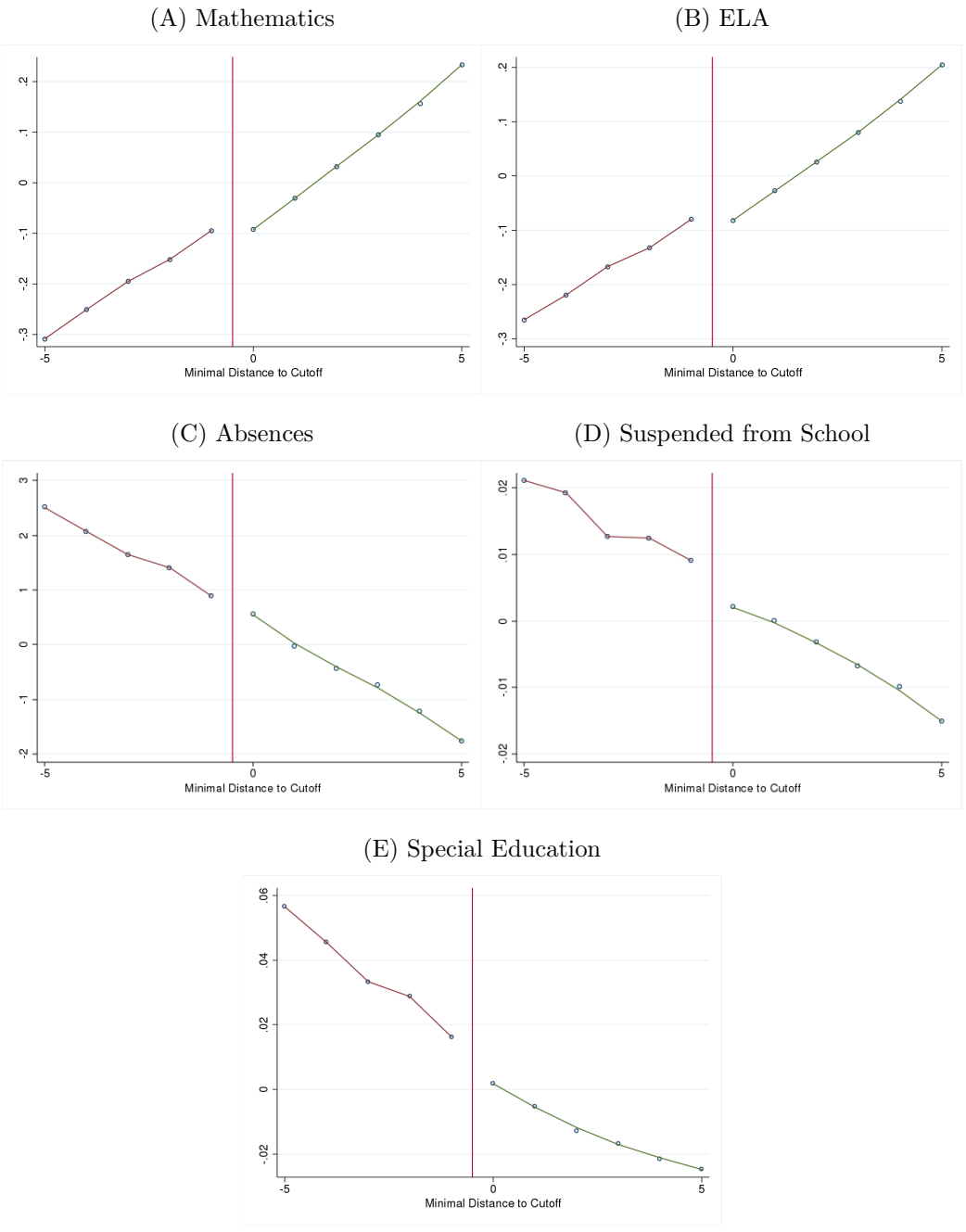
Note: Each point represents the percentage of students who are female (Panel A) or receive free/reduced price lunch (Panel B) at each index score.

Figure 2.5: Continuity of Current Test Scores, Absences, and Suspension



Note: These figures plot residuals from regressions of current test scores, absences, and an indicator for being suspended from school on test grade by test year fixed effects. ELA stands for the English Language Arts exam.

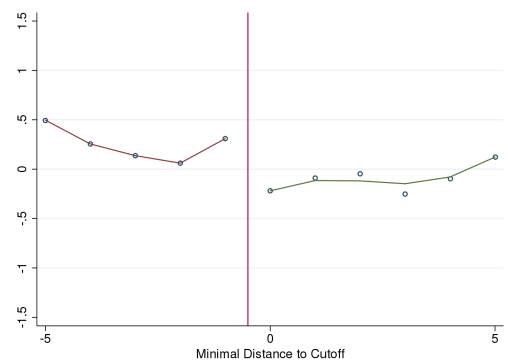
Figure 2.6: Evidence on Future Test Scores, Absences, Suspension, and Special Education



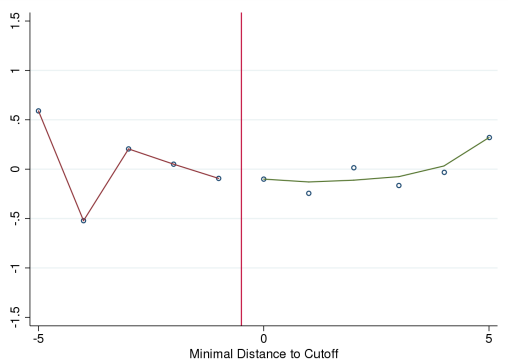
Note: These figures plot residuals from regressions of future test scores, absences, an indicator for being suspended from school, and probability of receiving special education on test grade by test year fixed effects. ELA stands for the English Language Arts exam.

Figure 2.7: Effects on Future Survey Responses

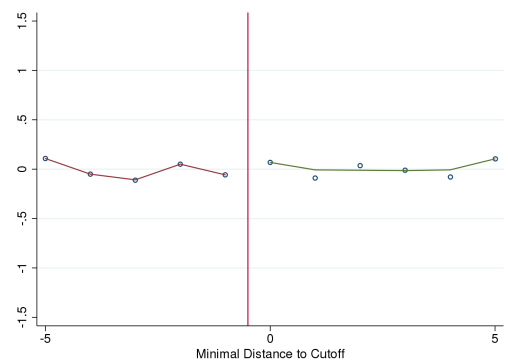
(A) Parental Satisfaction



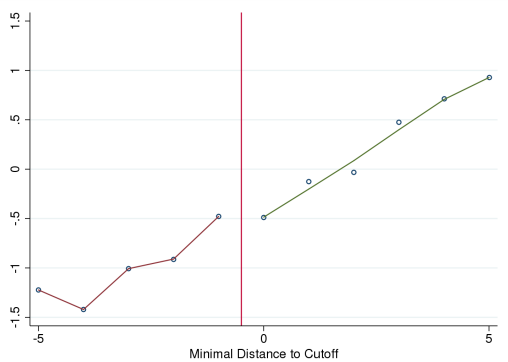
(B) Parent Sense of Overall Safety



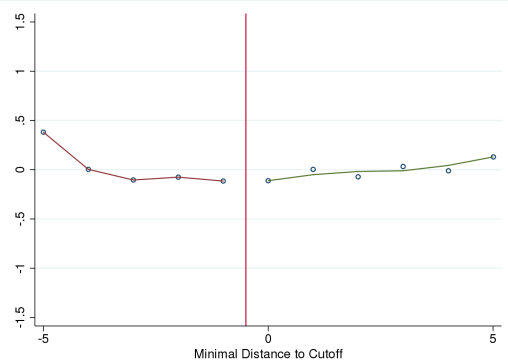
(C) Student Satisfaction



(D) Student Personal Safety



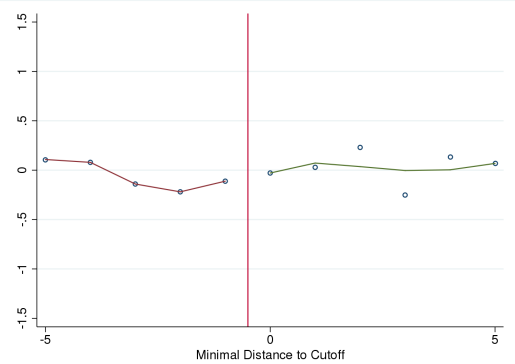
(E) Student Sense of Environment



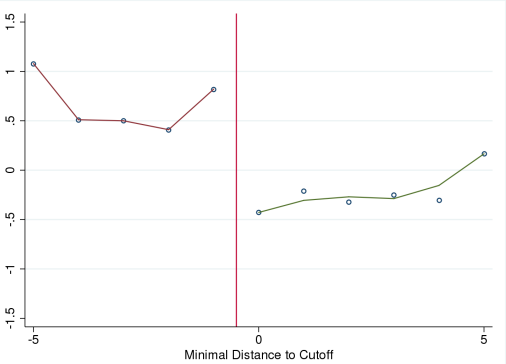
Note: These figures plot residuals from regressions of future parental satisfaction, parental sense of overall safety, student satisfaction, student safety, and student sense of environment on test grade by test year fixed effects and survey grade fixed effects.

Figure 2.8: Effects on Future Survey Responses by Policies

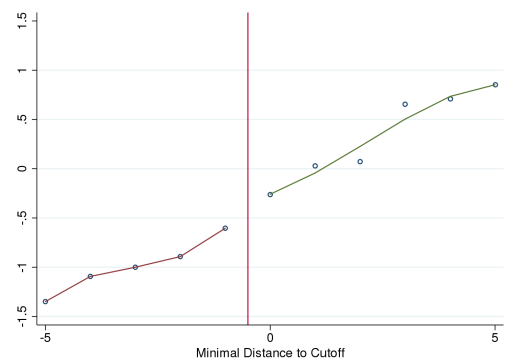
(A) Pre-policy Parental Satisfaction



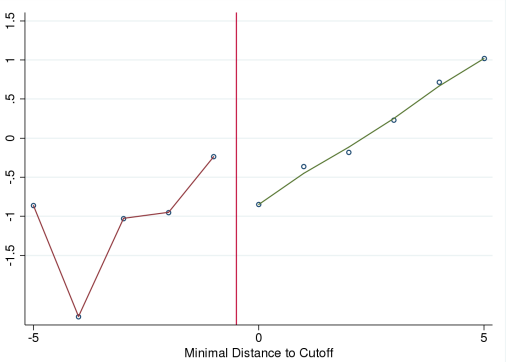
(B) Post-policy Parental Satisfaction



(C) Pre-policy Student Personal Safety



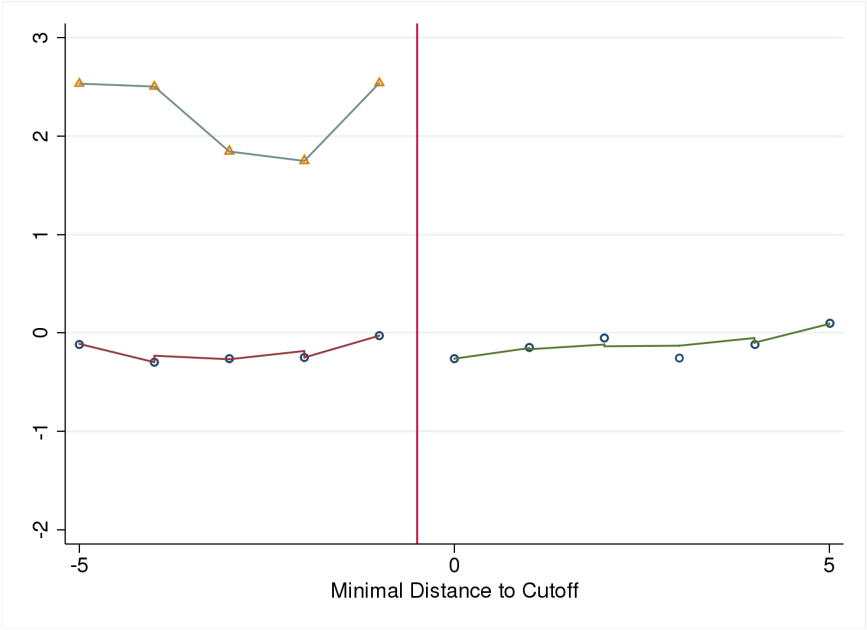
(D) Post-policy Student Personal Safety



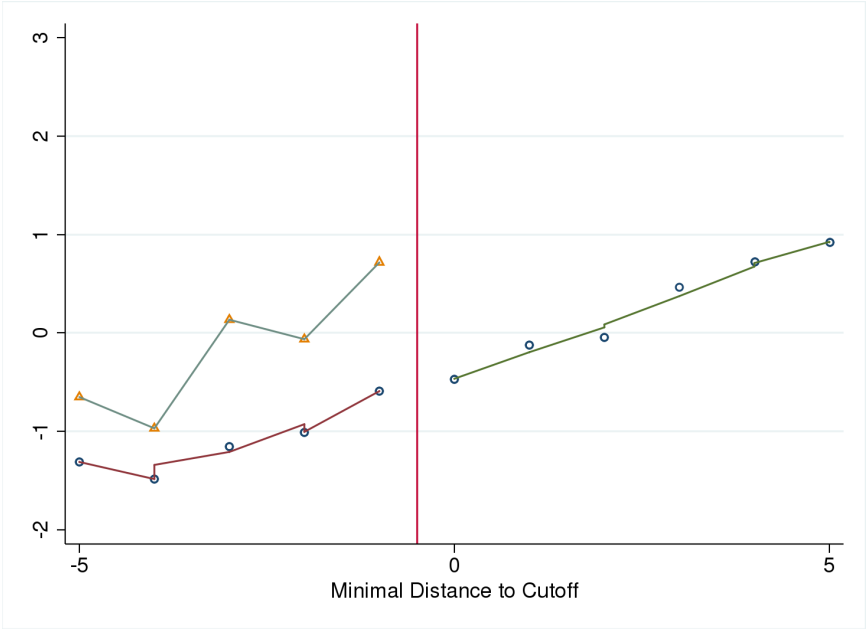
Note: These figures plot residuals from regressions of future parental satisfaction and student sense of personal safety on test grade by test year fixed effects and survey grade fixed effects. Plots are done separately for test grade-test year cells with and without the more stringent retention policy in effect.

Figure 2.9: Effects on Future Survey Responses by Actual Retention

(A) Parental Satisfaction

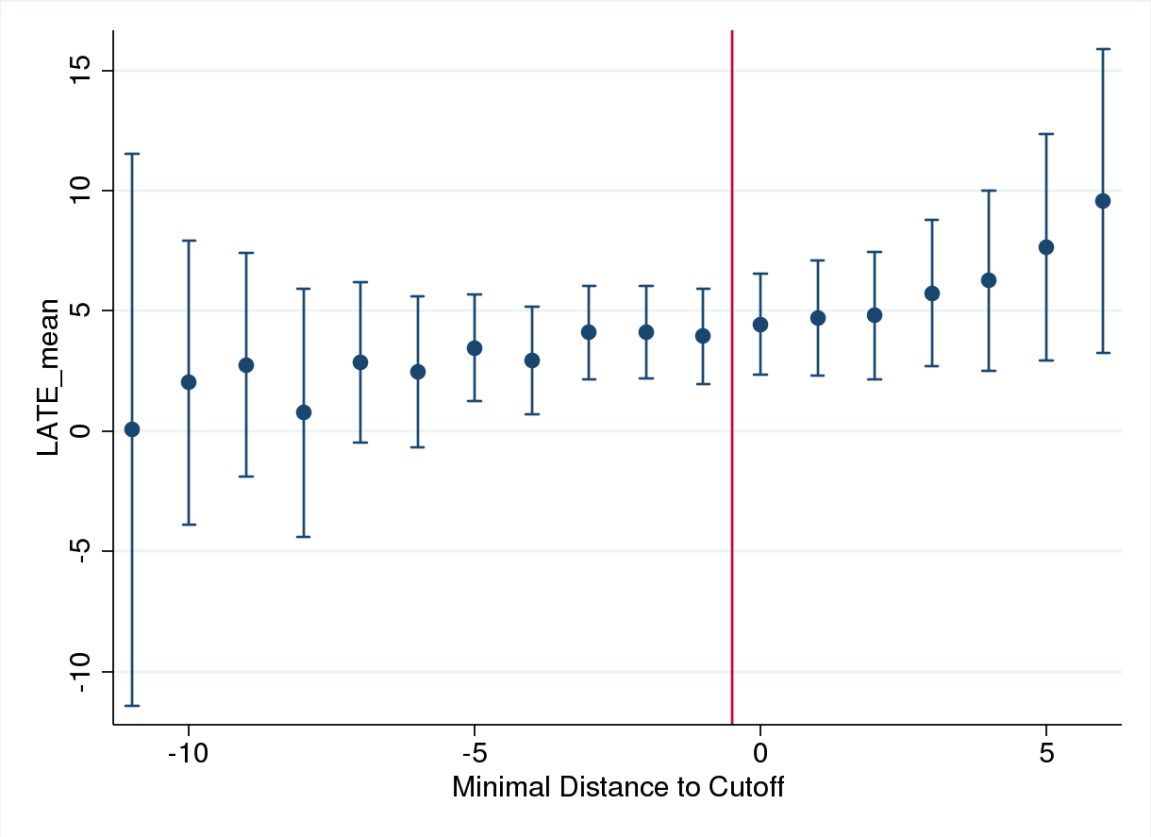


(B) Student Personal Safety



Note: These figures plot residuals from regressions of future survey responses on test grade by test year fixed effects. Plots are done separately for students who failed at least one test and were retained (yellow triangle), failed at least one of the tests but were not retained (circle on the left side of cutoff), and students who passed both tests (circle on the right side of cutoff).

Figure 2.10: CIA Estimates for Parental Satisfaction



Note: We use an estimator discussed in Kline (2011) to calculate a local average treatment effect of retention on future parental satisfaction; see the text of the paper for details. The figure plots the point estimate and its 95% confidence interval by index score.

2.8 Tables

Table 2.1: Summary Statistics

Variables	Full	RD Sample		Failing Sample	
		Below	Above	Retain	Promote
ELA Score	0.20 (0.88)	-0.8	-0.46	-1.06	-0.76
[Student-Test Observation]	[1,486,419]	[106,247]	[236,581]	[12,882]	[85,768]
Math Score	0.18 (0.89)	-0.77	-0.4	-1.01	-0.723
(Standard Deviation)	[1,493,253]	[106,247]	[236,581]	[12,882]	[85,768]
Failing ELA	6.10%	48%	0%	55%	47%
Failing Math	8.10%	65%	0%	68%	64%
Retained	2%	13%	0.80%	100%	0%
Ever Retained	5.70%	21%	9.20%	100%	10%
Ever Exempt	5.5%	10%	6.4%	15%	9.8%
Female	51%	49%	50%	47%	50%
Asian	13.20%	3.50%	5.10%	2.40%	3.70%
Hispanic	34.60%	37.80%	40%	35.60%	38.30%
Black	36.30%	52.50%	46.80%	58%	51.50%
White	15.20%	5.50%	7.40%	3.30%	5.80%
Other/Unknown	0.70%	0.70%	0.70%	0.70%	0.70%
Free Lunch	86%	95%	93%	96%	94%
Absences	12 (11.4)	17	14.7	20.6	16.3
Suspended from School	3%	5.40%	4%	7%	5%
Parental Satisfaction	74.16 (16.73)	70.13	71.82	68.42	70.54
[Number of Survey Responses]	[186,817]	[3,652]	[15,268]	[695]	[2,736]
Parent feels school is safe	80.64 (22.81)	73.97	76.6	72.59	74.34
	[170,160]	[3,289]	[13,774]	[622]	[2,466]
Student is satisfied	71.53 (15.46)	69.21	69.12	67.98	69.72
	[186,645]	[4,963]	[17,829]	[1,131]	[3,468]
Student feels safe	80.72 (20.87)	73.14	74.5	71.99	73.68
	[181,674]	[4,656]	[17,065]	[1,045]	[3,268]
Student likes environment	53.27 (17.86)	49.45	49.25	48.68	49.79
	[186,497]	[4,951]	[17,799]	[1,127]	[3,463]

Note: Test Scores are normalized within each grade × year to have a mean of zero and a standard deviation of one. Absences are capped at 50 days per year and suspension is an indicator for being suspended at least once during the school year. ELA stands for the English Language Arts exam. 2.1% of students are classified as special education in the following three years after being tested. Full Sample include every student except English learners and special education students. RD Sample includes the students in the 11 points window around the cutoff and serve as our main sample for analysis. Failing Sample includes those below the cutoff in the RD Sample.

Table 2.2: First Stage Regression Results

	Full	Survey	Parent	Student
Variables	Retention	Retention	Retention	Retention
Pre-Policy Failure	0.0334***	0.0285***	0.0302***	0.0262***
	(0.00189)	(0.00232)	(0.00369)	(0.00236)
Post-Policy Failure	0.211***	0.225***	0.215***	0.263***
	(0.0104)	(0.0131)	(0.0162)	(0.0186)
Observations	319,549	199,993	111,315	144,456
R-squared	0.167	0.170	0.188	0.141

Note: Each column reports coefficients from a single regression with grade × year control functions of student's index score. The full sample includes everyone in the RD sample, the survey sample includes everyone who or whose parent has ever responded to the survey, the parent sample includes everyone whose parent has ever responded to the survey, and the student sample includes everyone who has ever responded to the survey.

Table 2.3: Effects on Test Scores, Absences, Suspension, and Special Ed

Variable	ELA	Math	Absences	Suspension	Special Ed
Retention [placebo]	-0.00597	0.00129	-0.409	-0.00369	0
	(0.0391)	(0.0375)	(1.106)	(0.0186)	(0)
Retention [future]	0.546***	0.628***	0.533	0.0481*	0.0570**
	(0.0477)	(0.0556)	(1.537)	(0.0249)	(0.0231)
Observations	939,661	939,962	945,555	945,555	945,898
R-squared	0.190	0.214	0.035	0.021	0.043

Note: Each column reports coefficients from a single regression with grade × year control functions of student's index score. [placebo] stands for coefficients on current outcome and [future] stands for coefficients on the average outcome in the next three years after tests. Test Scores are normalized within each grade × year to have a mean of zero and a standard deviation of one. ELA stands for the English Language Arts exam. Absences are capped at 50 days per year and suspension is an indicator for being suspended at least once during the school year.

Table 2.4: Persistent Effects of Retention on Test Scores, Absences, and Suspension

Variable	Grade	ELA	Math
Retention [placebo]	0.000920	-0.00597	0.00129
	(0.00100)	(0.0391)	(0.0375)
Retention [$l = 1$]	-0.999***	0.664***	0.788***
	(0.00100)	(0.0481)	(0.0537)
Retention [$l = 2$]	-0.958***	0.447***	0.497***
	(0.0299)	(0.0694)	(0.0782)
Retention [$l = 3$]	-0.901***	0.362***	0.386***
	(0.0380)	(0.0836)	(0.100)
Observations	1,021,380	939,661	939,962
R-squared	0.991	0.194	0.215

Note: Each column reports coefficients from a single regression with grade × year control functions of student's index score. $l = 1, 2, 3$ stands for coefficients on next-year, two-year-later, and three-year-later outcomes, respectively. Test scores are normalized within each grade × year to have a mean of zero and a standard deviation of one. ELA stands for the English Language Arts exam.

Table 2.5: Effects on Survey Responses

Variable	Parent satisfied	Parent feels school is safe	Student satisfied	Student feels safe	Student likes environment
Retention [placebo]	0.235	-2.611	1.731	-6.091	-0.237
	(5.025)	(7.325)	(2.535)	(3.935)	(2.899)
Retention [future]	5.138**	-0.716	-0.113	6.133**	2.833
	(2.375)	(3.411)	(1.840)	(2.637)	(2.086)
Observations	163,594	148,330	319,109	307,465	318,533
R-squared	0.042	0.047	0.032	0.008	0.016

Note: Each column reports coefficients from a single regression with grade × year control functions of student's index score and survey grade fixed effects. [placebo] stands for coefficients on current outcome and [future] stands for coefficients on the average outcome in the next three years after tests.

Table 2.6: Persistent Effects on Survey Responses

Variable	Parent satisfied	Student feels safe
Retention [placebo]	0.235	-6.091
	(5.025)	(3.935)
Retention [$l = 1$]	-1.091	11.72*
	(6.563)	(6.577)
Retention [$l = 2$]	6.558	9.703*
	(4.862)	(4.989)
Retention [$l = 3$]	5.244	7.570
	(4.772)	(5.792)
Observations	163,594	307,465
R-squared	0.044	0.012

Note: Each column reports coefficients from a single regression with grade × year control functions of student's index score and survey grade fixed effects. [placebo] stands for coefficients on current outcome and [$l = 1, 2, 3$] stands for coefficients on next-year, two-year-later, and three-year-later outcomes, respectively.

Table 2.7: Effects on Parental Satisfaction and Students' Personal Safety between Policies

Variable	Parent satisfied	Student feels safe
Post-policy retention [placebo]	-0.211	-5.961
	(5.149)	(4.051)
Pre-policy retention [placebo]	10.63	-8.385
	(23.20)	(16.74)
Post-policy retention [future]	5.495**	6.527**
	(2.396)	(2.662)
Pre-policy retention [future]	-7.617	-8.829
	(11.82)	(12.41)
Observations	163,594	307,465
R-squared	0.038	0.006

Note: Each column reports coefficients from a single regression with grade × year control functions of student's index score and survey grade fixed effects. [future] stands for coefficients on the average outcome in the next three years after tests.

Chapter 3

Reviewing and Reassessing Children First in New York City

3.1 Introduction

When Michael Bloomberg took office as Mayor of New York City (NYC) in 2002, he was determined to restructure the existing education system and bring about radical changes to the system (Reid, 2003). He appointed Joel Klein, a lawyer from outside the education establishment, as the chancellor, and launched a series of policy changes, which are collectively known as Children First. During his 12-year tenure as the mayor, he dismantled the hierarchical education system, restructured school management, improved school finance, and reformed the workforce.

NYC is the largest school district in the United States with 1.1 million students in over 1,800 schools under the governance of New York City Department of Education (NYCDOE) (NYCDOE, 2017a). Children First significantly changed many aspects of public education in NYC, and its effect on such a large number of students is already an important subject. Moreover, Children First involved several types of education policies, and studying these policies and their interactions provides valuable lessons and experience for policy-makers in other urban school districts.

Understanding the impact of Children First is quite difficult for two reasons.

First, it affected the entire city over 12 years, and many concurrent events may confound the estimation of its impact. Second, Children First is a collection of policies, and understanding it requires examining these policies separately as well as their relationship, interconnection, and interaction.

To better understand Children First, I focus on three perspectives to analyze this education reform. First, this paper summarizes the literature on the impact of Children First as a whole and reviews studies examining the effectiveness of its associated policy changes. Second, I use the synthetic control method to provide new evidence of the overall impact of Children First on students' test scores. Third, I provide an economic analysis of how Children First affected the incentives, school management, and education inputs to understand the advantages and disadvantages of this reform.

In the first part of this paper, I present an overview of Children First and discuss studies of its overall impact. Children First can be understood as consisting of three phases. The first phase (2002-2006) mainly involved mayoral control over the city's schools and several pilot and supporting programs, which laid the foundation for further policy changes; the second phase (2007-2010) included large scale implementation of policies across several areas; and the third phase (2011-2013) was largely a continuation of the previous policies. The existing evidence (Blagg, 2016; Elwick, 2017; Kemple, 2011; Reback, 2014; UFT, 2015, p. ...) regarding the overall impact of Children First is limited, suggestive, and shows a mixed effect on student performance.

The paper then outlines three areas of key policy changes in Children First and

its associated studies. The first and main policy change provided schools with greater autonomy in school management in exchange for greater accountability pressure to improve student performance. The other two areas support the main policy by increasing financial resources and improving education workers in the district. Many studies (Rockoff and Turner, 2010; Unterman, 2014, p. ...) have examined these policies individually and tend to find positive effects.

I adopt the synthetic control method to reassess the overall impact of Children First and provide suggestive evidence of large positive effects on students' math test scores but not their English test scores. I compare NYC and a synthetic control group generated from the other counties in the New York state by matching test scores and student characteristics. However, the pre-treatment trend does not match well, and these results should be interpreted with great caution.

Lastly, I provide an economic analysis of Children First to understand its advantages and weaknesses. I show that the autonomy for accountability policy, the key component of Children First, could solve a key issue in public education: "a multi-task multi-principal nearly-monopoly organization with vague and poorly observable goals" (Dixit, 2002). Also, additional funding and human capital improvement could be complementary with autonomy, accountability, and other policies. Although the previous policies may have been shown effective in some studies, resistance from the unions and the large demand of Children First for resources may have diluted these effects and prevented this education reform from fully realizing its potential. In addition, some of these policies might be simply redistributional: they improved some students' performance while harming others'.

This paper makes three contributions. First, this paper provides up-to-date literature review of Children First and summarizes key studies on its associated policy changes. Second, this paper provides new evidence on the overall impact of Children First through the synthetic control method. Third, it includes an economic analysis to understand why some policies are successful but the overall impact is weak.

3.2 An Overview of Children First

Prior to the Bloomberg Administration, there were four levels of management in the public education system. From top to down, they were the mayoral office, the elected board of education, 32 community school districts, and the schools. The middle management was quite powerful: the board of education supervised the key policy changes, and the community school districts were directly involved in policy implementation and school management and decisions (Kelleher, 2014). This system has been criticized due to nepotism, mismanagement, and corruption (Rogers, 2009).

In 2002, Mayor Bloomberg took control of the public education in NYC, and reformed the existing district organization. The mayor applied his experience in managing private companies to the school and appointed a new chancellor, Joel Klein, to launch a large-scale school reform, known as Children First. The reform aimed to reduce bureaucracy, centralize policy-making process, and provide schools with support, autonomy, and accountability.

In the first phase of the reform (2002-2006), Mayor Bloomberg and Chancellor Klein centralized decision power and initiated a few supporting programs as prepa-

ration for launching a series of more drastic policy changes. They first replaced the elected board of education with the appointed Panel for Educational Policy, and replaced the 32 community school districts with 10 administrative regions, and later with a three-pronged organizational structure. This reorganization laid down the foundation for other policy changes. Afterward, they launched professional development programs to train new principals, reduce the prevalence of unlicensed teachers, and improve teacher quality. They also piloted a program known as Autonomy Zone to offer schools greater autonomy in exchange for greater accountability.

In the second phase (2007-2010), Mayor Bloomberg and Chancellor Klein expanded the Autonomy Zone to most of the city's schools. Mayor Bloomberg allowed the schools to have more flexibility in allocating budgets, hiring teachers, and determining professional development. However, schools, principals, and teachers were subject to performance reviews and faced severe consequences. They also adopted a new financing scheme (Fair Student Funding) to allocate additional financial resources to schools serving in-need students. A citywide data system and school support organizations were set up to support data-driven instructional practices and provide assistance to school management.

The third phase (2011-2013) was largely devoted to sustaining the prior policy changes, except for the sudden resignation of Chancellor Joel Klein, which was followed by a three-month controversial appointment of Cathleen Black, former chairwoman of Hearst Magazines (Gootman, 2010). Deputy Mayor Dennis Walcott took over and continued the reform, particularly in improving teachers' professional development and evaluation.

Children First was highly contentious, as it reformed the existing school system and promoted business practices in public education, relying on market-based competitive pressure and data-driven performance measures. For example, Giroux (2011) criticized the reform and warned that business culture could disempower students and teachers, shifting schools from developing students in a well rounded manner to focusing on test preparation. On the other hand, (Elwick, 2017; Kelleher, 2014) praised that the reform could reduce bureaucracy and provide higher quality education. Traver (2006) emphasized that this reform failed to influence school and teacher culture, a key aspect driving school effectiveness.

Quantitative evidence may provide a clearer sense of the effectiveness of Children First, but caveats apply to these pieces of evidence. Because Children First was implemented citywide, its effects may be confounded by other activities and events, such as a stronger economic recovery after the 2001 financial recession (Reback, 2014). In addition, NYC is unique in many aspects (such as size, demographics, and location), and it is difficult to find comparable cities/counties to form a reasonable comparison group.

Kemple (2011) provided probably the most compelling causal analysis of Children First. He used a comparative interrupted time series analysis to show that students' test scores and high school graduation rate increased compared to a regression-adjusted counterfactual using four other large school districts in New York State after Children First was launched in 2003. Besides the above caveats, the results on high school graduation rates are subject to miscalculation and alternative ways of accumulating high school credits (Burke, Chapman, and Monahan, 2013; Office of the State

Comptroller, 2014).

The results do not look as optimistic when we look at the test scores from the National Assessment of Educational Progress (NAEP). Blagg (2016) compared the change in students' test scores from 2005 to 2013 for all urban districts under the Trial Urban District Assessment program (TUDA) in NAEP. The average scale score across 4th and 8th grade mathematics and reading tests in NYC increased 4 points. This progress seems fine but places the city 11th out of 12 urban school districts in TUDA. Based on this result, students in NYC did not experience faster growth in test scores during this key period of Children First. It is noteworthy that other urban districts such as Chicago and Boston also implemented similar reforms, so this result might suggest that such reforms work better in other school districts.

Another caveat of such quantitative evidence is that it cannot evaluate the impact of Children First on outcomes that are difficult to observe and measure. For example, Giroux (2011) and Scott and DiMartino (2009) argue that such a reform might undercut schools' focus and efforts on improving students' social values, imagination, and civic courage, which are considered as key purposes of public education as well.

To conclude, Mayor Bloomberg and Chancellor Klein borrowed experience from business practices to design Children First, which changed many aspect of public education in NYC, from district organization to school responsibility. Because of the massive scale of Children First, it is difficult to isolate its impact from other concurrent events, and existing evidence suggests a mixed impact of this reform on student performance.

3.3 Policy Changes in Children First

This section summarizes the main policy changes in Children First and provides relevant studies to understand their impact. Table 3.1 presents the timeline of the major policies in Children First. These policy changes can be divided into three categories. The centerpiece focused on restructuring school management and promoting new school models. The two supporting pillars are remaking school budget and reforming school workforces.

Restructuring School Management

A key aspect of Children First focused on restructuring the schools and intensifying the competition across schools by closing low-performing schools, opening new and more effective schools, and increasing school choice for students. These policies may impose pressure on school leaders and induce greater school effort on students' test scores.

The city launched two key policies to provide accountability pressure to schools: Quality Review provided qualitative evaluations from external consultants, and Progress Report measured student performance in each school. Schools receiving low grades from these two policies faced severe consequences such as removal of the principal or closing the school. Rockoff and Turner (2010) used a regression-discontinuity design and showed that receipt of a low grade from Progress Report significantly increased students' test scores and improved parental evaluations of school quality.

Moreover, Mayor Bloomberg and Chancellor Klein believed that large high schools

can be inefficient and started replacing low-performing, large high schools with smaller ones. Between 2003 and 2013, NYC replaced 63 pre-existing high schools with 337 small new high schools (Robbins and Meyer, 2013). To minimize disruption in student learning, the city to the effort to gradually phase out the pre-existing schools and build up the new schools. Bloom and Unterman (2012) and Unterman (2014) used randomized assignment of students to 108 new, small high schools with excess demand to estimate the effect of attending these schools. They found that attendees experienced a 9.4 percentage points higher on-time graduation rates and are 8.4 percentage more likely to attend a postsecondary education program. Abdulkadiroğlu, Hu, and Pathak (2013) used the same technique and identified positive effects on test scores, credit accumulation, student engagement, and teacher feedback. However, Hemphill et al. (2009) reported declining attendance and graduation rates over time as well as high teacher turnover rates in these small high schools, and these schools may have diverted more low-achieving students to existing large high schools.

These results are overall encouraging but caution is required before generalizing that small high schools are effective and should replace all large high schools. First, these 108 new, small high schools in the analyses are a subset of all 337 small new high schools, and they faced excess demand from students. This suggests that these 108 schools may be higher quality, and students attending other small new high schools might not experience the same benefit. Also, Bloom, Thompson, and Unterman (2010) noted that these small schools are authorized through a competitive proposal process and received assistance and policy protections from the district. Therefore, if the number of small high schools scales up, the additional schools may be lower

quality. Finally, Schwartz, Stiefel, and Wiswall (2013) used school proximity as an instrumental variable strategy and showed that attending new small high schools, rather than the old ones, present positive effects on students, and suggest that being new, instead of being small, may be a more important factor for the success of the new, small high schools.

In addition to the new high schools, charter schools had mushroomed during the Bloomberg administration, and the number grew from 14 in 2001 to 183 in 2013. The mayor's advocate for charter schools is consistent with the philosophy behind Children First: giving schools greater autonomy, increasing school choice, and intensifying competition among schools. Hoxby, Murarka, and Kang (2009) used the lottery-based admission policy in NYC charter schools, and find that lottery winners showed substantial improvement in their academic performance, which might be due to a longer school year in these schools. The caution for the small new high schools also applies to this case: scaling up charter schools faces many challenges (Garcia, 2010) and might be subject to diminishing returns to scale.

To provide students and their parents with more choices in high schools, the city eliminated school zones and allowed them to actively make high school choices and to take the exam for the city's most selective high schools. Nevertheless, Hemphill et al. (2009) suggested that A challenge in this choice system is that many students, especially those require special education and foreign language assistance, and their parents found it difficult to navigate the choice process, and school counselors were overwhelmed by the complicated and burdensome application process. In addition, Nathanson, Corcoran, and Baker-Smith (2013) found that low-achieving students

were matched to, on average, lower-performing schools, partially because students tend to favor geography, eligibility constraint, and personal preference instead of school quality. Also, the school choice program failed to close the gaps between low-achieving students and their higher-achieving peers.

Another key policy is not as related to school restructuring but significantly changed the organization of student body and student incentives. The policy eliminated social promotion and imposed academic standards for students to be promoted to the next grade. As result, students who failed the test may repeat a grade with the next cohort and have another year to comprehend the knowledge required to be promoted. Geng and Rockoff (2016) showed that repeating a grade improved parental satisfaction of school quality and students' perceived safety in school; Geng (2018) found evidence that the policy may induce an incentive effects on students to study harder and avoid failing the test.

NYC has made sizable progress in restructuring schools but faced significant challenges in deepening these policies mainly due to the massive size of the district and the large number of disadvantaged and minority students there. Small new high schools and charter schools in NYC did show positive benefits to students, but scaling up these programs while maintaining the same quality would require prohibitive financial resources and supply of high-quality education talent. These two features also compromised the intended effects of the school choice program on more equitable access to high-achieving schools.

Remaking School Budget

Establishing new schools, providing professional development, and many other programs in Children First are resource-consuming and required more funding. NYC's strong economic recovery since 2001 and efforts in attracting private philanthropy provided strong support to these program. In addition, the city changed the school financing scheme to better accommodate schools serving in-need and disadvantaged students.

During 2002 and 2008, NYC experienced an annual increase of 10% (on average) in the city fund and tax revenues (DiNapoli, 2015). In 2007, NYC received additional funding through Contract for Excellence from New York State, which was phased-in over time. Stiefel and Schwartz (2011) showed that per-pupil revenue in the city grew dramatically during this period, from $13,290 in 2002 to $19,075 in 2008, making the city the second highest per-pupil spending school district among the largest 100 public school districts in the United States. They also found that this change reflected an increased number of special education students as well as higher teacher salaries and benefits.

In addition, Mayor Bloomberg and former Chancellor Klein made considerable efforts to attract private philanthropy as a means to support new initiatives. Between 2003 and 2009, they raised $255 million through the Fund for NYC Public Schools (ibid.), which were used to support the small new high schools, professional training, and other programs through intermediary organizations. However, total private funds only comprised of 1.3% of education expenditure in NYC, and was too small to have

a large influence on total spending.

In 2007, the city implemented Fair Student Funding, changing the school financing scheme. The old formula was largely based on teacher salaries. Because more experienced (expensive) teachers tend to work in higher-achieving schools, low-achieving schools ended up with less spending per student. The new formula depend on student characteristics and favored disadvantaged and in-need students. Therefore, schools serving students were able to receive additional funding from this scheme, and funding between low- and high-achieving schools became more equitable. The new formula was planned to phase in across years: the "winning" schools initially only received part of the additional funding while the "losing" schools were guaranteed the original funding for a few years (hold-harmless provision). Dinerstein and Smith (2014) found a 0.039 standard deviation increase in schools' value-added for math but not for English for every $1,000 increase in projected per-student funding.

However, the 2008 Great Recession brought the growth in funding to a halt. Both the state and the city underwent a sharp reduction in their budget and had to cap or even cut some of the education funding. As a result, per-pupil revenue grew by only 2% between 2009 and 2012, which was mainly driven by increasing pension payments to teachers (New York City Independent Budget Office, 2011). The freeze in school funding and the hold-harmless provision together resulted in that 94 percent of NYC schools were receiving too little money based on student need (Subramanian, 2013).

To sum up, more education funding became available due to the city's budget surpluses, philanthropy, and additional state aid. Fair Student Funding also made school funding more equitable. Nevertheless, the financial crisis in 2008 disrupted the

full phase-in of the state aid and new city funding formula, and in-need schools still fell short of funding until the next mayor took office.

Reforming School Workforce

Better trained principals and teacher may provide higher-quality education to students and facilitate the school restructuring process. Mayor Bloomberg and Chancellor Klein redesigned the personnel policies to provide high quality professional development opportunities, open the market for hiring and transferring teachers, increase teacher salaries, and impose rigorous evaluation systems on teachers and principals.

The core training program for principals was the NYC Leadership Academy, which trained new principals in business-style management and prepare them to support the low-performing schools under the reforms of Children First (O'Day and Bitter, 2010). Its flagship Aspiring Principals Program (APP) aimed at training new principals to work in lower-performing elementary and middle schools. Clark, Martorell, and Rockoff (2009) and Corcoran, Schwartz, and Weinstein (2012) found suggestive evidence that principals trained through APP had a positive impact on students' test scores over time.

The city also launched Leaders in Education Apprenticeship Program to identify eligible teachers and prepare them to transition into a principal position. In addition, the city participated in New Leaders, a national program to develop school leaders, and Principal Pipeline Initiative. Despite the tremendous effect in coaching new principals, the city still fell behind in filling the principal vacancies (Turnbull et al.,

2013). This shortage reflects an earlier comment in this paper: the scale of NYC school district is too large to fully satisfy its demand for resources.

On the teacher side, NYC scaled up NYC Teaching Fellow and recruited more teachers from Teach for America, both of which are highly selective programs. On the one hand, Boyd et al. (2008) find that the gap between the qualifications of New York City teachers in high-poverty schools and low-poverty schools has narrowed substantially since 2000. On the other hand, Kane, Rockoff, and Staiger (2008) showed that teachers from these two programs were not associated with greater teacher effectiveness, casting doubts on using these certification as a means to select teachers. Three teacher residency programs were also launched to prepare teachers to teach certain in-need student groups, such as English Language Learners.

The city also streamlined the hiring and transferring process of teachers, and gave principals greater autonomy in teacher hiring. Under the new policy, the hiring process starts earlier in the year and help minimize delays in placing new teachers. Also, veteran teachers no longer have priorities in the hiring and transferring process, and principals are able to hire based on merit instead of seniority.

To win support from United Federation of Teachers (UFT) for these changes, Mayor Bloomberg and Chancellor Klein implemented across-the-board raises in teachers' salaries with other financial incentives. For example, NYC's starting teacher salaries increased 13 percent from 2000 to 2008, and eligible veteran teachers may also become Lead Teachers to coach other teacher colleagues and receive $10,000 of additional pay (Goertz, Loeb, and Wyckoff, 2011). The city and UFT also piloted a teacher incentive program in over 200 high-need schools, but evidence shows no

effects from this program (Fryer Jr, 2013; Marsh et al., 2011).

Principals and teachers were also subject to more stringent evaluation on student performance. Principals faced possible removal of their position if their schools received poor ratings from the Progress Report and Quality Review. It was much more difficult to implement an evaluation system for teachers. In 2013, with the intervention of the New York State, the city and the union came into agreement on imposing a teacher evaluation metric, which was partially based on students' improvements in test scores (Medina, 2010).

NYC clearly made great efforts in training, hiring, evaluating, and providing incentives to teachers and principals. The evidence on the effects of these program was largely suggestive and shows at best weakly positive. These results demonstrate substantial difficulty in effective recruitment and development of education workforce. Also, the city faced strong resistance from the teacher union and part of the additional funding was used to resolve this conflict.

3.4 Reassessment of Children First

In this section, I first show that how NYC improved several important inputs to improve student learning, then present the change in student performance over this period, and lastly use the synthetic control method to estimate the overall impact of Children First on student performance.

Figure 3.1 presents the improvement in several inputs over the period of Children First. Panel A shows that since 2002, much more teachers who scored in the top-

third of state-wide SAT scores have entered NYC education workforce (Lankford et al., 2014), improving the NYC teachers' quality during this period. Consistent with Stiefel and Schwartz (2011), Panel B shows constant growth in per-pupil spending from 2002 to 2008 (when the financial crisis took place), and stagnated until 2012. Still, this growth significantly improved the funding to students. Corresponding the Panel B, class sizes in Panel C steadily decreased from 2002 to 2009, but bounced back rapidly ever since. Panel D shows that the number of charter schools increased dramatically since 2004 and quadrupled during this period. Clearly, the amount and quality of education inputs increased under Children First, although the 2008 financial crisis led to a halt in further improvement in funding and class sizes.

I use 4th and 8th test scores from the public data site in New York State Education to measure student performance. The data contain basic information of all school districts in the New York state from 1999 to 2012. Test scores are standardized at the year-grade-subject level to have zero mean and unit standard deviation.[1] I also rely on the report from NYCDOE (2015) to present the trend of high school graduation rates, but exclude them from the impact analysis because it is subject to manipulation (Burke, Chapman, and Monahan, 2013; Office of the State Comptroller, 2014).

Figure 3.2 shows the average test scores in 4th and 8th grade between 1999 and 2012. Panel A and B present the math and English test scores in 4th grade, and they had been increasing greatly over this period (from the bottom in the state to be around the average). However, the growth occurred before 2002, so the continual growth could be a result of this pre-reform trend. Panel C and D show the math and

[1] Average test scores are unavailable for 2005.

English test scores in 8th grade. The math scores increased steadily while the English scores had been flat. Figure 3.3 shows that various measures of high school graduation rates also went up during this period. Overall speaking, student performance had been improving when Children First was in place, but these figures cannot tell if these improvements are due to Children First.

Evaluating the impact of Children First is difficult. Besides potential confounding factors, choosing a reasonable comparison group is also challenging because NYC is quite different in several aspects. Table 3.2 presents the average of student characteristics and performance between NYC and other counties in the New York state. Students in NYC are much more likely to be free lunch recipient, special education recipient, minority, and low-achieving.

To account for this large difference, I adopt the synthetic control method in the hope of finding some counties or their linear combinations in the New York state that resemble NYC before Children First was implemented. I consider NYC as a single school district and other counties in the New York state as potential control groups in the donor pool. The matching covariates are average test score, percentage of students receiving free lunch, percentage of special education students, and percentage of African American, Hispanic, Asian, and White students before Children First was implemented in 2003.[2] Inference is based on assigning a treatment status to each member of the donor pool and comparing the treatment effects on the actual treated group with the placebo treatment effects on the members of the donor pool.

[2]The matching algorithm uses a Stata package developed by Abadie, Diamond, and Hainmueller (2014), which minimizes the pre-treatment mean square prediction error (MSPE).

Figure 3.4 presents the graphical evidence on the impact of Children First on students' Math and English test scores in 4th and 8th grade. The red dots plot the difference between the treatment group and the synthetic control group in each year; the gray lines plot the difference between each member in the donor pool and its synthetic control group in each year. The horizontal line is at zero; the vertical line indicates when Children First was implemented.

Panel A and B show the result on test scores in 4th grade. Before Children First was implemented in 2003, the matching between NYC and the synthetic control group is poor, indicating a lack of counties comparable to NYC. After Children First was implemented, the math scores experienced a sizable jump and continued to increase ever since, but English scores seem to stagnate. Panel C and D show the result on test scores in 8th grade. The matching works better in this case but still is not ideal. We may also observe an improvement in 8th-grade math scores but not English scores after Children First. The placebo tests show a better matching within the donor pool but not with NYC. A rough comparison between the placebo effects and the real effect seems to suggest that the effects on math scores are significant from zero.

One caveat of this analysis is that the matching covariates may not be sufficient for controlling certain variables that may drive changes in NYC after 2002. A compelling factor is that NYC experienced a stronger recovery after 2001, which might contribute to a higher growth of student performance in NYC and overstate the impact of Children First.

To conclude, I use a synthetic control method to estimate the impact of Children First on students' 4th-grade and 8th-grade test scores. There appears to be large

improvements in students' math test scores in 4th and 8th grades, but not in their English test scores. However, poor matching of the synthetic control method and some potential confounding factors limit the validity of these results.

3.5 An Economic Analysis of Children First

In this section, I provide an economic analysis to understand the impact of various policies on student learning through providing incentives and autonomy, improving inputs in education production, and potential complementarity among these factors. I also discuss challenges and obstacles facing this reform, which may counteract its effectiveness.

Incentives and Autonomy

According to Dixit (2002), public education is characterized as "a multi-task multi-principal nearly-monopoly organization with vague and poorly observable goals". NYC school district was clearly in this category: schools' enrollment was guaranteed with school zones, there were no consequences for failing to improve students' test scores, multiple stakeholders were responsible for decision-making in school operation. In this case, the goal of education was unclear, no one was fully accountable for improving schools, and no incentives were provided to do so. Efforts were weak and diversified into multiple goals; free-riding among different agents was likely.

The autonomy for accountability policy under Children First clearly changed this situation. The goal became clearer – improving students' test scores; principals be-

came solely responsible for achieving this goal; and the consequences were severe if the principals failed to achieve the goal. Therefore, principals would shoulder most of the responsibilities and respond to the punishment by increasing and concentrating efforts in improving students' test scores. Decentralization of management and greater autonomy would also provide principals with more flexibility in allocating resources, managing staff, and adjusting school operation, which may complement the accountability policy.[3] Bloom et al. (2015) also note that school autonomy, accountability, and leadership (three aspects emphasized in Children First) are associated with higher management scores and better educational outcomes.

When principals were given greater autonomy, their managerial skills became more important in affecting school performance, and differences in principals' quality may widen the performance between schools. Rice (2010) reported that principals' effectiveness does vary from one to another, and more importantly, principals possessing characteristics that are associated with higher effectiveness are less likely to work in high-poverty and low-achieving schools. This pattern raises the concern that greater autonomy might result in a widening gap in student performance between low- and high-achieving schools. The various professional development programs under Children First may alleviate this concern, which will be discussed in the next section.

Children First also increased competition across schools and further strengthened schools' incentives to improve student performance. The threat of closing low-

[3]Hong, Kueng, and Yang (2016) suggests a similar idea by showing a complementarity between performance pay and decentralized decision-making for a sample of Canadian firms.

performing schools, opening up new and charter schools, and providing a wide range of choices essentially eliminated the monopoly of zoned schools and magnified the punishment for poorly-performing schools. The real effect might be attenuated, since schools' academic performance is usually not students' and parents' foremost concern (Hastings, Kane, and Staiger, 2006; Hastings and Weinstein, 2008; Nathanson, Corcoran, and Baker-Smith, 2013).

Because public education is multi-task, accountability based on math and English may improve schools' effort on these two subjects but harm students' test scores in other subjects and development of non-cognitive skills.[4] One solution is to incorporate these measures into the accountability system as well, but West (2016) pointed out that measuring non-cognitive skills is mainly based on self-reported surveys and subject to manipulation, an issue yet to be addressed to be included in any accountability system. Also, improvements in math and English test scores do translate into better adulthood outcomes (Chetty, Friedman, and Rockoff, 2014b), which relieves the concern of over-emphasis of these tests.

To sum up, an important piece of Children First made principals the main stakeholder and provided them with incentives to improve students' test scores. Coupled with greater autonomy in school management, the principals were expected to exert greater effort and adjust management to focus on students' test scores. Although there are concerns about teaching to the test, focusing on math and English test scores seems to be the optimal goal under various constraints.

[4]For example, West et al., 2016 found that attending charter schools improve students' academic performance but harm their conscientiousness, self-control, and grit.

Inputs and Complementarity

The supporting policies under Children First focused on providing and improving several key inputs in education production. More importantly, these inputs may possibly complement the autonomy for accountability policy as well as one another, magnifying the overall impact of Children First.

A key component of Children First was to increase the overall education expenditure for students, especially those requiring additional education support. Although it is impossible to directly assess the impact of this additional funding on NYC education outcomes, research has shown that increasing education spending has a positive impact on student achievement and closing achievement gaps (Jackson, Johnson, and Persico, 2015; Lafortune, Rothstein, and Schanzenbach, 2016). This additional funding loosed the financial constraint for schools and allowed them to have the resources to, for example, provide better facilities to teachers and more remedial programs to students, which might complement the autonomy for accountability policy.[5]

Principal and teacher quality also constitutes an important part of education production, and several programs in Children First focused on recruiting more effective principal and teachers as well as improving the existing ones. Numerous studies have shown the importance of more effective principals (Clark, Martorell, and Rockoff, 2009; Day, Gu, and Sammons, 2016; Dhuey and Smith, 2014; Rice, 2010) and more effective teachers (Chetty, Friedman, and Rockoff, 2014a,b; Rivkin, Hanushek, and Kain, 2005). In addition, greater autonomy also enabled more effective principals

[5]The complementarity between education spending and school accountability makes intuitive sense but there lack evidence of it.

and teachers to better utilize their resources, adjust managerial strategy, and choose optimal instructional models. In other words, there can be complementarity between the autonomy for accountability policy and improvements in teacher/principal quality. Certainly, an open question is whether the programs in Children First improved the quality of teachers and principals in NYC, and the existing evidence provided only suggestive evidence of the effectiveness of these programs.

These supporting programs clearly targeted a few essential inputs in education production and have the potential of improving education quality in NYC. Moreover, these programs may complement the autonomy for accountability policy, further improving the overall effectiveness of Children First. However, the empirical results of these programs and their complementarity are largely unknown and worth of future research.

Challenges and Obstacles

The previous analyses have shown that the overall design of Children First and various policies targeted at the right areas and pointed at the right direction. However, the sheer size, diversity, and unions in NYC posed great challenges to Children First.

First, NYC school district is massive, and it requires an enormous amount of resources to improve it. School spending is certainly a key component, and NYC made substantial progress prior to the 2008 financial crisis. A more challenging resource is the workforce. Establishing new schools required more principals, teachers, and administrative staff. However, the short-run supply of these education workers

is inelastic, and thus expanding the demand for them would result in excessive demand, raising price (impossible due to union contracts), and decreased quality. This constraint limited the scale of Children First and might have resulted in resource competitions among schools, leading to redistribution instead of improvement in overall student performance in the city. One solution is provide students with more access to high-quality education through online learning and virtual schools (Barbour and Reeves, 2009).

Second, New York City enrolls a socioeconomically and racially diverse student population. As a result, one size can hardly fit all. For example, Herrmann (2011) showed that curriculum standardization in NYC did not produce positive effects on student performance. Also, the school choice system also imposed challenges to disadvantaged students and counselors in schools serving these students (Hemphill et al., 2009). A more tailored system may be more effective. Some other policies in Children First have done a better job at tailoring to different schools and students. For example, Fair Student Funding assigned more weight to in-need students in the funding formula; Progress Report mainly compares schools serving students of similar demographics and need.

Third, administrators' union and teacher union imposed substantial resistance to Children First, particularly with respect to workforce renumeration, recruiting, and evaluation. The support from principals and teachers came with 23 percent increase in principals' base pay (Herszenhorn, 2007) and 13 percent increase in entry-level teachers' base pay (Goertz, Loeb, and Wyckoff, 2011) as well as performance-based bonus payment. This means much of the education expenditure was used to cover

higher salaries for existing education workers, which has no direct benefits to the students in the short run. Also, since teachers and principals are rarely fired after they are tenured (Algar, 2016; Edelman, 2013), the incentives for them to exert greater effort into improving student performance are weak.

3.6 Conclusion

Mayor Bloomberg's Children First gave rise to many radical changes to school management, principals' incentives, school funding, and workforce development in New York city. To better understand this large-scale education reform in NYC, I first outline the key components of Children First and summarize important studies on its overall effectiveness. However, the evidence is suggestive, and the estimates are mixed.

Since Children First is a collection of many policy changes, understanding these policies is essential for understanding Children First. I examine these policies through reviewing key studies on the effectiveness of them. I found that these studies tend to demonstrate positive effects, which suggests a positive impact of Children First.

Adopting the synthetic control method by comparing NYC and other counties in the New York state shows that Children First may have large positive effects on math but not on English test scores. These results are consistent with the results in Chapter 1, in which the complementarity exhibits a large improvement in math scores rather than in English scores. Nevertheless, NYC is quite different from other counties, and the pre-treatment matching does not look ideal. Therefore, these results

should be considered with great caution.

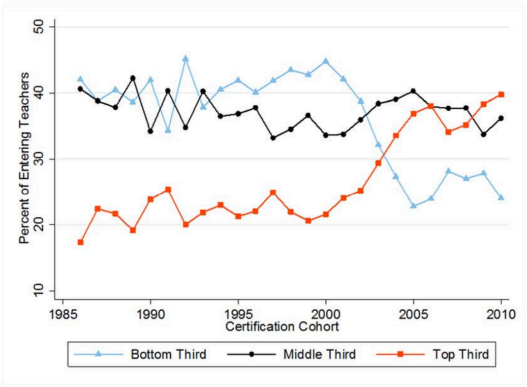
Through an economic analysis, I show that the autonomy for accountability policy, additional financial resources, and human capital development may have benefited student learning, but resource constraints, diversity of student composition, and resistance from the unions may have limited the effectiveness of Children First.

The analysis of Children First provides several suggestions to policy-makers in pubic education. These results show that certain educational policies are effective, especially when they are able to address key issues in public education (Dixit, 2002). Also, education reforms similar to Children First might be more successful in school districts where the district size is smaller, the student composition is homogeneous, unions are not as powerful, and the resource constraint is not as tight. However, each school district is different, and policy-makers need to carefully consider local students' demand, school organization, and culture to analyze and judge the effectiveness of such a reform.

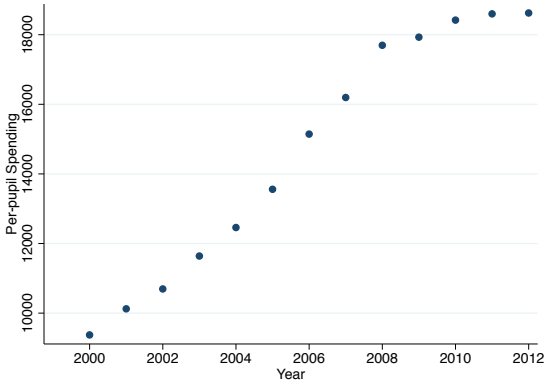
3.7 Figures

Figure 3.1: Trend of Several Key Educational Inputs

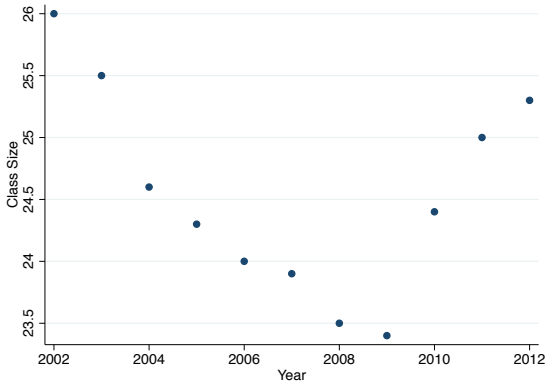
(A) Teacher SAT Scores



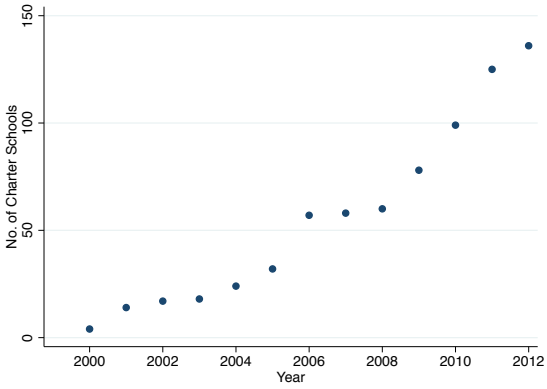
(B) Per-pupil Spending



(C) Average Class Size

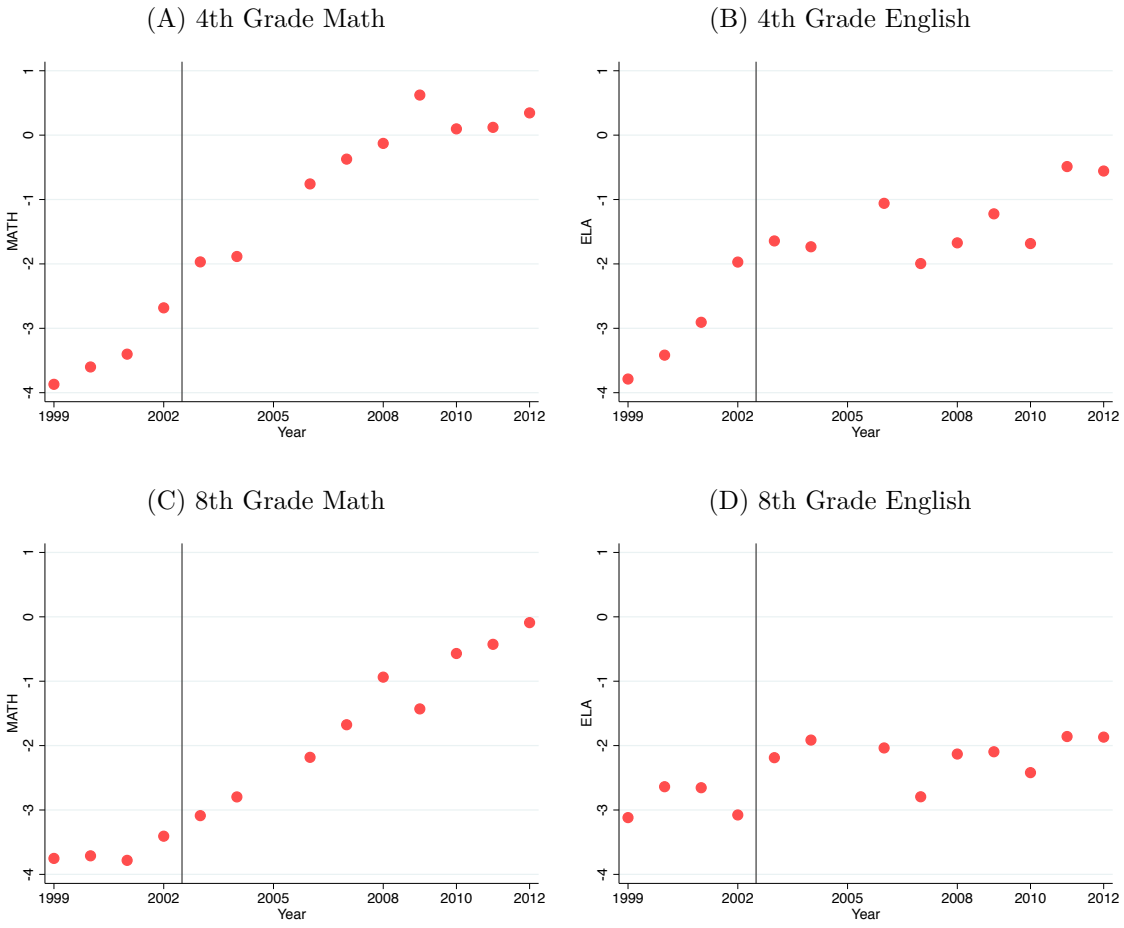


(D) Charter Schools



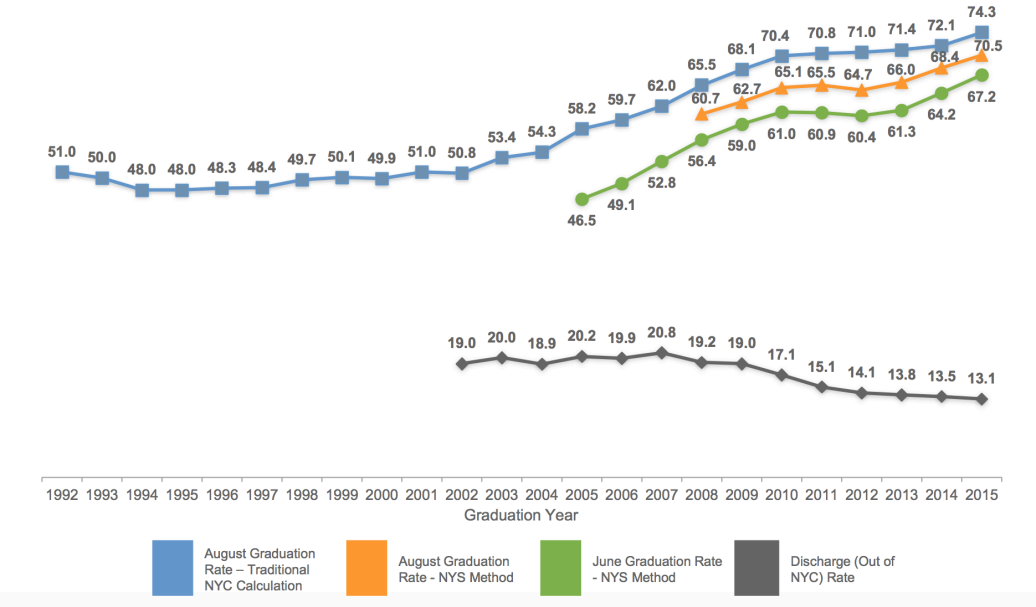
Notes: Panel A plots the percentage of entering teachers in NYC drawn from bottom, middle, and top thirds of state-wide score distribution (on national SATs) (Lankford et al., 2014). Panel B plots the average student spending from the school-based expenditure reports (NYCDOE, 2016). Panel C plots the average class size in 4th grade from the Class Size Report (NYCDOE, 2017b). Panel D plots the number of charter schools each year in NYC (The New York City Charter School Center, 2012).

Figure 3.2: Trend of NYC Student Test Scores



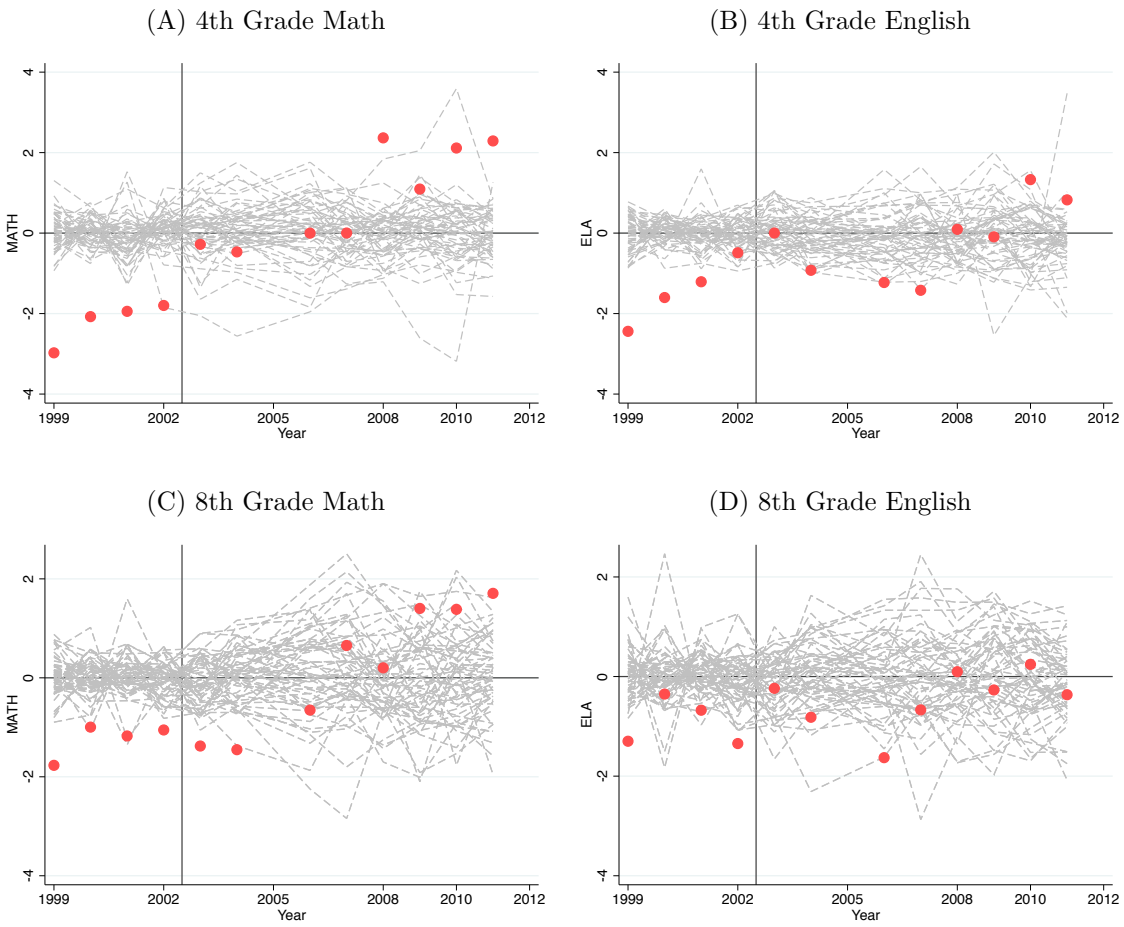
Notes: All panels are based on data from 1999 to 2012. The red dots plot the raw data of NYC student outcomes in each year. The dependent variables in Panels A and B are math and English test scores in 4th grade; the dependent variables in Panels C and D are math and English test scores in 8th grade. To the right of the black line are years after Children First was implemented.

Figure 3.3: Trend of NYC Student High School Graduation Rate



Notes: Figure plots the graduation rates by various standards in each year (NYCDOE, 2015).

Figure 3.4: Overall Impact of Children First



Notes: All panels are based on data from 1999 to 2012. The red dots plot the difference between the treatment group and the synthetic control group in each year; the gray lines plot the difference between each member in the donor pool and its synthetic control group in each year. The dependent variables in Panels A and B are math and English test scores in 4th grade; the dependent variables in Panels C and D are math and English test scores in 8th grade. The horizontal line is at zero. To the right of the black line are years after Children First was implemented.

3.8 Tables

Table 3.1: Chronology of Major Policies under Children First

Year	Major Policies
2002	Mayor Bloomberg took office and gained control of NYC's schools Small high school movement and charter school expansion initiated
2003	10 administrative regions replaced 32 community school districts Math and reading curricula were standardized The NYC Leadership Academy to train and support school leaders
2004	Autonomy Zone as a pilot program of the accountability for autonomy policy A universal high school choice process Grade retention policy holds back students who fall behind academically
2006	Quality Review to provide schools with qualitative evaluation
2007	School Accountability Scheme to evaluate schools based on student performance Fair Student Funding to allocate more funding based on student needs
2013	Teacher ratings tied with student growth on state tests

Table 3.2: Summary Statistics

	New York City	Other Counties in NY State
Percent of Free Lunch Recipient	66.23	25.24
Percent of Special Education Students	14.85	1.458
Percent of African American Students	32.42	5.979
Percent of Hispanic Students	39.19	4.355
Percent of Asian Students	13.50	1.757
Percent of White Students	14.88	87.94
Average Math Score	-1.744	0.0306
Average ELA Score	-2.113	0.0371

Notes: The table presents the means of variables for New York City vs. other counties in the New York state.

Bibliography

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller (2010). "Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program." In: Journal of the American Statistical Association 105.490, pp. 493–505.

— (2014). SYNTH: Stata module to implement Synthetic Control Methods for Comparative Case Studies.

— (2015). "Comparative politics and the synthetic control method." In: American Journal of Political Science 59.2, pp. 495–510.

Abdulkadiroğlu, Atila, Weiwei Hu, and Parag A Pathak (2013). Small high schools and student achievement: Lottery-based evidence from New York City. Tech. rep. National Bureau of Economic Research.

Algar, Selim (2016). It's basically impossible to fire a New York City school teacher. https://nypost.com/2016/12/08/its-basically-impossible-to-fire-a-new-york-city-school-teacher/.

Almond, Douglas and Bhashkar Mazumder (2013). "Fetal Origins and Parental Responses." In: Annu. Rev. Econ. 5.1, pp. 37–56.

Anderson, Gabrielle E., Angela D. Whipple, and Shane R. Jimerson (2005). "Student Ratings of Stressful Experiences at Home and School: Loss of a Parent and Grade Retention as Superlative Stressors." In: Journal of Applied School Psychology 21.1, pp. 1–20.

Andrew, Megan (2014). "The scarring effects of primary-grade retention? A study of cumulative advantage in the educational career." In: Social Forces, p. 74.

Angrist, Joshua and Miikka Rokkanen (2015). "Wanna Get Away? Regression Discontinuity Estimation of Exam School Effects Away from the Cutoff." In: Journal of the American Statistical Association 110.512, pp. 1331–1344.

Athey, Susan and Scott Stern (1998). An empirical framework for testing theories about complimentarity in organizational design. Tech. rep. National Bureau of Economic Research.

Barbour, Michael K and Thomas C Reeves (2009). "The reality of virtual schools: A review of the literature." In: Computers & Education 52.2, pp. 402–416.

Behrman, Jere R et al. (2015). "Aligning learning incentives of students and teachers: Results from a social experiment in Mexican high schools." In: Journal of Political Economy 123.2, pp. 325–364.

Blagg, Kristin (2016). Making the Grade in America's Cities: Assessing Student Achievement in Urban Districts. Tech. rep. Urban Institute.

Bloom, Howard S, Saskia Levy Thompson, and Rebecca Unterman (2010). "Transforming the High School Experience: How New York City's New Small Schools Are Boosting Student Achievement and Graduation Rates." In: MDRC.

Bloom, Howard S and Rebecca Unterman (2012). "Sustained Positive Effects on Graduation Rates Produced by New York City's Small Public High Schools of Choice. Policy Brief." In: MDRC.

Bloom, Nicholas et al. (2015). "Does management matter in schools?" In: The Economic Journal 125.584, pp. 647–674.

Boyd, Donald et al. (2008). "The narrowing gap in New York City teacher qualifications and its implications for student achievement in high-poverty schools." In: Journal of Policy Analysis and Management 27.4, pp. 793–818.

Brynjolfsson, Erik and Paul Milgrom (2013). "Complementarity in organizations." In: The Handbook of Organizational Economics, pp. 11–55.

Bui, Sa A, Steven G Craig, and Scott A Imberman (2014). "Is Gifted Education a Bright Idea? Assessing the Impact of Gifted and Talented Programs on Students." In: American Economic Journal: Economic Policy 6.3, pp. 30–62.

Burke, K, B Chapman, and R Monahan (2013). "Critics blast credit recovery as city data reveals frequent use by public high school students." In: The New York Daily News.

Byrnes, Deborah A (1989). "Attitudes of students, parents and educators toward repeating a grade." In: Flunking grades: The policies and effects of retention.

Campanile, Carl (2004). Pols Give School Reform An F. English. Copyright - (Copyright 2004, The New York Post. All Rights Reserved; Last updated - 2012-01-26.

Card, David et al. (2012). "Inequality at Work: The Effect of Peer Salaries on Job Satisfaction." In: The American Economic Review 102.6, pp. 2981–3003.

Cascio, Elizabeth U and Douglas O Staiger (2012). Knowledge, Tests, and Fadeout in Educational Interventions. Tech. rep. National Bureau of Economic Research.

Chetty, Raj, John N Friedman, and Jonah E Rockoff (2014a). "Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates." In: The American Economic Review 104.9, pp. 2593–2632.

— (2014b). "Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood." In: The American Economic Review 104.9, pp. 2633–2679.

Chetty, Raj et al. (2011). "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project Star*." In: Quarterly Journal of Economics 126.4.

Chiang, Hanley (2009). "How accountability pressure on failing schools affects student achievement." In: Journal of Public Economics 93.9, pp. 1045–1057.

Clark, Damon, Paco Martorell, and Jonah Rockoff (2009). "School Principals and School Performance. Working Paper 38." In: National Center for Analysis of longitudinal data in Education research.

Corcoran, Sean P, Amy Ellen Schwartz, and Meryle Weinstein (2012). "Training your own: The impact of New York City's aspiring principals program on student achievement." In: Educational Evaluation and Policy Analysis 34.2, pp. 232–253.

Crego, Al et al. (2009). Ending Social Promotion Without Leaving Children Behind: The Case of New York City. Tech. rep. Santa Monica, CA: RAND Corporation.

Day, Christopher, Qing Gu, and Pam Sammons (2016). "The impact of leadership on student outcomes: How successful school leaders use transformational and instructional strategies to make a difference." In: Educational Administration Quarterly 52.2, pp. 221–258.

Dee, Thomas S et al. (2016). The Causes and Consequences of Test Score Manipulation: Evidence from the New York Regents Examinations. Tech. rep. National Bureau of Economic Research.

Deming, David J et al. (2016). "School accountability, postsecondary attainment, and earnings." In: Review of Economics and Statistics 98.5, pp. 848–862.

Deming, David and Susan Dynarski (2008). "The Lengthening of Childhood." In: Journal of Economic Perspectives 22.3, pp. 71–92.

Dhuey, Elizabeth and Justin Smith (2014). "How important are school principals in the production of student achievement?" In: Canadian Journal of Economics/Revue canadienne d'économique 47.2, pp. 634–663.

DiNapoli, Thomas P. (2015). Review of the Financial Plan of the City of New York. Tech. rep. New York State Office of the State Comptroller.

Dinerstein, Michael, Troy Smith, et al. (2014). Quantifying the Supply Response of Private Schools to Public Policies. Tech. rep.

Dixit, Avinash (2002). "Incentives and organizations in the public sector: An interpretative review." In: Journal of human resources, pp. 696–727.

Dobbs, Michael (2004). Ready for Fourth Grade? Not So Fast, New York Says; Policy Against Social Promotion Draws Fire From Some Groups. English. Copyright - Copyright The Washington Post Company Jul 7, 2004; People - Bloomberg, Michael; Last updated - 2010-08-05.

Duflo, Esther, Pascaline Dupas, and Michael Kremer (2011). "Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya." In: The American Economic Review 101, pp. 1739–1774.

Edelman, Susan (2013). Failing NYC school principals are rarely fired. https://nypost.com/2013/06/16/failing-nyc-school-principals-are-rarely-fired/.

Eide, Eric R. and Mark H. Showalter (2001). "The Effect of Grade Retention on Educational and Labor Market Outcomes." In: Economics of Education Review 20, pp. 563–576.

Elwick, Alex (2017). "Education reform in New York City (2002–2013)." In: Oxford Review of Education 43.6, pp. 677–694.

Eren, Ozkan, Briggs Depew, and Stephen Barnes (2017). "Test-Based Promotion Policies, Dropping Out, and Juvenile Crime." In: Journal of Public Economics.

Figlio, David, Susanna Loeb, et al. (2011). "School accountability." In: Handbook of the Economics of Education 3.8, pp. 383–417.

Fryer Jr, Roland G (2013). "Teacher Incentives and Student Achievement: Evidence from New York City Public Schools." In: Journal of Labor Economics 31.2, pp. 373–407.

Garcia-Pereza, J. Ignacio, Marisa Hidalgo-Hidalgoa, and J. Antonio Robles-Zurita (2014). "Does Grade Retention Affect Students' Achievement? Some Evidence from Spain." In: Applied Economics 46.12, pp. 1373–92.

Garcia, Peggie (2010). Scaling up High-Quality Charter Schools. Tech. rep. National Charter School Resource Center.

Geng, Tong (2018). The Complementarity of Incentive Policies in Education: Evidence from New York City. Tech. rep. Columbia University.

Geng, Tong and Jonah E Rockoff (2016). Does Repeating a Grade Make Students (and Parents) Happier? Regression Discontinuity Evidence from New York City. Tech. rep. Columbia University.

Giroux, Henry A (2011). "Business Culture and the Death of Public Education: Mayor Bloomberg, David Steiner, and the politics of corporate 'leadership'." In: Policy Futures in Education 9.5, pp. 553–559.

Goertz, Margaret, Susanna Loeb, and Jim Wyckoff (2011). "Recruiting, evaluating and retaining teachers: The children first strategy to improve New York City's teachers." In: Education reform in New York City: Ambitious change in the nation's most complex school system, pp. 157–177.

Gootman, Elissa (2004). Test Policy for 3rd Graders Is Met by More Resistance. English. Copyright - Copyright New York Times Company Feb 11, 2004; Document feature - photographs; People - Bloomberg, Michael; Last updated - 2010-06-29; CODEN - NYTIAO.

— (2010). "Frustrations with Mayor Are Backdrop to Nominee Uproar." In: The New Tork Times (November 25, 2010) A 28.

Hastings, Justine S, Thomas J Kane, and Douglas O Staiger (2006). Preferences and heterogeneous treatment effects in a public school choice lottery. Tech. rep. National Bureau of Economic Research.

Hastings, Justine S and Jeffrey M Weinstein (2008). "Information, school choice, and academic achievement: Evidence from two experiments." In: The Quarterly journal of economics 123.4, pp. 1373–1414.

Hemphill, Clara et al. (2009). The new marketplace: How small-school reforms and school choice have reshaped New York City's high schools. New School Center for New York City Affairs.

Herrmann, Mariesa (2011). "One Size Fits All? The effect of curriculum standardization on student achievement." In:

Herszenhorn, DM (2007). "Bloomberg reaches deal with principals." In: New York Times.

Herszenhorn, David M. (2004). Stricter Standards in New York May Hold 15,000 in 3rd Grade. English. Copyright - Copyright New York Times Company Jan 9, 2004; Last updated - 2010-06-29; CODEN - NYTIAO.

Holmstrom, Bengt and Paul Milgrom (1994). "The firm as an incentive system." In: The American Economic Review, pp. 972–991.

Hong, Bryan, Lorenz Kueng, and Mu-Jeung Yang (2016). "Complementarity of Performance Pay and Task Allocation."

Hoxby, Caroline M, Sonali Murarka, and Jenny Kang (2009). "How New York City's charter schools affect achievement." In: Cambridge, MA: New York City Charter Schools Evaluation Project, pp. 1–85.

Jackson, C Kirabo, Rucker C Johnson, and Claudia Persico (2015). "The effects of school spending on educational and economic outcomes: Evidence from school finance reforms." In: The Quarterly Journal of Economics 131.1, pp. 157–218.

Jacob, Brian A. and Lars Lefgren (2004a). "Remedial Education and Student Achievement: A regression-discontinuity analysis." In: The Review of Economics and Statistics 86.1, pp. 226–244.

— (2004b). "Remedial Education and Student Achievement: an RD Analysis." In: The Review of Economics and Statistics 86.1, pp. 226–244.

— (2009). "The Effect of Grade Retention on High School Completion." In: American Economic Journal: Applied Economics 1.3, pp. 33–58.

Jacob, Brian A and Jonah E Rockoff (2012). "Organizing schools to improve student achievement: Start times, grade configurations, and teacher assignments." In: The Education Digest 77.8, p. 28.

Johnson, Rucker C and C Kirabo Jackson (2017). Reducing Inequality Through Dynamic Complementarity: Evidence from Head Start and Public School Spending. Tech. rep. National Bureau of Economic Research.

Kane, Thomas J, Jonah E Rockoff, and Douglas O Staiger (2008). "What does certification tell us about teacher effectiveness? Evidence from New York City." In: Economics of Education review 27.6, pp. 615–631.

Kelleher, Maureen (2014). "New York City's Children First: Lessons in School Reform." In: Center for American Progress.

Kemple, James J (2011). "Children First and student outcomes: 2003-2010." In: Education reform in New York City: Ambitious change in the nation's most complex school system, pp. 255–292.

Kline, Patrick (2011). "Oaxaca-Blinder as a Reweighting Estimator." In: American Economic Review: Papers and Proceedings 101, pp. 532–537.

Koles r, Michal and Christoph Rothe (2016). "Inference in Regression Discontinuity Designs with a Discrete Running Variable."

Koppensteiner, Martin Foureaux (2014). "Automatic Grade Promotion and Student Performance: Evidence from Brazil." In: Journal of Development Economics 107, pp. 277–290.

Ladd, Helen F and Douglas L Lauen (2010). "Status versus growth: The distributional effects of school accountability policies." In: Journal of Policy Analysis and Management 29.3, pp. 426–450.

Lafortune, Julien, Jesse Rothstein, and Diane Whitmore Schanzenbach (2016). School finance reform and the distribution of student achievement. Tech. rep. National Bureau of Economic Research.

Lankford, Hamilton et al. (2014). "Who enters teaching? Encouraging evidence that the status of teaching is improving." In: Educational Researcher 43.9, pp. 444–453.

Lavy, Victor, M Daniele Paserman, and Analia Schlosser (2012). "Inside the black box of ability peer effects: Evidence from variation in the proportion of low achievers in the classroom." In: The Economic Journal 122.559, pp. 208–237.

Lee, David S. and David Card (2008). "Regression discontinuity inference with specification error." In: Journal of Econometrics 142, pp. 655–674.

Macartney, Hugh, Robert McMillan, and Uros Petronijevic (2015). Incentive design in education: An empirical analysis. Tech. rep. National Bureau of Economic Research.

Malamud, Ofer, Cristian Pop-Eleches, and Miguel Urquiola (2016). Interactions Between Family and School Environments: Evidence on Dynamic Complementarities? Tech. rep. National Bureau of Economic Research.

Manacorda, Marco (2012). "The Cost of Grade Retention." In: The Review of Economics and Statistics 94.2, pp. 596–606.

Mariano, Louis T. and Paco Martorell (2013). "The Academic Effects of Summer Instruction and Retention in New York City." In: Educational Evaluation and Policy Analysis 35.1, pp. 96–117.

Marsh, Julie A et al. (2011). A big apple for educators: New York City's experiment with schoolwide performance bonuses: Final evaluation report. Rand Corporation.

Mbiti, Isaac et al. (2016). "Inputs, Incentives, and Complementarities in Primary Education: Experimental Evidence from Tanzania." In: Unpublished Paper.

McCombs, Jennifer Sloan, Sheila Nataraj Kirby, and Louis T Mariano (2009). Ending social promotion without leaving children behind: The case of New York City. Rand Corporation.

Medina, Jennifer (2010). "Agreement will alter teacher evaluations." In: New York Times, p. 17.

Milgrom, Paul and John Roberts (1995). "Complementarities and fit strategy, structure, and organizational change in manufacturing." In: Journal of Accounting and Economics 19.2, pp. 179–208.

Murphy, Richard and Felix Weinhardt (2014). "Top of the Class: The Importance of Ordinal Rank." In: CESifo Working Paper Series.

NYCDOE (2015). New York City Graduation Rates Class of 2015 (2011 Cohort). http : / / schools . nyc . gov / NR / rdonlyres / AB773209 - 5CDE - 49F2 - 87CE - 2B10F5837F7B/0/2015GraduationRatesWebsite11116.pdf.

— (2016). School-Based Expenditure Reports. http : / / schools . nyc . gov / Offices / DBOR/SBER/default.htm.

— (2017a). About Us. http://schools.nyc.gov/AboutUs/default.htm.

— (2017b). Class Size Report. http : / / schools . nyc . gov / AboutUs / schools / data / classsize/classsize.htm.

Nathanson, Lori, Sean Corcoran, and Christine Baker-Smith (2013). "High School Choice in New York City: A Report on the School Choices and Placements of Low-Achieving Students." In: Research Alliance for New York City Schools.

Nathanson, Lori, Meghan McCormick, and James J. Kemple (2013). Strengthening Assessments of School Climate: Lessons from the New York City School Survey. Brief. Research Alliance for NYC Schools.

Neal, Derek and Diane Whitmore Schanzenbach (2010). "Left behind by design: Proficiency counts and test-based accountability." In: The Review of Economics and Statistics 92.2, pp. 263–283.

New York City Independent Budget Office (2011). New York City Public School Indicators: Demographics, Resources, Outcomes. Tech. rep.

O'Day, Jennifer and Catherine Bitter (2010). "Improving Instruction in New York City Schools: An Evolving Strategy." In: Education Reform in New York City. Harvard Education Press. Cambridge, MA.

Office of the State Comptroller (2014). New York City Department of Education: Accuracy of Reported Discharge Data. Tech. rep. Division of State Government Accountability, New York State.

Ozek, Umut (2015). "Hold Back To Move Forward? Early Grade Retention And Student Misbehavior." In: Education Finance and Policy 10.3, pp. 350–377.

Reback, Randall (2008). "Teaching to the rating: School accountability and the distribution of student achievement." In: Journal of Public Economics 92.5, pp. 1394–1415.

— (2014). "Review of New York City's Children First." In: National Education Policy Center–Great Lakes Center.

Reback, Randall, Jonah Rockoff, and Heather L Schwartz (2014). "Under pressure: Job security, resource allocation, and productivity in schools under No Child Left Behind." In: American Economic Journal: Economic Policy 6.3, pp. 207–241.

Reid, Karla Scoon (2003). "Mayor outlines major overhaul of NYC system." In: Education Week 22.19, pp. 1–1.

Rice, Jennifer King (2010). "Principal Effectiveness and Leadership in an Era of Accountability: What Research Says. Brief 8." In: National center for analysis of longitudinal data in education research.

Rivkin, Steven G, Eric A Hanushek, and John F Kain (2005). "Teachers, schools, and academic achievement." In: Econometrica 73.2, pp. 417–458.

Robbins, Liz and Theodoric Meyer (2013). Nine High Schools, One Roof. http://www.nytimes.com/2013/03/17/nyregion/at-the-stevenson-campus-nine-high-schools-one-roof.html?pagewanted=all.

Rockoff, Jonah and Cecilia Speroni (2008). "Reliability, Consistency, and Validity of the NYC DOE Environmental Surveys: A Preliminary Analysis."

Rockoff, Jonah and Lesley J Turner (2010). "Short-run impacts of accountability on school quality." In: American Economic Journal: Economic Policy 2.4, pp. 119–147.

Rogers, David (2009). Mayoral control of the New York City schools. Springer Science & Business Media.

Rose, Stephanie (2012). "Third Grade Reading Policies." In: Education Commission of the States (NJ3).

Rouse, Cecilia Elena et al. (2013). "Feeling the Florida Heat? How Low-Performing Schools Respond to Voucher and Accountability Pressure." In: American Economic Journal: Economic Policy, pp. 251–281.

Schwartz, Amy Ellen, Leanna Stiefel, and Matthew Wiswall (2013). "Do small schools improve performance in large, urban districts? Causal evidence from New York City." In: Journal of Urban Economics 77, pp. 27–40.

Schwerdt, Guido, Martin R. West, and Marcus A. Winters (2015). "The Effects of Test-based Retention on Student Outcomes over Time: Regression Discontinuity Evidence from Florida." National Bureau of Economic Research working papers.

Scott, Janelle and Catherine DiMartino (2009). "Public education under new management: A typology of educational privatization applied to New York City's restructuring." In: Peabody Journal of Education 84.3, pp. 432–452.

Stiefel, Leanna and Amy Ellen Schwartz (2011). "Financing K-12 education in the Bloomberg years, 2002-2008." In: Education reform in New York City: Ambitious change in the nation's most complex school system, pp. 55–83.

Subramanian, Sabrita (2013). "Is it getting fairer? Examining five years of school allocations under fair student funding." In: New York: New York City Independent Budget Office. Retrieved June 17, p. 2013.

The New York City Charter School Center (2012). The State of the NYC Charter School Sector. Tech. rep.

Todd, Petra E and Kenneth I Wolpin (2003). "On the specification and estimation of the production function for cognitive achievement." In: The Economic Journal 113.485.

— (2012). "Estimating a Coordination Game Within the Classroom." In: Manuscript, Univ. Pennsylvania.

Topel, Robert (1991). "Specific Capital, Mobility, and Wages: Wages Rise with Job Seniority." In: Journal of Political Economy, pp. 145–176.

Traver, Amy (2006). "Institutions and organizational change: Reforming New York City's public school system." In: Journal of Education Policy 21.5, pp. 497–514.

Turnbull, Brenda J et al. (2013). Six districts begin the principal pipeline initiative. Tech. rep. The Wallace Foundation.

UFT (2015). New York City's Performance on NAEP 2015. Tech. rep. United Federation of Teachers.

Unterman, Rebecca (2014). "Headed to College the Effects of New York City's Small High Schools of Choice on Postsecondary Enrollment. Policy Brief." In: MDRC.

West, Martin R (2016). "Should non-cognitive skills be included in school accountability systems? Preliminary evidence from California's CORE districts." In: Evidence Speaks Reports 1.13.

West, Martin R et al. (2016). "Promise and paradox: Measuring students' non-cognitive skills and the impact of schooling." In: Educational Evaluation and Policy Analysis 38.1, pp. 148–170.

Wu, Wei, Stephen G. West, and Jan N. Hughes (2010). "Effect of Grade Retention in First Grade on Psychosocial Outcomes." In: Journal of Educational Psychology 102.1, pp. 135–152.

Appendix

Conceptual Framework

I apply a simple framework to describe the mechanism of a student's response to additional teacher effort into testing in order to understand the effects found in the empirical analysis. The framework focuses on a student's maximization problem and abstracts away from the joint determination of teacher input and student effort, which is discussed by Todd and Wolpin (2012). The framework shows that the interactive effects arise from two sources: the change in the pattern of the student's behavioral response due to greater student incentives, and additional marginal returns to student effort due to teacher effort.

Setup

Consider a student in a classroom with a teacher, where their joint effort affects the student's test scores. The education production involves two aspects: the technology of producing test scores and the costs associated with student effort, both of which are subject to the teacher effort. Specifically, the student's maximization problem is defined as:

$$\max_{s}[\alpha A - C(s,t)] \tag{1}$$

where A is the student's test score and is defined as $A = F(s,t)$, in which $F(s,t)$ represents the technology of producing test scores, and the first-order partial differentials, $F_s(s,t)$ and $F_t(s,t)$, are assumed to be both positive, indicating positive returns to student effort and teacher effort in terms of test scores; s and t are student effort and teacher effort, respectively; $\alpha > 0$ measures the student's preference for test scores; and $C(s,t)$ indicates the student's costs of exerting effort and is assumed to be positive. $F_{ss}(s,t) < 0$ is assumed to capture diminishing marginal returns to student effort; $C_{ss}(s,t) = 0$ is assumed for simplicity.

Teacher effort is determined exogenously and assumed to be equal to the strength of teacher incentives. Mathematically, $t = \beta$, where β measures teacher incentives. When neither policies is in effect, the baseline parameters are denoted as α_0 and β_0. Given the assumptions, the solution to this problem is equivalent to solving the first-order condition, $\alpha F_s(s,\beta) - C_s(s,\beta) = 0$.

The Retention Policy

When the retention policy is in place, student incentives increase from α_0 to α_1, and A can be shown to increase with α. The change in a student's test score is described by $\Delta A(\Delta\alpha, \beta_0) = F(s(\alpha_1, \beta_0), \beta_0) - F(s(\alpha_0, \beta_0), \beta_0)$, in which $\Delta\alpha = \alpha_1 - \alpha_0$. A first-order Taylor approximation shows that the change in the test score is roughly:

$$\Delta A(\Delta\alpha, \beta_0) \approx F_s(s(\alpha_0, \beta_0), \beta_0)\frac{ds}{d\alpha}\Delta\alpha \tag{2}$$

When $\Delta\alpha > 0$, the sign of ΔA is equivalent to the sign of $F_s(s(\alpha_0, \beta_0), \beta_0)ds/d\alpha$. $F_s(s(\alpha_0, \beta_0), \beta_0)$ is assumed to be positive, and $ds/d\alpha$ can be shown as $-f_s/(\alpha f_{ss})$, which is also positive based on the assumptions. As a result, when student incentives increase, the student exerts more effort, and his or her test score increases.

The Accountability Scheme

When a teacher directs additional effort to the student after the implementation of the accountability scheme, the change in the test score is more complicated. When β_0 increases to β_1, $\Delta A(\alpha_0, \Delta\beta) = F(s(\alpha_0, \beta_1), \beta_1) - F(s(\alpha_0, \beta_0), \beta_0)$, where $\Delta\beta = \beta_1 - \beta_0$. This change is approximated as:

$$\Delta A(\alpha_0, \Delta\beta) \approx \underbrace{F_t(s(\alpha_0, \beta_0), \beta_0)\frac{dt}{d\beta}\Delta\beta}_{\text{Direct Effect}} + \underbrace{F_s(s(\alpha_0, \beta_0), \beta_0)\frac{ds}{d\beta}\Delta\beta}_{\text{Behavioral Response}} \tag{3}$$

The direct effect is clearly positive, since the assumptions state that $F_t(s(\alpha_0, \beta_0), \beta_0) > 0$, and $dt/d\beta = 1 > 0$. The sign of the behavioral response is determined by $ds/d\beta$, which is not necessarily positive. $ds/d\beta$ can be shown as:

$$\frac{ds}{d\beta} = -\frac{\alpha F_{st} - C_{st}}{\alpha F_{ss}} \tag{4}$$

In this equation, since F_{ss} is assumed to be negative, the sign is determined by the relative magnitude of αF_{st} and C_{st}. These two factors may represent the two counteracting effects described in the section on a possible mechanism. $F_{st}(s, t)$ is assumed to be positive and captures the first effect — that is, teacher effort increases the return to student effort in terms of test scores. $C_{st}(s, t)$ is also assumed to be positive and represents the effect whereby additional teacher effort on increase student laziness and resistance.

When student incentives are low, C_{st} dominates, and the student exhibits a negative behavioral response. A small α results in a relatively larger C_{st}, and therefore $\alpha F_{st} - C_{st} < 0$, leading to $ds/d\beta < 0$. Therefore, the student reduces the amount of effort. If the reduction of student effort is large enough, the change in his or her test score can be negative, as is found in the empirical analysis of the accountability scheme.

The Interactive Effects

The interactive effects identified in the empirical analysis are equivalent to subtracting the individual effect of each policy from the combined effects of the two policies. In other words, the interactive effects are defined as $\Delta A(\Delta\alpha, \Delta\beta) - \Delta A(\alpha_0, \Delta\beta) - \Delta A(\Delta\alpha, \beta_0)$. The combined effects, $\Delta A(\Delta\alpha, \Delta\beta)$, are approximately:

$$\Delta A(\Delta\alpha, \Delta\beta) \approx \underbrace{F_s(s(\alpha_0, \beta_0), \beta_0)\frac{ds}{d\alpha}\Delta\alpha}_{\text{Direct Effect from } \Delta\alpha} + \underbrace{F_t(s(\alpha_0, \beta_0), \beta_0)\frac{dt}{d\beta}\Delta\beta}_{\text{Direct Effect from } \Delta\beta} + \underbrace{F_s(s(\alpha_0, \beta_0), \beta_0)\frac{ds}{d\beta}\Delta\beta}_{\text{Behavioral Response to } \Delta\beta} \quad (5)$$

In this equation, the direct effect from $\Delta\beta$ is equal to the direct effect of the accountability scheme, $\Delta A(\alpha_0, \Delta\beta)$.

As a result, the interactive effects arise from the change in $ds/d\alpha$ with respect to β_1 and the change in $ds/d\beta$ with respect to α_1. The change in $ds/d\alpha$ can be shown as:

$$\frac{\partial(ds/d\alpha)}{\partial\beta} = \frac{-\alpha F_{st} f_{ss} + \alpha F_s F_{sst}}{(\alpha F_{ss})^2} \quad (6)$$

Since the assumptions have determined the signs of F_s, F_{st}, and F_{ss}, the sign of F_{sst} needs to be assumed in order to determine the sign of $\frac{\partial(ds/d\alpha)}{\partial\beta}$. Since additional teacher effort should not make the marginal returns to student effort diminish faster, F_{sst} is expected to be weakly positive. All these assumptions indicate that $\frac{\partial(ds/d\alpha)}{\partial\beta} > 0$, which means that the growth in student effort with respect to α increases with β.

The change in $ds/d\beta$ is equal to:

$$\frac{\partial(ds/d\beta)}{\partial\alpha} = \frac{\partial s}{\partial\alpha} \times \frac{-F_{ss}(F_{sts} - C_{sst}/\alpha + C_{st}/\alpha^2) - F_{sss}(F_{st} - C_{st}/\alpha)}{(F_{ss})^2} \quad (7)$$

The assumption $C_{ss} = 0$ and other assumptions on the sign of other factors indicate that $-F_{ss}(F_{sts} - C_{sst}/\alpha + C_{st}/\alpha^2) = -F_{ss}(F_{sts} + C_{st}/\alpha^2) > 0$. The sign of F_{sss} is more difficult to determine.[6] If F_{sss} is negative and large, F_{ss} decreases quickly with respect to s, and student incentives are expected to have a small overall impact, which is not supported by the effects of the retention policy found in the empirical analysis. The negative effects of the accountability scheme suggest $F_{st} - C_{st}/\alpha < 0$; $\frac{\partial s}{\partial\alpha}$ has been shown to be positive. As a result, $\frac{\partial(ds/d\beta)}{\partial\alpha} > 0$.

In short, the interactive effects arise from two sources: the increase in student effort due to additional teacher effort and the reduction in the student's negative behavioral response due to greater student incentives.

[6] If $F(s,t)$ follows a Cobb-Douglas form, $F_{sss} > 0$.

Probability of Retention

Figure 1.2 suggests that students' probability of retention may have changed after 2007, the year when the accountability scheme was implemented. Since retained students do not count toward the student progress scores in the accountability scheme, this change may be due to the accountability scheme.

It is challenging to causally estimate the interactive effect on the probability of retention, since there lacks a control group within each grade for students subject to the retention policy. Many factors may have resulted in a change in retention patterns. For example, teachers and principals may have promoted students who failed the test but could be counted favorably in the accountability scheme if promoted. Changes in retention patterns could also be due to changes in students' academic portfolios or behaviors.

Appendix Figure A20 examines change in retention patterns after the accountability scheme was implemented. Panels A and B show that, during the two years between the retention policy and the accountability scheme, the grade subject to the retention policy imposed a greater probability of retention, especially for students who failed the test.[7] Panels C and D show the probability of retention after the accountability scheme was implemented. Retention risks conditional on failing math tests increased for grades not subject to the retention policy, which may due to more selective retention decisions on these grades. Retention risks conditional on failing English tests decreased for the grade subject to the policy. Examining the summer school outcomes suggests that this decrease may be due to higher August English test scores.

[7]There is a jump two points to the right of the black line. The jump is due to adoption of the state tests and redefinition of the cutoff in 2006.

Survey Questions in Each Category

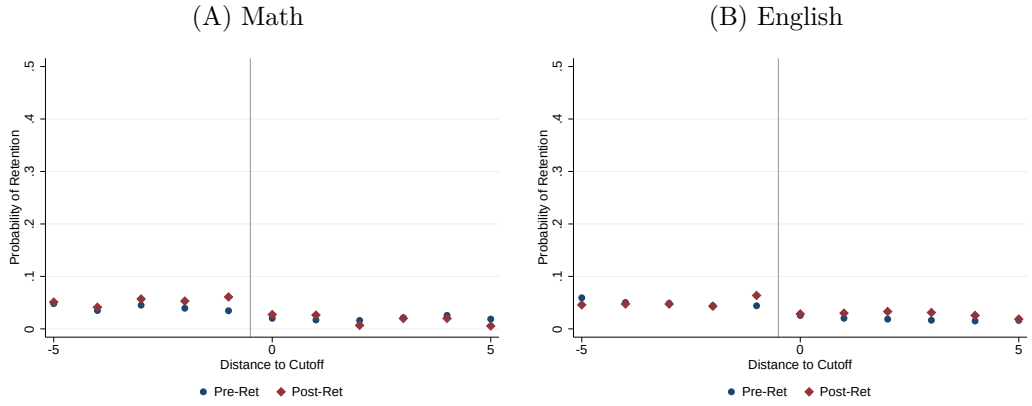
- Parents' overall satisfaction includes questions 2, 5, 9 and 13.
- Parents' sense of overall safety includes question 11.
- Students' overall satisfaction includes questions 2a, 3e, 3f, 3g, 6a, 6c-6g, 14a
- Students' sense of personal safety includes 13a, 13e, 13f, 13g
- Students' perception of environment includes 3d, 6b, 12a, 12b, 12c, 13b, 13c, 13d, 14b-14f.

Continuity of Personal Characteristics

In order to validate our Regression Discontinuity Design, we test continuity of characteristics other than percent of women, percent of reduce/free-price lunch recipients, and density of observations (shown in Figures 2.3 and 2.4). Appendix Figure A27 shows that the percentage of each ethnicity is continuous across the cutoff. Appendix Figure A28 presents the percentage of students who stay at NYC public schools next year at each index and there is no discontinuity at the cutoff. Appendix Figure A29 shows the percentage of students and parents who responded to surveys by index score and both rates are smooth through the cutoff. We also test continuity by regression analysis. These results are in Appendix Table A18. This supports the notion that our results are not driven by any discontinuity of other student characteristics across the cutoff.

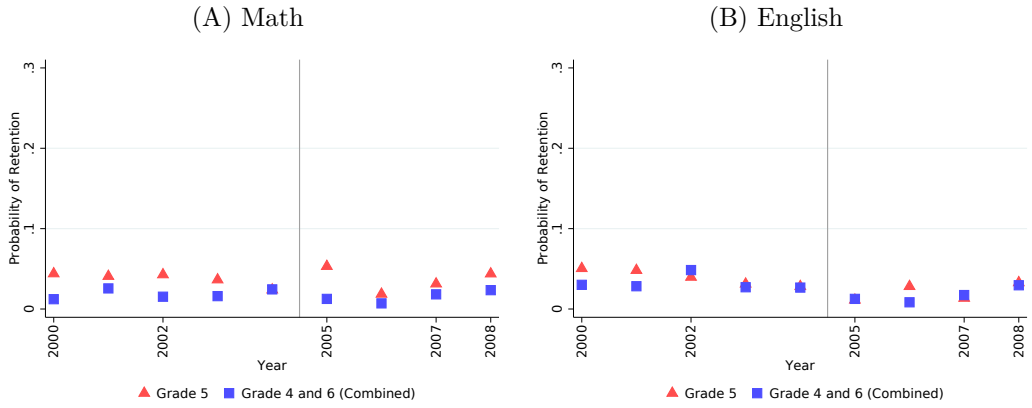
Appendix Figures

Figure A1: Probability of Retention for Exempt Students



Notes: Both panels are restricted to the years prior to the accountability scheme (prior to 2007) and to students exempt from the retention policy. Each point represents the probability of being retained at each value of the index. The index is defined as the difference between a student's spring test score and the cutoff in each subject. Students on the left of the gray vertical line failed the test. Pre-Ret combines the grades/years not subject to the retention policy, and Post-Ret combines the grades/years subject to the retention policy.

Figure A2: The Probability of Retention for Exempt Students: Time Series



Notes: Both panels focus on students exempt from the retention policy. Each point restricts the observations to the students in Figure 1.1 and represents the probability of retention conditional on failing the test in each subject-grade-year cell — that is, $Prob(Retention|Fail) - Prob(Retention|Pass)$. Blue triangles present the probability of retention for 5th grade;Gray squares present the probability of retention for 4th and 6th grades. To the right of the black line are years after the retention policy was implemented.

Figure A3: The Accountability Grade Rubric

Final Calculation of Progress Report Grade

Category Scores are calculated by weighting the values within each category of the Proximity to Peer Horizon (x3) and Proximity to Peer Horizon (x1) measures for School Environment, Student Performance, and Student Progress. As the weighting indicates, Proximity to Peer Horizon counts three times as much as Proximity to City Horizon. These weighted values within each category are then averaged to create scores for School Environment, Student Performance, and Student Progress. The school's overall score is a weighted average of School Environment (15%), Student Performance (25%), and Student Progress (60%) plus any additional credit earned by the school.

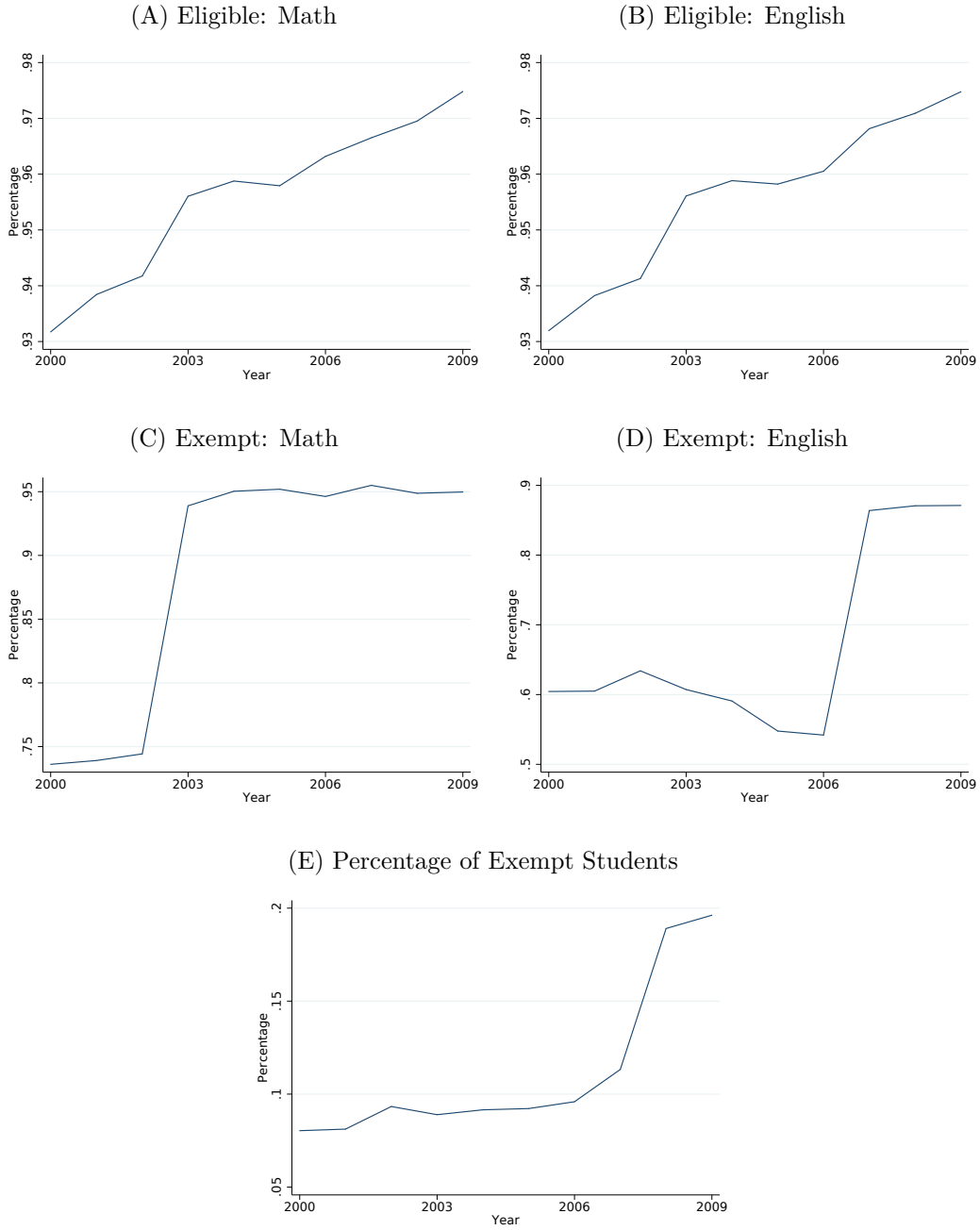
The maximum point values for each measure are indicated in the table below:

Category Measure	*Total points*	*Peer Horizon point values (75% of total)*	*City Horizon point values (25% of total)*
School Environment	**15.0**	**11.25**	**3.75**
Academic Expectations	2.5	1.875	0.625
Communication	2.5	1.875	0.625
Engagement	2.5	1.875	0.625
Safety and Respect	2.5	1.875	0.625
Attendance	5.0	3.75	1.25
Student Performance	**25.0**	**18.75**	**6.25**
ELA – Percentage of Students at Proficiency	6.25	4.6875	1.5625
ELA – Median Student Proficiency	6.25	4.6875	1.5625
Math – Percentage of Students at Proficiency	6.25	4.6875	1.5625
Math – Median Student Proficiency	6.25	4.6875	1.5625

Category Measure	*Total points*	*Peer Horizon point values (75% of total)*	*City Horizon point values (25% of total)*
Student Progress	**60.0**	**45.0**	**15.0**
ELA – Percentage of Students Making at Least 1 Year of Progress	7.5	5.625	1.875
ELA – Percentage of Students in School's Lowest Third Making at Least 1 Year of Progress	7.5	5.625	1.875
ELA – Average Change in Student Proficiency for Level 1 and Level 2 students	15.0 (school-specific based on the % of students reflected in each measure)	11.25 (school-specific)	3.75 (school-specific)
ELA – Average Change in Student Proficiency for Level 3 and Level 4 students			
Math – Percentage of Students Making at Least 1 Year of Progress	7.5	5.625	1.875
Math – Percentage of Students in School's Lowest Third Making at Least 1 Year of Progress	7.5	5.625	1.875
Math – Average Change in Student Proficiency for Level 1 and Level 2 students	15.0 (school-specific based on the % of students reflected in each measure)	11.25 (school-specific)	3.75 (school-specific)
Math – Average Change in Student Proficiency for Level 3 and Level 4 students			

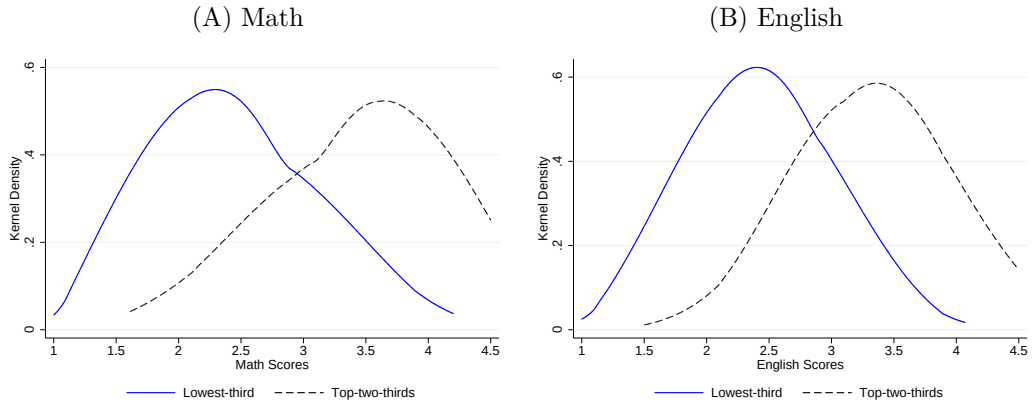
Notes: See the website of the New York City Department of Education for full documentation: http://schools.nyc.gov/Accountability/tools/report/ProgressReport_2007-2013.htm.

Figure A4: Selection



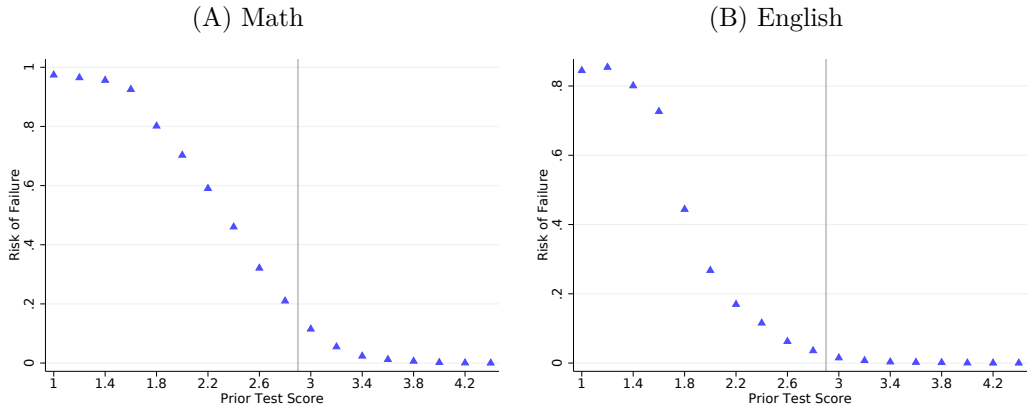
Notes: Panels A, B, C, and D plot the percentage of students who took the exam separately by subject and exemption status for the retention policy. Panel E plots the percentage of exempt students conditional on having both current and prior test scores.

Figure A5: Distribution of Test Scores for Lowest-Third and Top-Two-Thirds Students



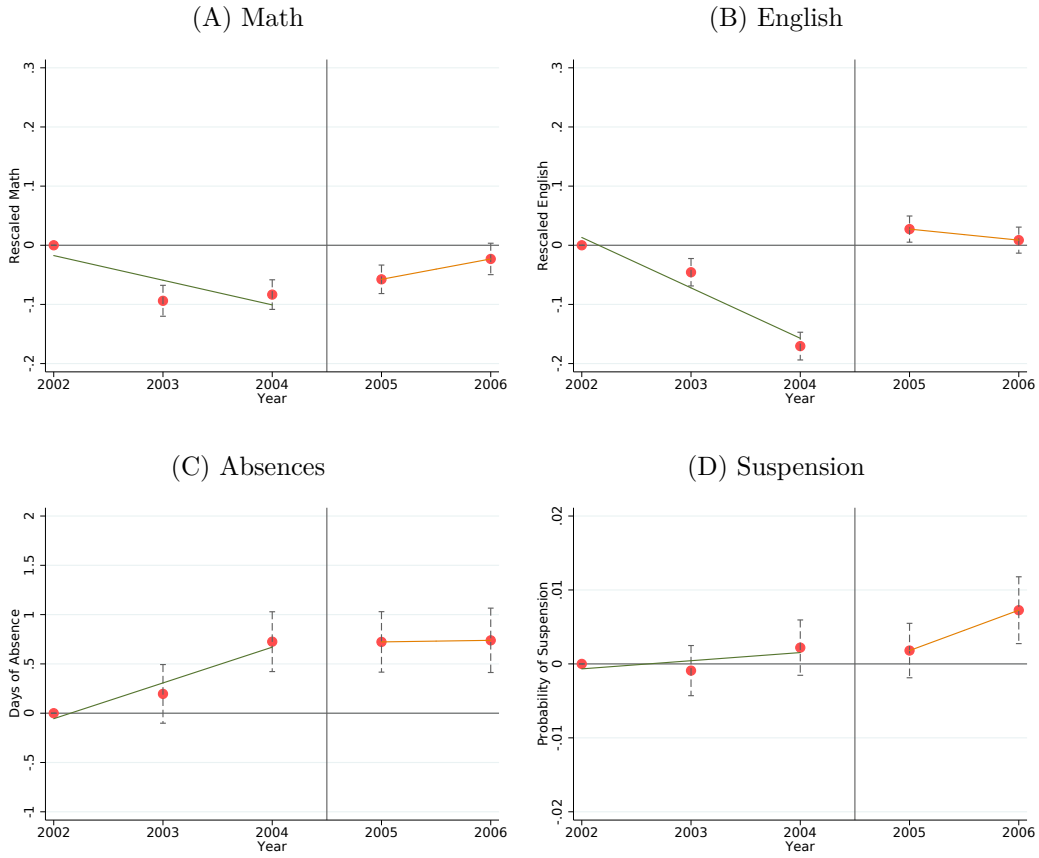
Notes: Each panel plots the kernel density of test scores separately for lowest-third and top-two-thirds students in each subject.

Figure A6: Empirical Risk of Failure



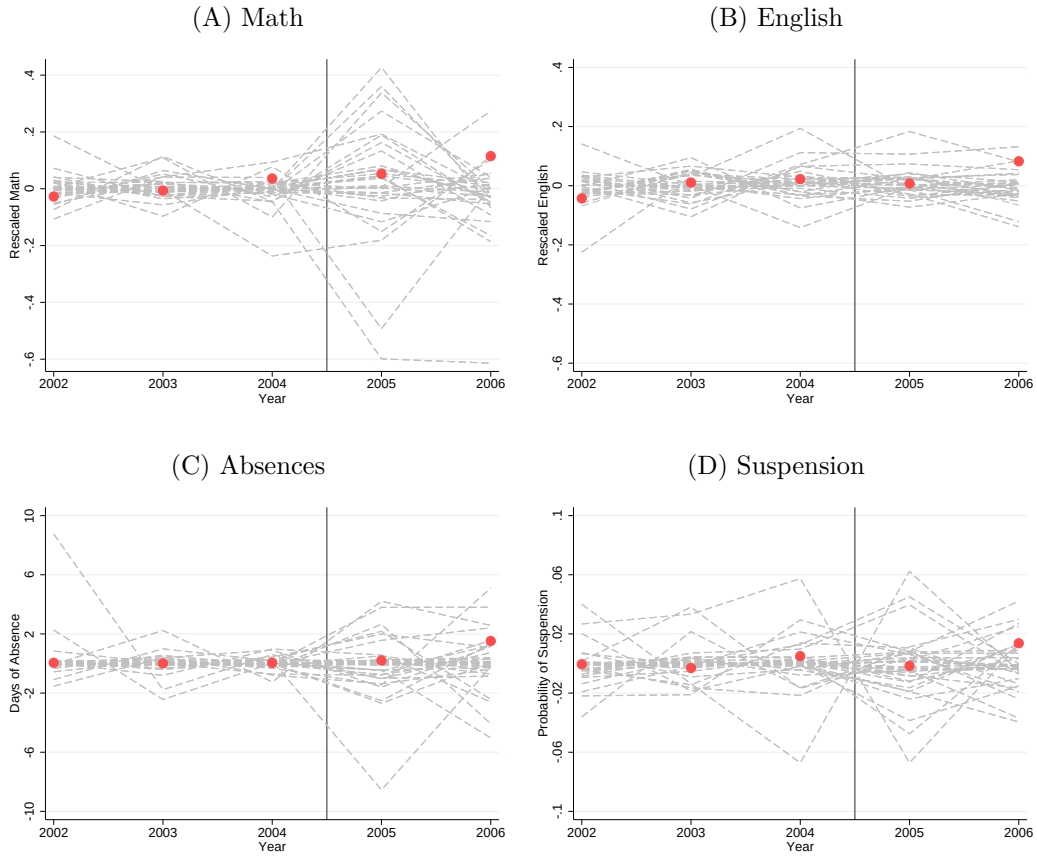
Notes: Both panels are restricted to years when neither policies was implemented (prior to 2004) and divide prior test scores into bins of 0.2 points each. Each point represents the average probability of failing the test at each bin of prior test scores.

Figure A7: Effects of the Retention Policy (DID)



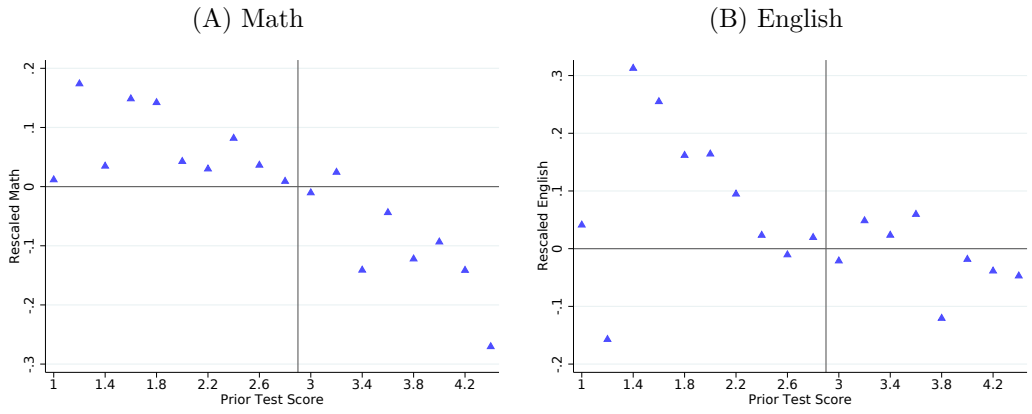
Notes: All panels are based on data from 2002 to 2006 and use the grade subject to the retention policy. This figure plots coefficients β_2 for each year from an event-study version of Equation 1.1. The dependent variables in Panels A and B are test scores in math and English. The dependent variables in Panels C and D are the number of days absent from school and an indicator of ever being suspended from school. To the right of the black line are years after the retention policy was implemented.

Figure A8: Effects of the Retention Policy (Synthetic Control): Placebo



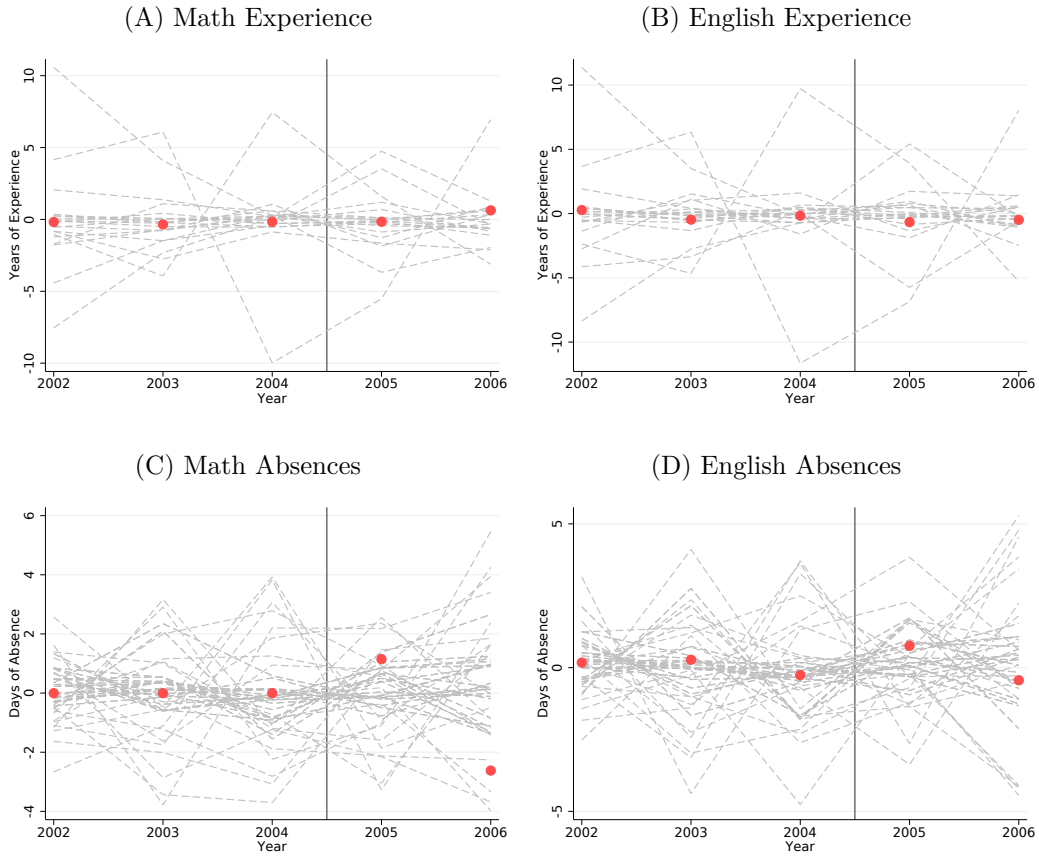
Notes: All panels are based on data from 2002 to 2006 and use grades not subject to the retention policy. The red line plots the difference between the treatment group and the synthetic control group in each year; the gray lines plot the difference between each member in the donor pool and its synthetic control group in each year. The dependent variables in Panels A and B are test scores in math and English; the dependent variables in Panels C and D are the number of days absent from school and an indicator of ever being suspended from school. To the right of the black line are years after the retention policy was implemented.

Figure A9: Distributional Effects of the Retention Policy



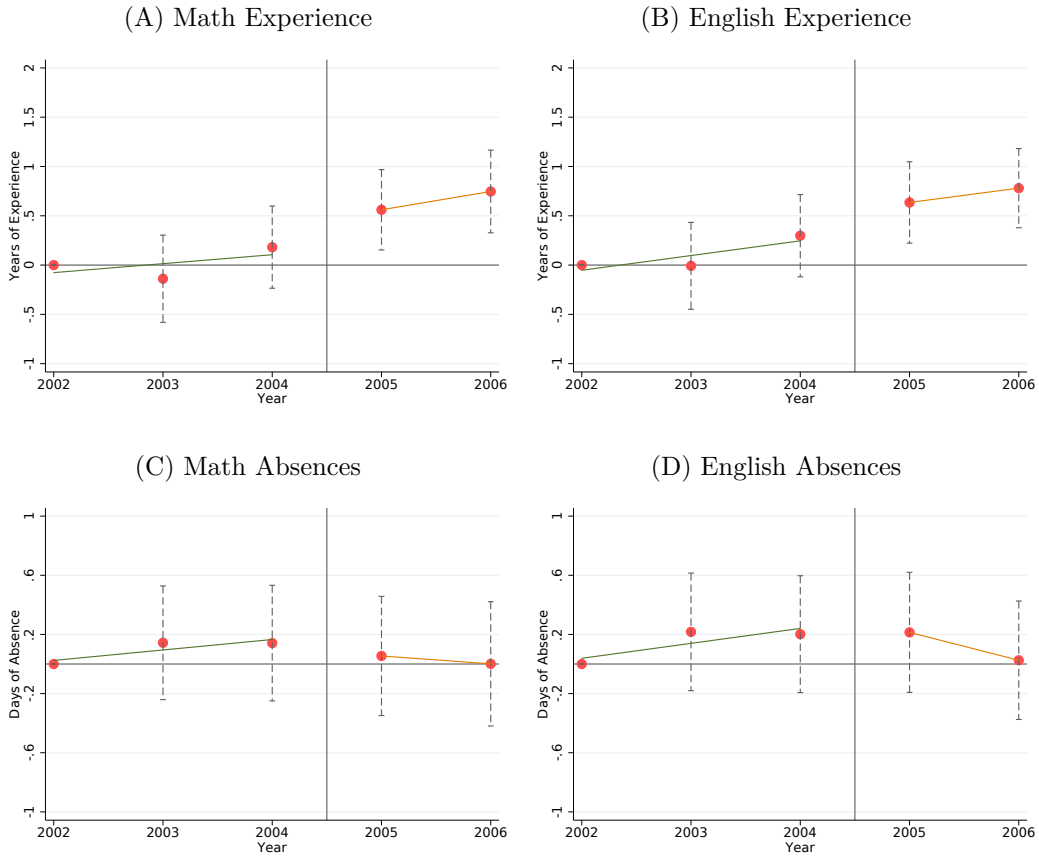
Notes: Both panels are based on students in the grade subject to the retention policy between 2003 and 2005 and divide prior test scores into bins of 0.2 points each. Each point represents a difference-in-difference estimate of the retention policy for each bin of students, using exempt students as a control group. Above the horizontal line stands for improvements in the outcome. To the right of the black line are students who faced little risk of failure.

Figure A10: Effects of the Retention Policy on Teachers (Synthetic Control)



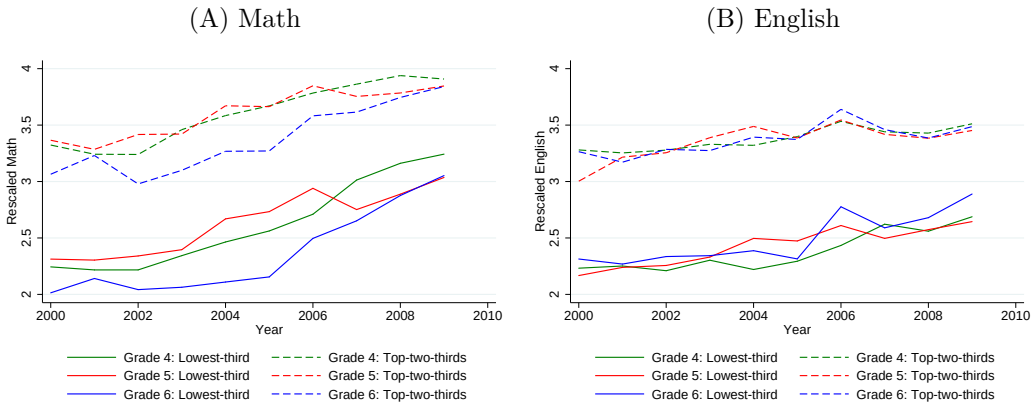
Notes: All panels are based on data from 2002 to 2006 and use the grade subject to the retention policy. The red line plots the difference between the treatment group and the synthetic control group in each year; the gray lines plot the difference between each member in the donor pool and its synthetic control group in each year. The dependent variables in Panels A and B are teachers' years of experience in math and English; the dependent variables in Panels C and D are teachers' days of absence from school in each subject. To the right of the black line are years after the retention policy was implemented.

Figure A11: Effects of the Retention Policy on Teachers (DID)



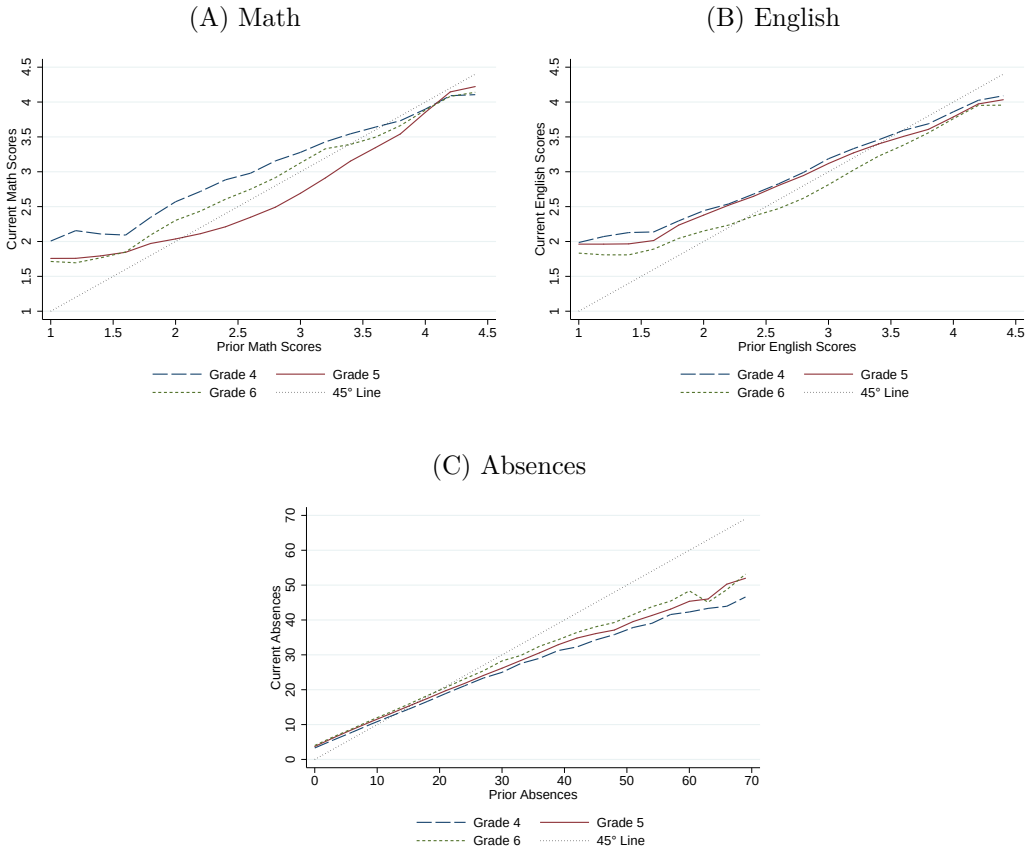
Notes: All panels are based on teacher data from 2002 to 2006 and use the grade subject to the retention policy. This figure plots coefficients β_2 for each year from an event-study version of Equation 1.1. The dependent variables in Panels A and B are teachers' years of experience in math and English; the dependent variables in Panels C and D are teachers' days of absence from school in each subject. To the right of the black line are years after the retention policy was implemented.

Figure A12: Trends in Prior Outcomes



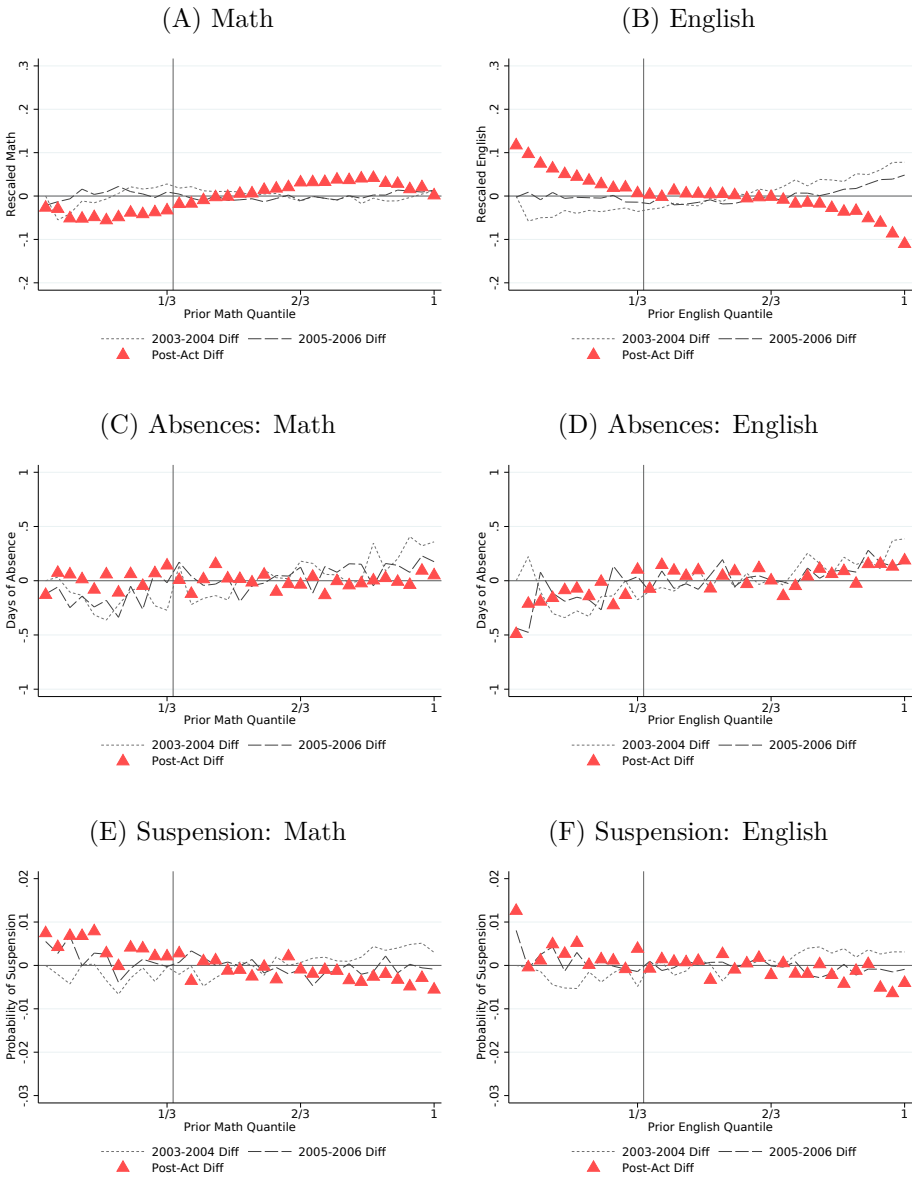
Notes: Both panels plot average prior outcomes in each year separately for lowest-third and top-two-thirds students.

Figure A13: Relationships between Current and Prior Outcomes



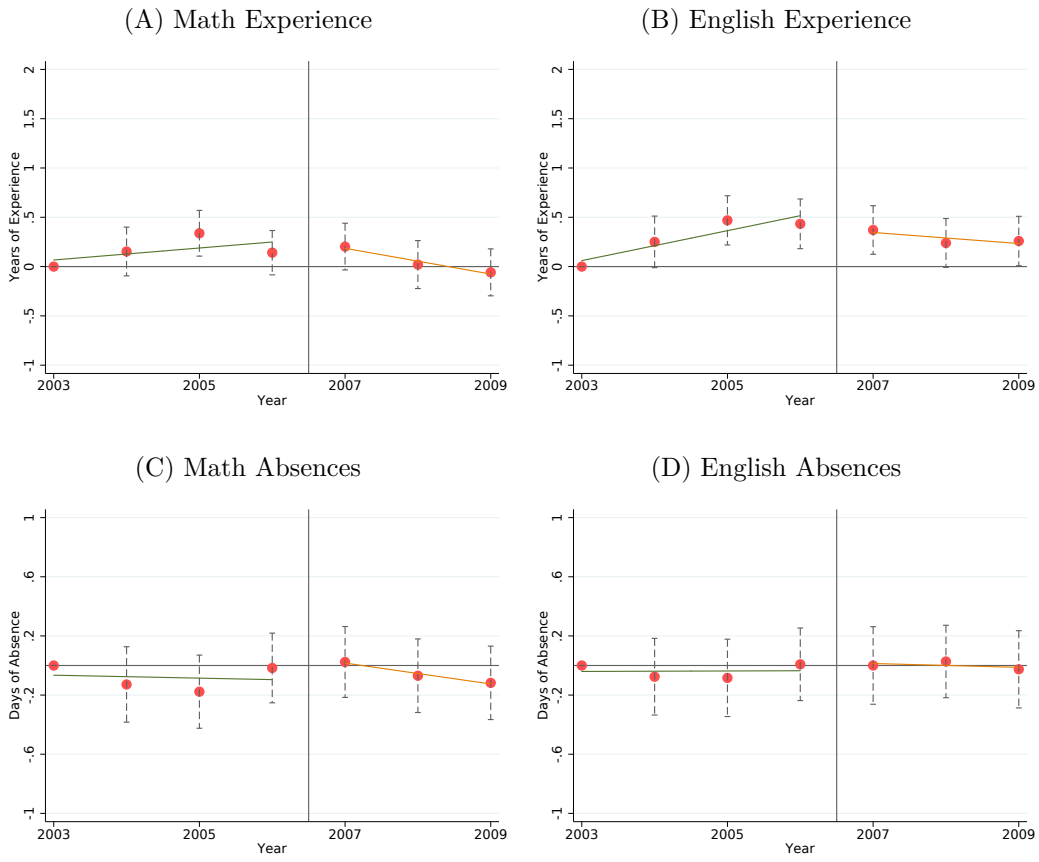
Notes: All panels are restricted to years when neither policies was in effect (prior to 2005).

Figure A14: Distributional Effects of the Accountability Scheme



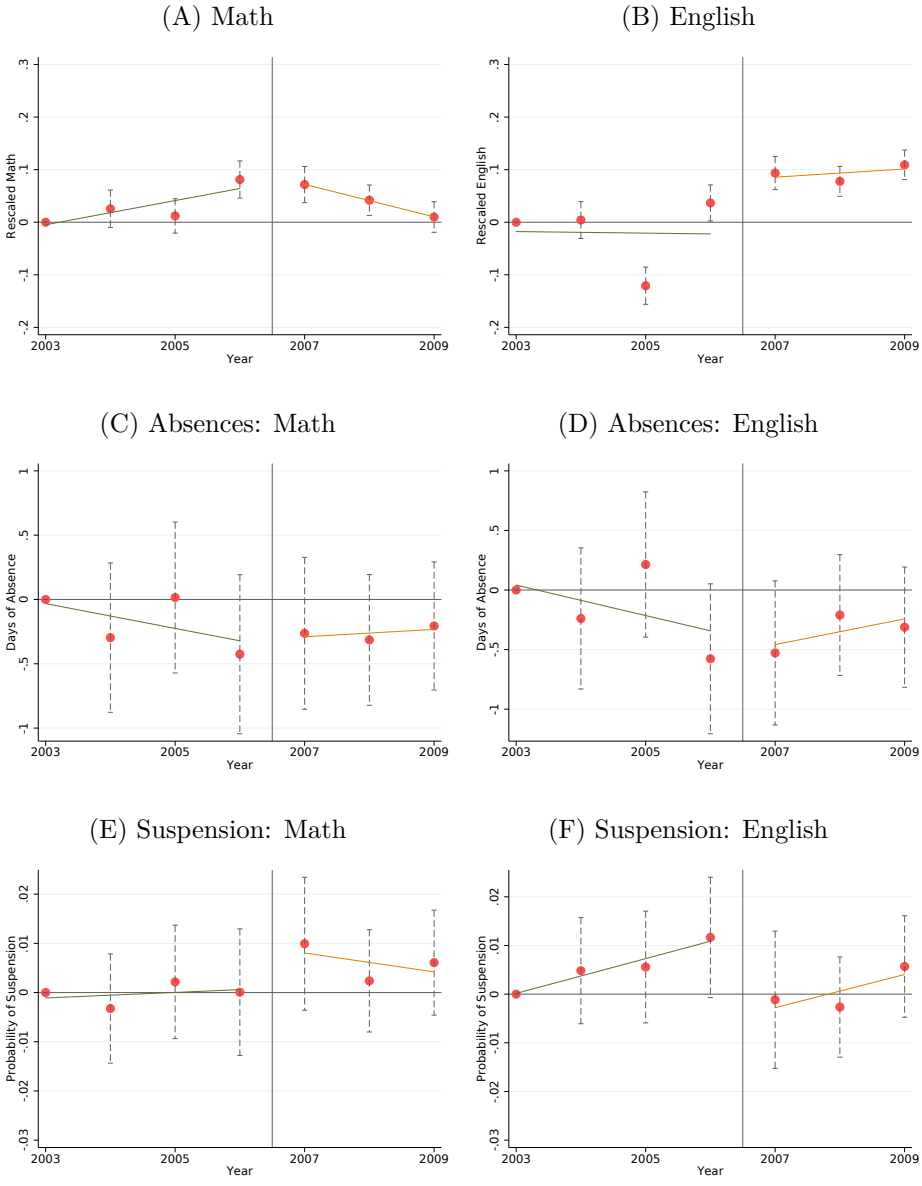
Notes: All panels are based on data from 2003 to 2009 and focus on students subject to the retention policy but in the grades not subject to the retention policy. This figure plots residuals obtained from regressing the outcomes on Equation 1.3. The x-axis in the left column uses prior math ranks; the x-axis in the right column uses prior English ranks. Triangles plot the means of the residuals in the post-accountability era; the lighter dashed line plots the means of the residuals in the years 2003 and 2004, and the darker and longer one plots those in the years 2005 and 2006. The gray vertical line indicates the cutoff for being in the lowest third and the gray horizontal line is at the value of zero. The dependent variables in Panels A and B are test scores in each subject; the dependent variables in Panels C and D are days absent from school; the dependent variables in Panels E and F are probability of suspension.

Figure A15: Effects of the Accountability Scheme on Teachers



Notes: All panels are based on data from 2003 to 2009 and focus on students subject to the retention policy but in the grades not subject to the retention policy. This figure plots coefficients β_2 for each year from an event-study version of Equation 1.3. Each point represents the average of the residuals at each year. The dependent variables in Panels A and B are teachers' years of experience in math and English; the dependent variables in Panels C and D are teachers' days of absence from school in each subject. To the right of the black line are years after the accountability scheme was implemented.

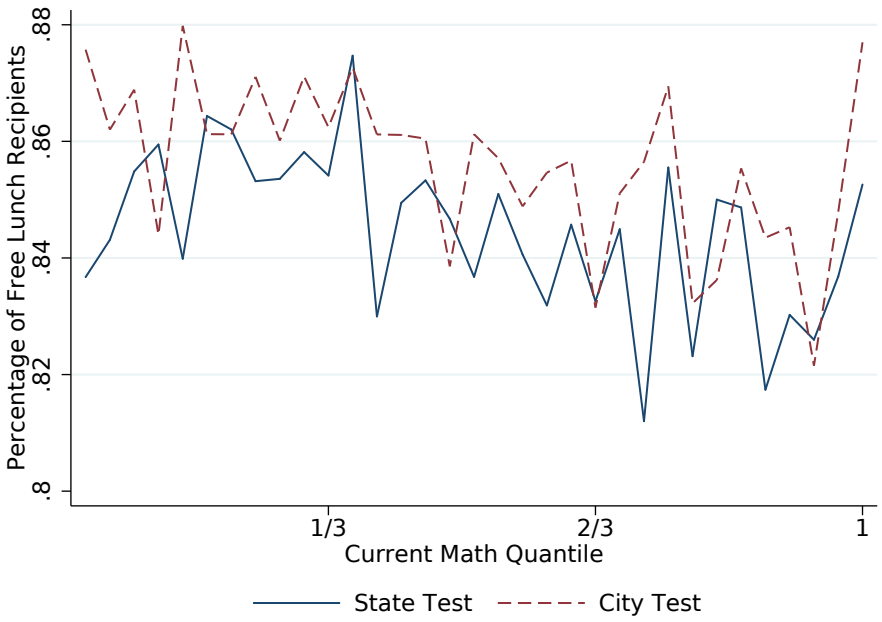
Figure A16: Effects of the Accountability Scheme: Special Ed/ELL



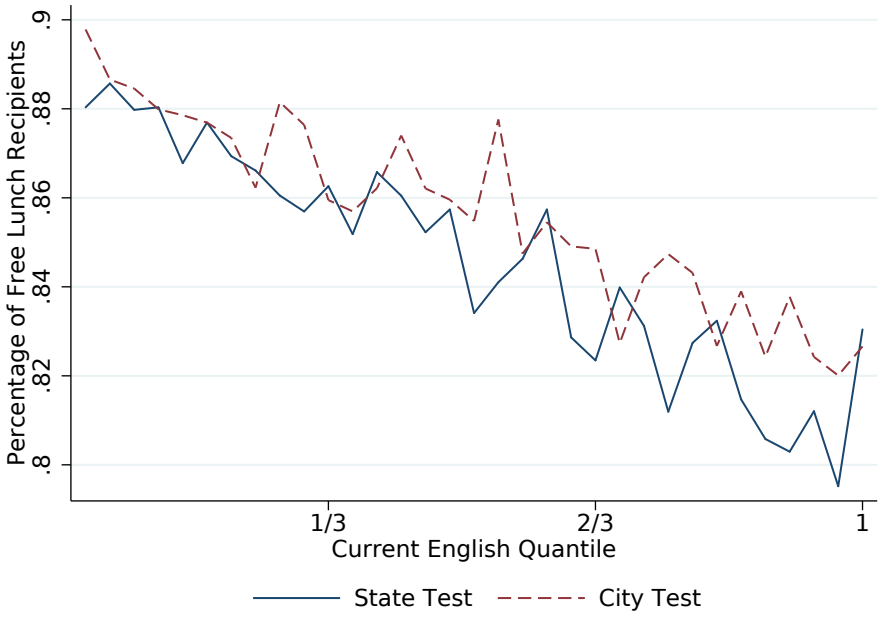
Notes: All panels are based on data from 2003 to 2009 and focus on special education/ELL students in 4th and 6th grades. This figure plots coefficients β_2 for each year from an event-study version of Equation 1.2. The dependent variables in Panels A and B are test scores in each subject; the dependent variables in Panels C and D are days absent from school; the dependent variables in Panels E and F are probability of suspension. To the right of the black line are years after the accountability scheme was implemented.

Figure A17: Distribution of Free Lunch Recipients: City vs. State Tests

(A) Math

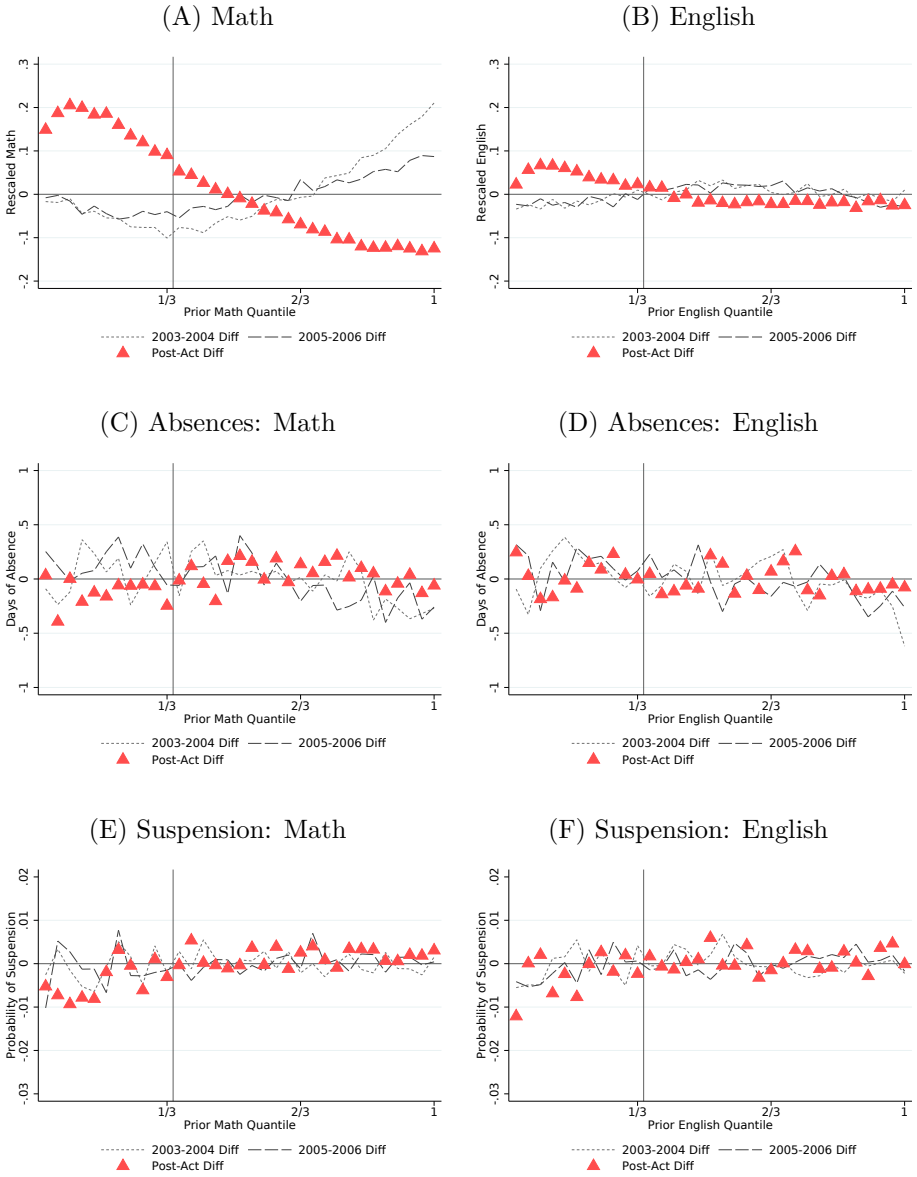


(B) English



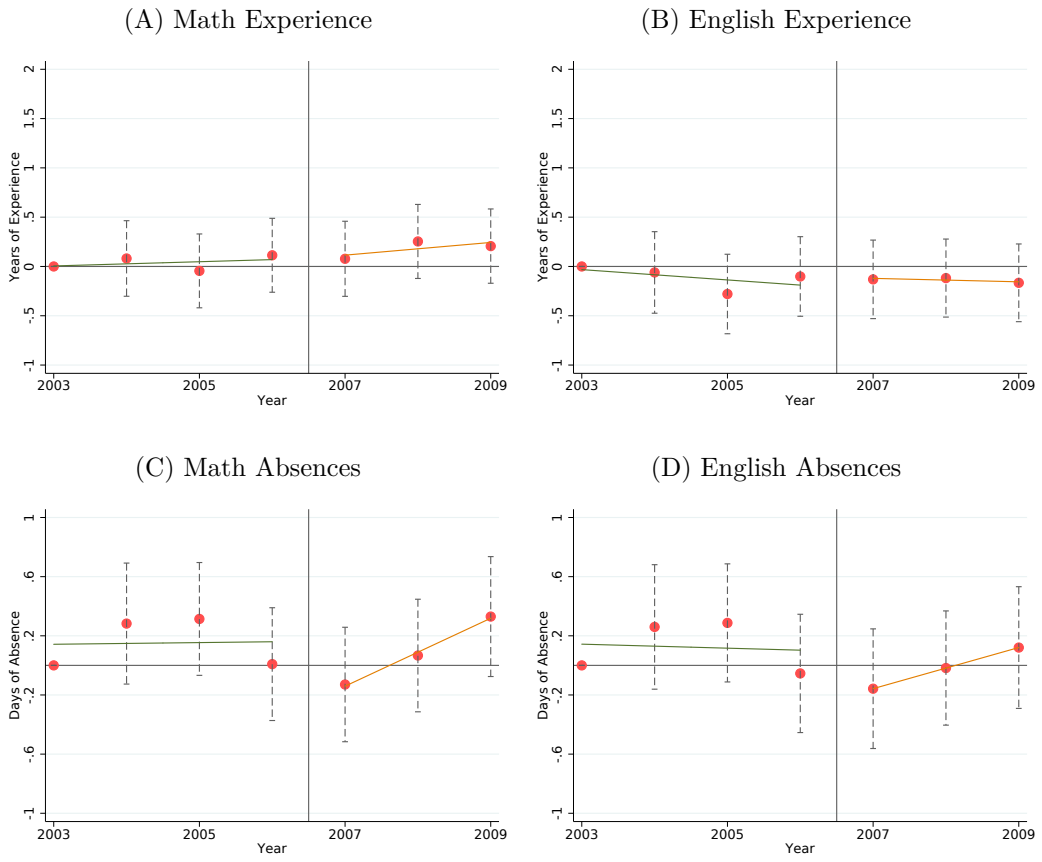
Notes: Both panels plot the percentage of free lunch recipients at each quantile immediately before (2005) and after adopting the state tests (2006) in grades 4 and 6. The solid line stands for the state tests (2006) and the dashed line stands for the city tests (2005).

Figure A18: Distributional Effects of the Policy Interaction



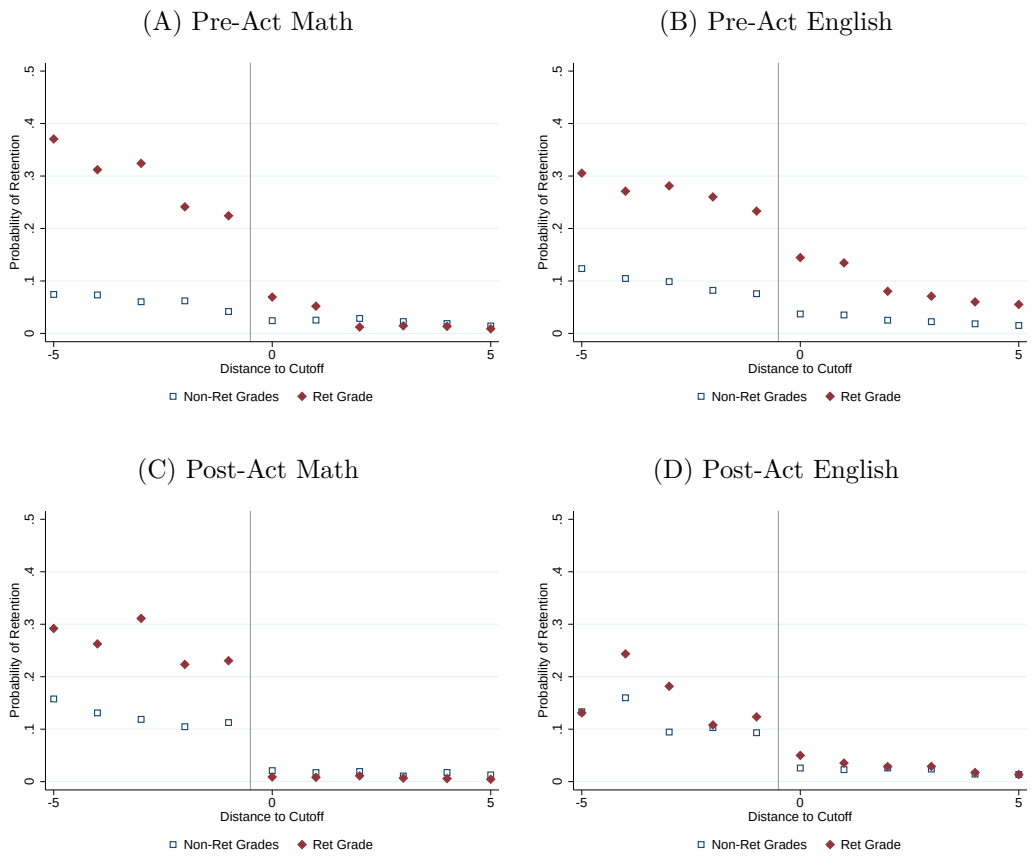
Notes: All panels use data from 2003 to 2009, focus on students subject to the retention policy, and plot the differences in the residuals obtained from regression 1.3 between the grade subject to the retention policy and other grades. The x-axis in the left column uses prior math ranks; the x-axis in the right column uses prior English ranks. The short dashed line plots the years 2003 and 2004, the black and longer dashed line plots the years 2005 and 2006, and triangles plot the post-accountability years. The gray vertical line indicates the cutoff for being in the lowest third and the gray horizontal line is at the value of zero. The dependent variables in Panels A and B are test scores in each subject; the dependent variables in Panels C and D are days absent from school; the dependent variables in Panels E and F are probability of suspension.

Figure A19: Effects of the Policy Interaction on Teachers



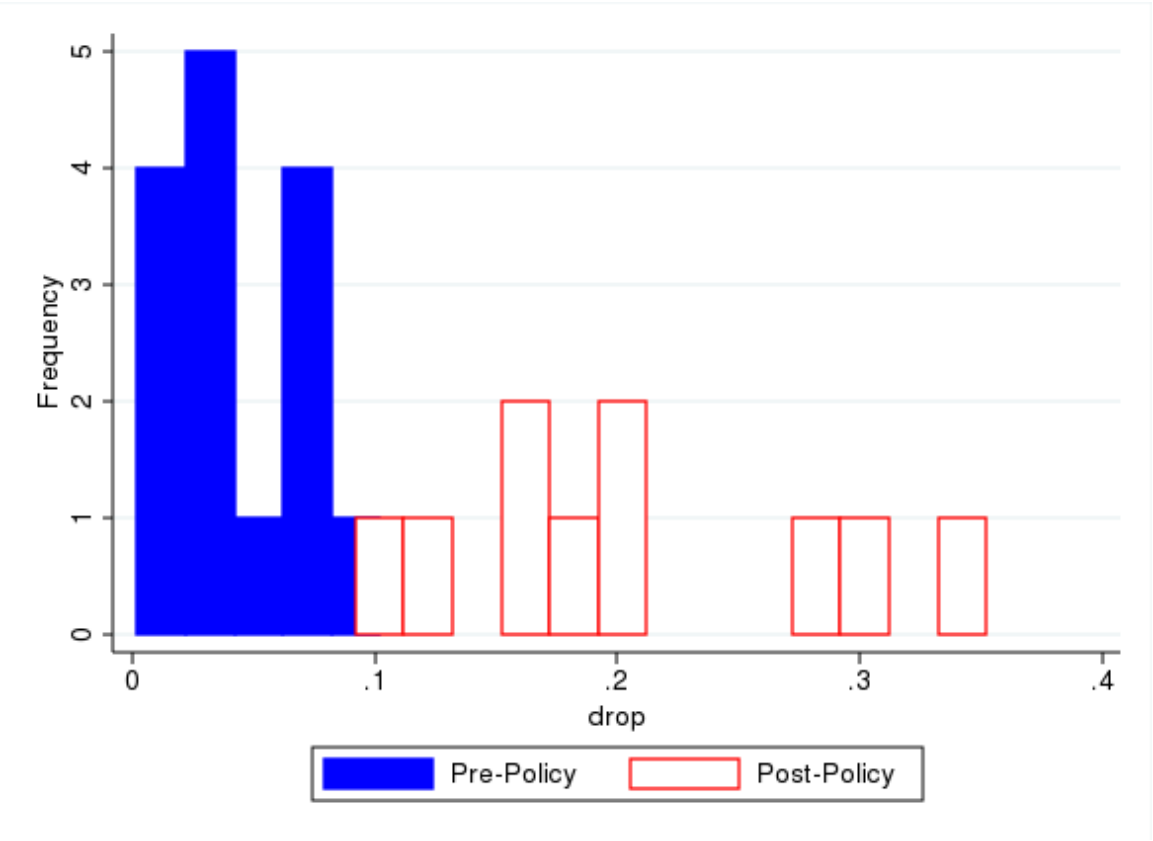
Notes: All panels use data from 2003 to 2009, focus on students subject to the retention policy, and plot a time series of DID estimates that use the residuals generated from regression 1.3 to measure the effect of being a lowest-third student in the grade subject to the retention policy. The left panels focus on lowest-third students in math, and the right panels examine lowest-third students in English. The dependent variables in Panels A and B are teachers' years of experience in math and English; the dependent variables in Panels C and D are teachers' days of absence from school in each subject. To the right of the black line are years after the accountability scheme was implemented.

Figure A20: Changes in the Probability of Retention



Notes: All panels restrict the data as described in the Data section. Panels A and B are restricted to the two years prior to the accountability scheme (2005 and 2006); Panels C and D are restricted to the years after the accountability scheme was implemented (after 2007). Each point represents the probability of being retained at each value of the index. The index is defined as the difference between a student's test score and the cutoff in each subject. Students to the left of the gray vertical line failed the test. "Non-Ret Grades" combines grades not subject to the retention policy, and "Ret Grade" stands for the grade subject to the retention policy.

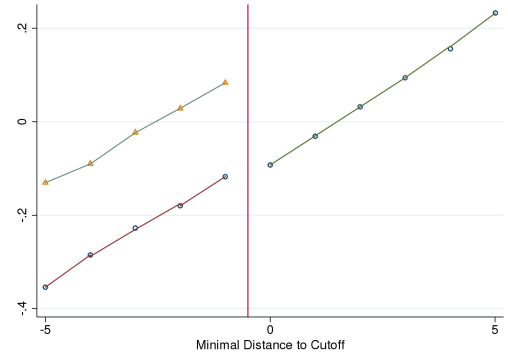
Figure A21: Frequency of Pre- and Post-Policy Retention Rates



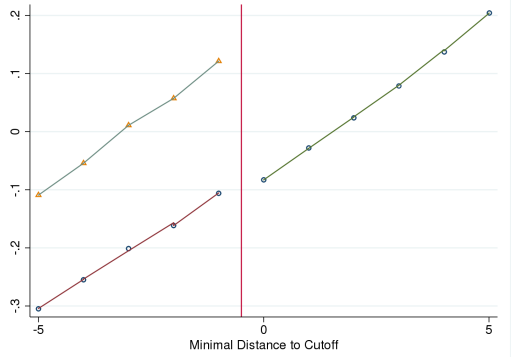
Note: The retention rates are the discontinuity in the probability of being retained at the cutoff. This figure plots the histogram of change in retention rate at the cutoff at each grade-year cell.

Figure A22: Effects on Future Test Scores and Special Education

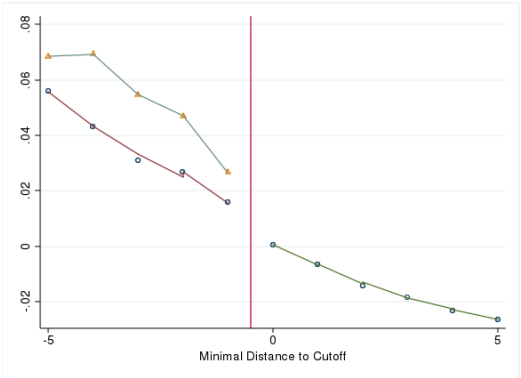
(A) Mathematics



(B) ELA



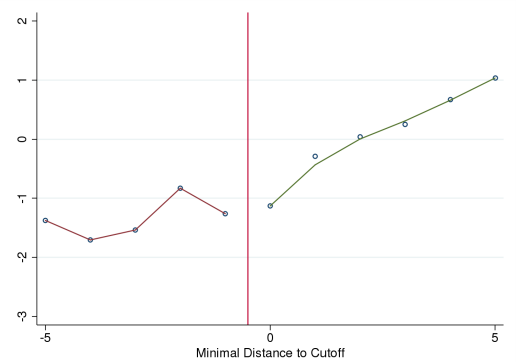
(C) Special Education



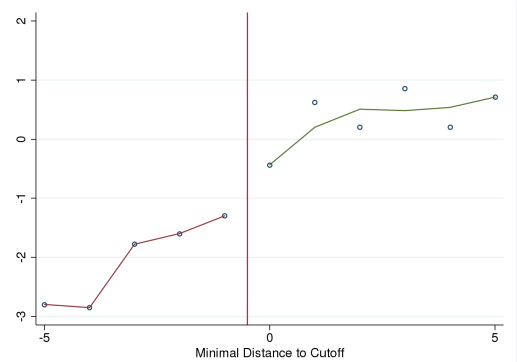
Note: These figures plot residuals from regressions of future test scores and probability of receiving special education on fixed effects for test grade by test year. ELA stands for the English Language Arts exam. Average residuals by index score are plotted separately by students who were retained (yellow triangle), failed at least one of the tests but were not retained (circle on the left side of cutoff), and passed both tests (circle on the right side of cutoff). ELA stands for the English Language Arts exam.

Figure A23: Placebo Effects on Survey Responses

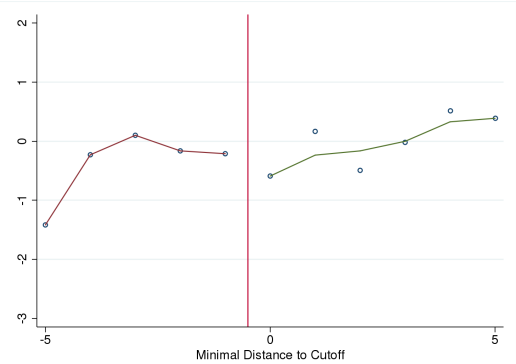
(A) Parental Satisfaction



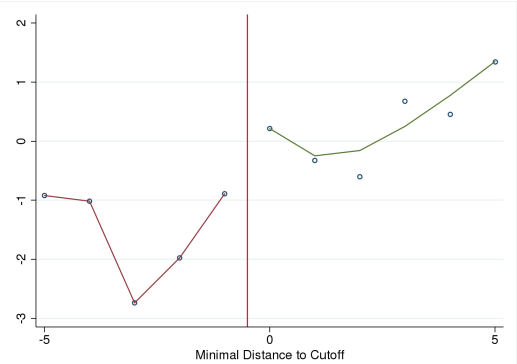
(B) Parental Sense of Overall Safety



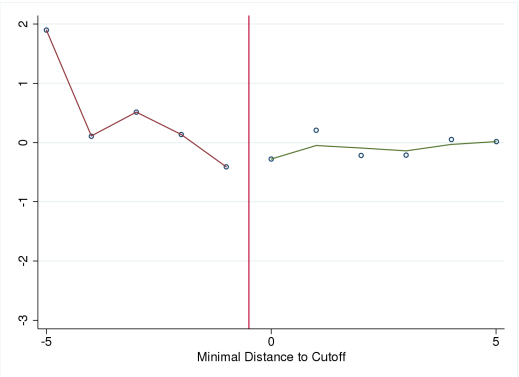
(C) Student Satisfaction



(D) Student Safety



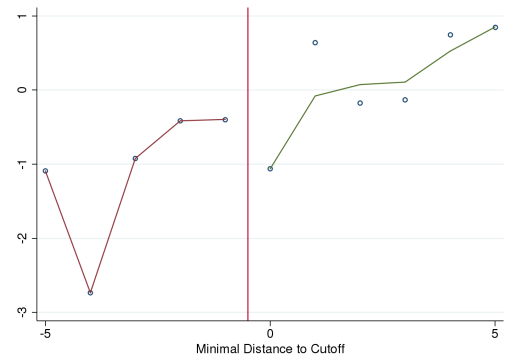
(E) Student Sense of Environment



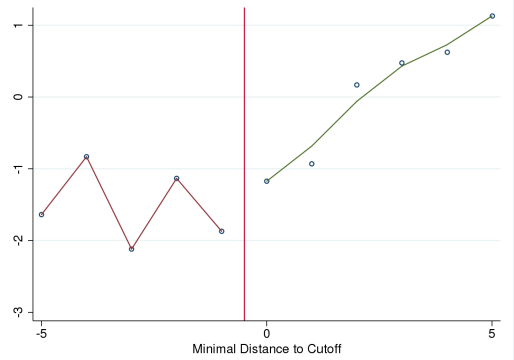
Note: These figures plot residuals from regressions of current (i.e. prior to retention) values of parental satisfaction, parental sense of overall safety, student satisfaction, student safety, and student sense of environment on fixed effects for test grade by test year.

Figure A24: Placebo Effects on Survey Responses by Policies

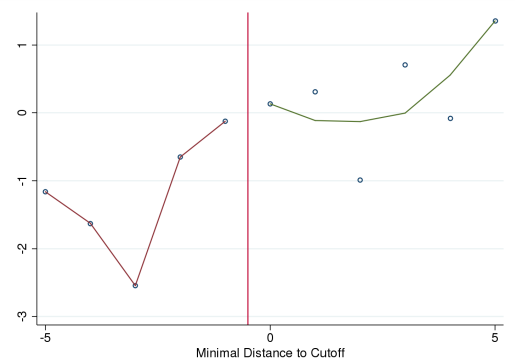
(A) Pre-policy Parental Satisfaction



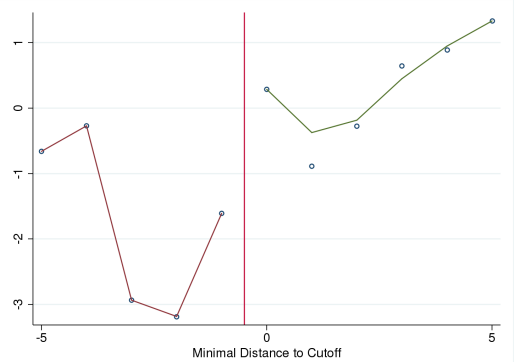
(B) Post-policy Parental Satisfaction



(C) Pre-policy Student Safety



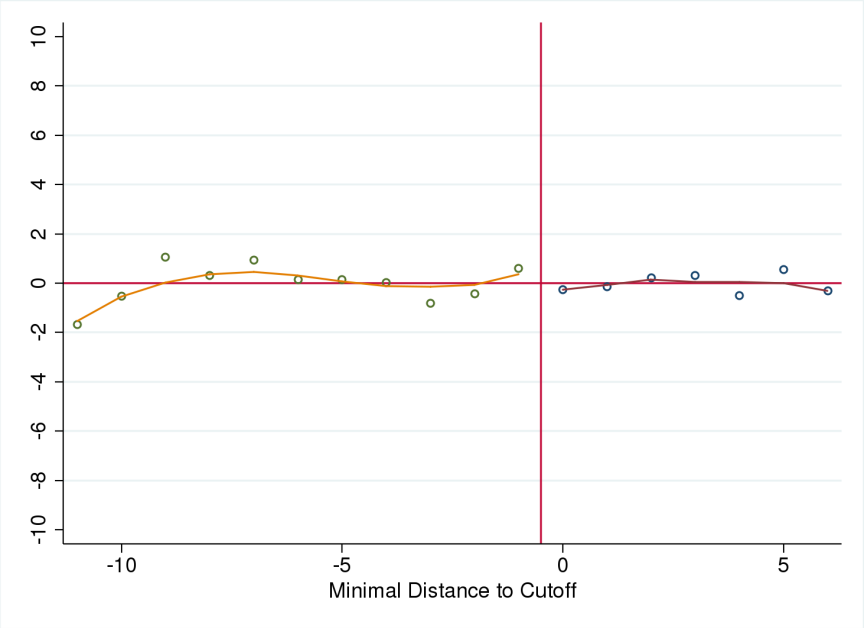
(D) Post-policy Student Safety



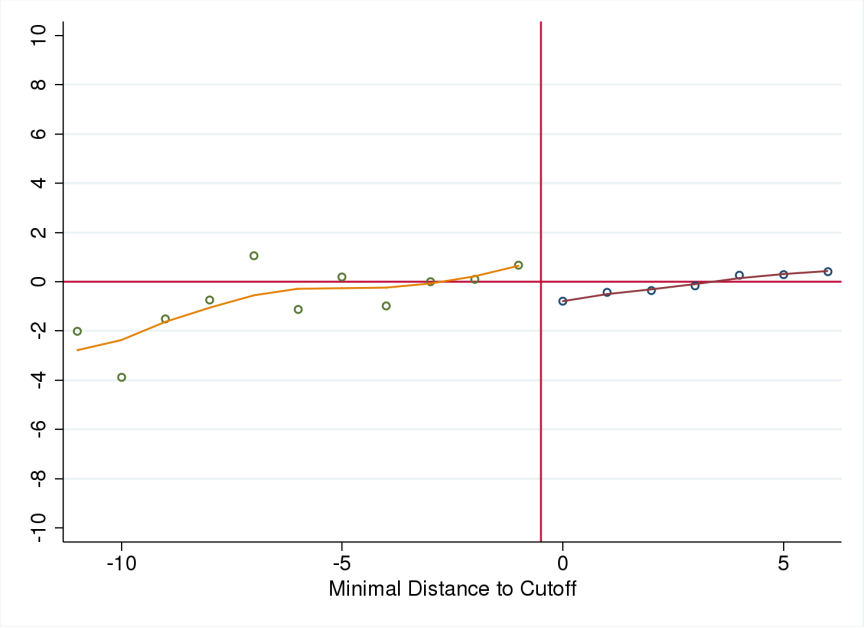
Note: These figures plot residuals from regressions of current (i.e. prior to retention) values of parental satisfaction, parental sense of overall safety, student satisfaction, student safety, and student sense of environment on fixed effects for test grade by test year. Plots are done separately by retention policy regime

Figure A25: CIA Visual Test

(A) Parental Satisfaction

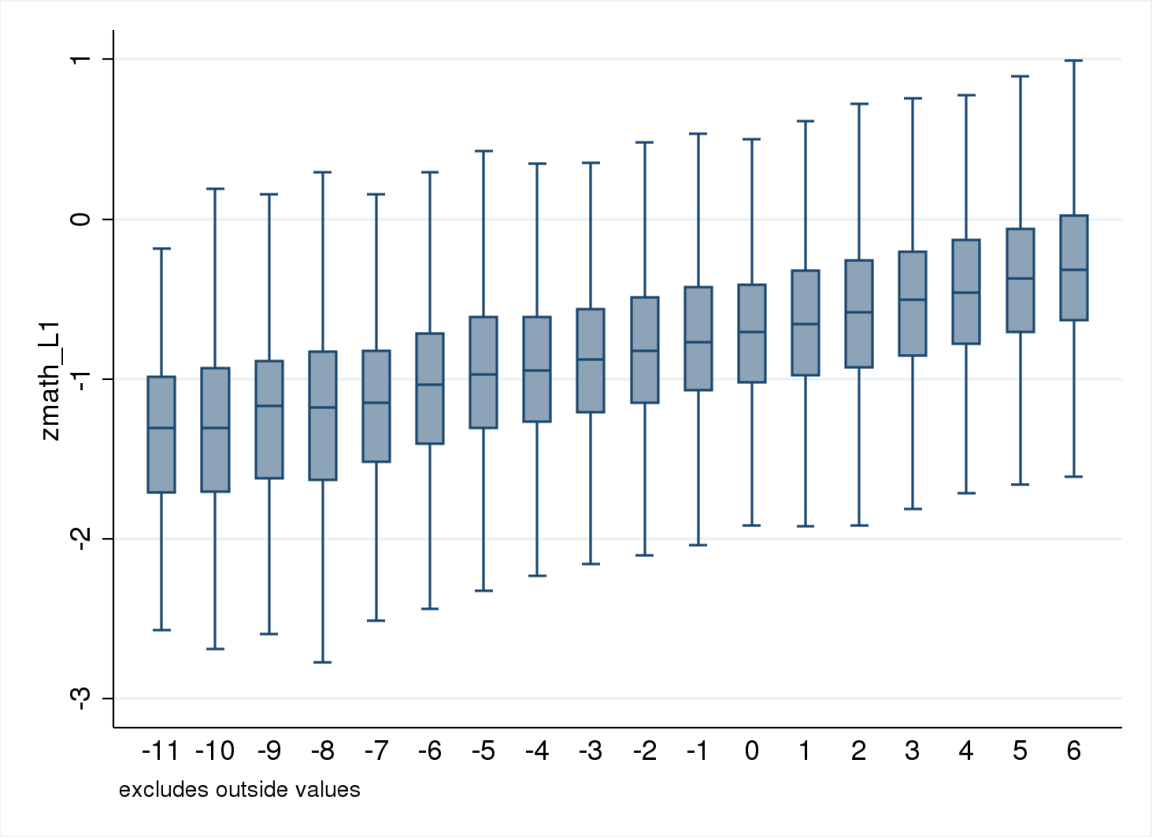


(B) Student Safety



Note: These figures plot residuals from regressions on pre-determined covariates by index score for parental satisfaction and student safety. Comparing the LOWESS and the horizontal line at each side of the cutoff only supports CIA with respect to parental satisfaction outcome. Appendix Table A17 presents regression results and suggests the same conclusion.

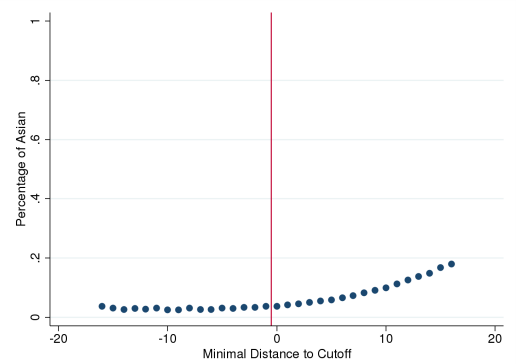
Figure A26: Distribution of Two-year Prior Mathematics Score



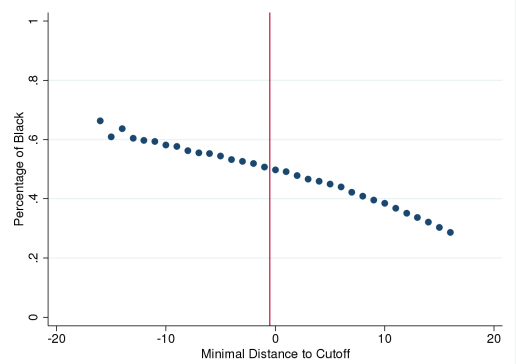
Note: The box plot summarizes the distribution of two-year prior normalized mathematics scores at each index score.

Figure A27: Continuity of Other Personal Characteristics

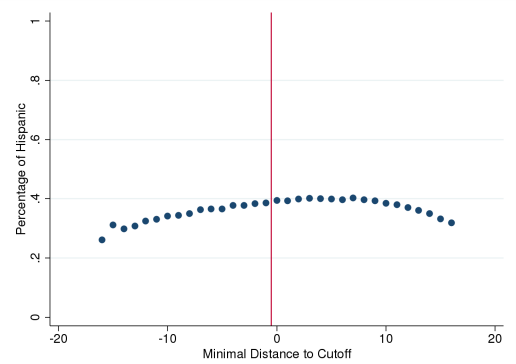
(A) Asian



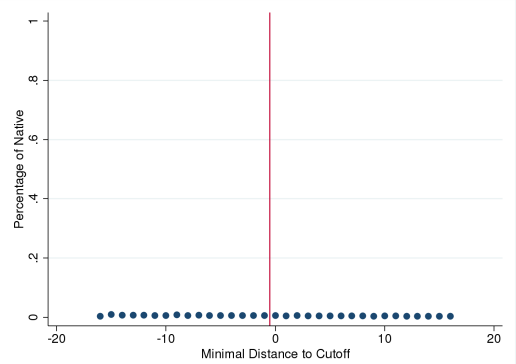
(B) Black



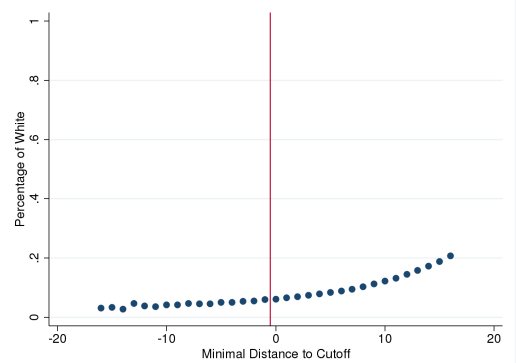
(C) Hispanic



(D) Native

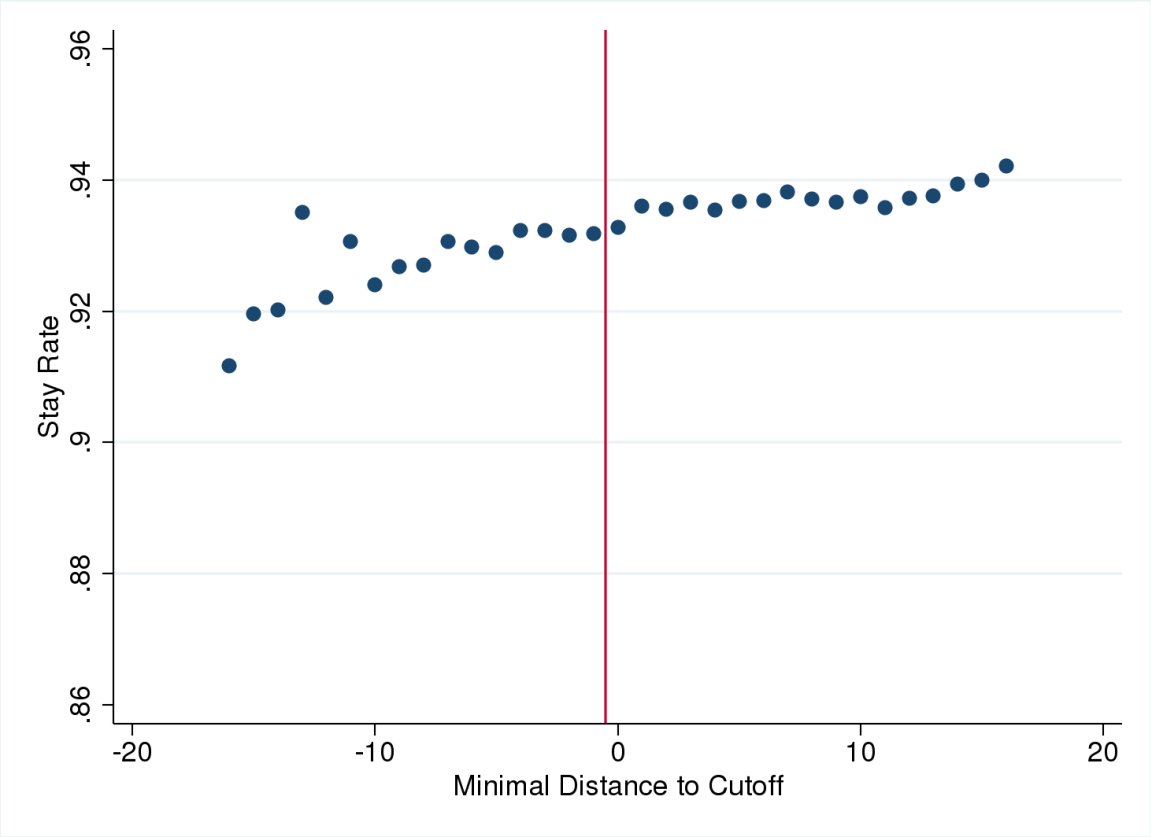


(E) White



Note: These figures plot average percent of Asian (Panel a), Black (Panel b), Hispanic (Panel c), Native (Panel d) , and White (Panel e) students by index score.

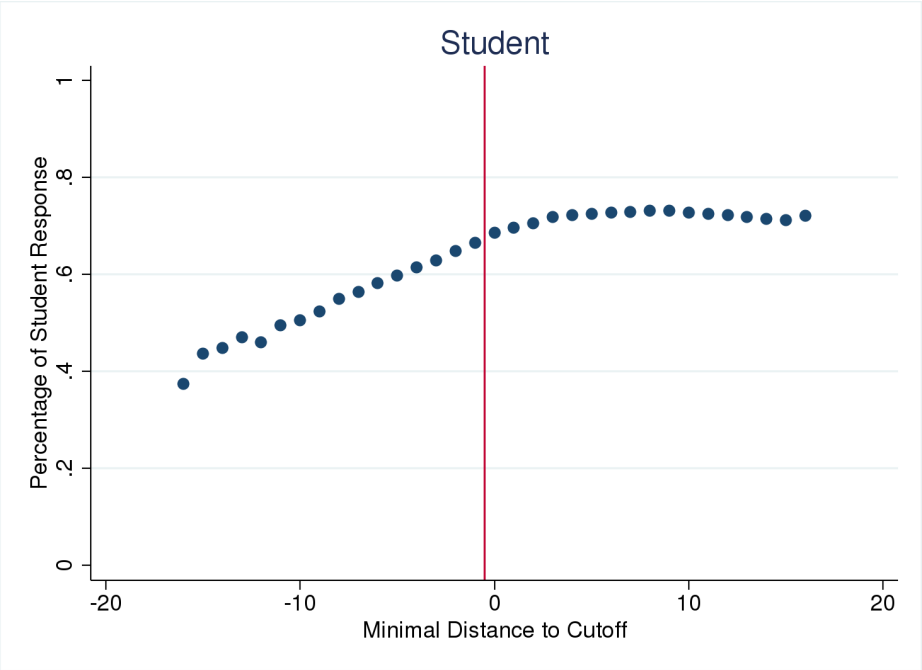
Figure A28: Continuity of Attrition Rate



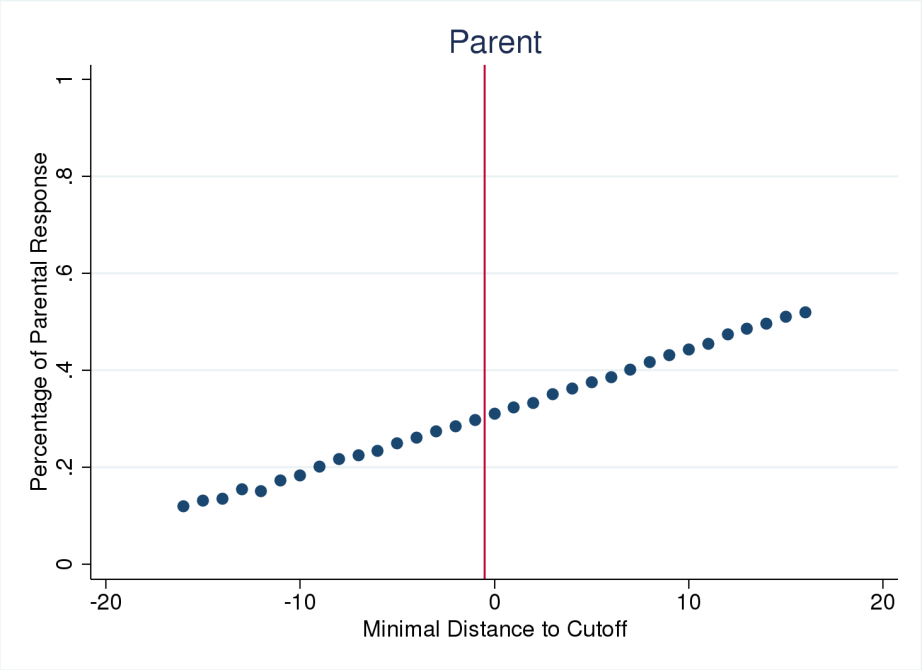
Note: This figure plots average probability for appearing in the datasets next year at each index score.

Figure A29: Continuity of Response Rates against Indexes

(A) Student Response Rate



(B) Parent Response Rate



Note: Each point represents the raw response rate at each index for parents and students. These figures plot percentage of students (a) and parents (b) who respond to the survey.

Appendix Tables

Table A1: Effects of the Retention Policy on Teachers

	Math		English	
	Experience	Absences	Experience	Absences
RetPol	-0.21	0.33	-0.22	0.34
	(0.30)	(0.29)	(0.30)	(0.29)
Observations	192,829	189,689	192,840	189,643

Notes: All regressions implement specification 1.1 and display the coefficient of $RetPol_{igt}$, an indicator of the retention policy. Standard errors are clustered at the school-year level in parentheses.
* $p < .05$, ** $p < .01$, *** $p < .001$.

Table A2: Effects of the Accountability Scheme on Teachers

	Math		English	
	Experience	Absences	Experience	Absences
Low*Act	-0.097	0.027	0.0078	0.037
	(0.065)	(0.070)	(0.063)	(0.070)
Observations	730,520	721,679	731,565	722,223

Notes: All regressions restrict observations to grades not subject to the retention policy and implement specification 1.2. The coefficient of the interaction term $Low_{ist'} * Act_{it}$ is displayed. The interaction term in columns 1 and 2 (columns 3 and 4) is a dummy for the interaction of being a lowest-third student in math (English) and being in the post-accountability era. Standard errors are clustered at the school-year level in parentheses. * $p < .05$, ** $p < .01$, *** $p < .001$.

Table A3: Policy Interaction on Students: Placebo

	Test Scores	Absences	Suspension
Panel A: Math-lowest-third			
Low*Ret*Act	-0.0041	-0.49	-0.0057
	(0.025)	(0.40)	(0.0094)
Panel B: English-lowest-third			
Low*Ret*Act	0.037	0.32	0.014
	(0.022)	(0.41)	(0.010)
Observations	90,916	90,916	90,916

Notes: All regressions implement specification 1.4 for the years between 2003 and 2007 and focus on students exempt from the retention policy. The coefficient of the triple-interaction term $Low_{ist'} * Act_{it} * RetPol_{igt}$ is displayed. The triple-interaction term is a dummy for the triple interaction of being a lowest-third student in math or English, being in post-accountability era, and being subject to the retention policy. Standard errors are clustered at the school-year level in parentheses. * $p < .05$, ** $p < .01$, *** $p < .001$.

Table A4: Policy Interaction on Teachers

	Math		English	
	Experience	Absences	Experience	Absences
Low*Ret*Act	0.059	-0.067	-0.11	-0.15
	(0.11)	(0.12)	(0.11)	(0.12)
Observations	1,110,237	1,095,866	1,111,176	1,096,305

Notes: All regressions implement specification 1.4. The coefficient of the triple-interaction term $Low_{ist'} * Act_{it} * RetPol_{igt}$ is displayed. The triple-interaction term is a dummy for the triple interaction of being a lowest-third student in math or English, being in post-accountability era, and being subject to the retention policy. Standard errors are clustered at the school-year level in parentheses.

* $p < .05$, ** $p < .01$, *** $p < .001$.

Table A5: Policy Interaction on Students: Accountability Robustness

	Test Scores	Absences	Suspension
Panel A: Math-lowest-third			
Low*Ret*Act	0.27***	-0.45***	-0.0066***
	(0.0065)	(0.089)	(0.0019)
Panel B: English-lowest-third			
Low*Ret*Act	0.071***	-0.14	-0.0040*
	(0.0052)	(0.084)	(0.0019)
Observations	1,155,107	1,155,107	1,155,107

Notes: All regressions implement specification 1.4, including year-specific covariates of being a citywide lowest-third student, categorical dummies of ethnicity groups, and an indicator of having prior test scores between 2.5 and 3.5. The coefficient of the triple-interaction term $Low_{ist'} * Act_{it} * RetPol_{igt}$ is displayed. Standard errors are clustered at the school-year level in parentheses. * $p < .05$, ** $p < .01$, *** $p < .001$.

Table A6: Policy Interaction on Students: High-Achieving Schools

	Test Scores	Absences	Suspension
Panel A: Math-lowest-third			
GenEd*Low*Ret*Act	0.022	-0.52	-0.0089
	(0.061)	(0.87)	(0.020)
Observations	227,649	227,649	227,649
Panel B: English-lowest-third			
GenEd*Low*Ret*Act	-0.052	-0.97	0.0060
	(0.060)	(0.95)	(0.024)
Observations	231,714	231,714	231,714

Notes: All regressions focus on students in schools with average test scores above the 75th percentile and implement specification 1.4 interacting with an indicator of being a general education student who is subject to the retention policy. The coefficient of the triple-interaction term $Low_{ist'} * Act_{it} * RetPol_{igt}$ interacting with the indicator of being a general education student is displayed. Standard errors are clustered at the school-year level in parentheses. * $p < .05$, ** $p < .01$, *** $p < .001$.

Table A7: Effects of Retention with Additional Grade and Year

Variable	Parent satisfied	Student feels safe
Retention	-1.920	-6.404*
[placebo]	(4.837)	(3.785)
Retention	5.242**	5.501**
[future]	(2.367)	(2.614)
Observations	189,807	395,442
R-squared	0.042	0.008

Note: Each column reports coefficients from a single regression with grade × year control functions of student's index score and survey grade fixed effects. [placebo] stands for coefficients on current outcome and [future] stands for coefficients on the average outcome in the next three years after tests. We include students who were in 8th grade or tested in school year 2009-2010 in addition to our main RD sample.

Table A8: Persistent Effects of Retention with Additional Grade and Year

Variable	Parent satisfied	Student feels safe
Retention	-1.920	-6.404*
[placebo]	(4.837)	(3.785)
Retention $[l = 1]$	-1.477	10.40*
	(3.429)	(5.897)
Retention $[l = 2]$	4.196	9.618*
	(3.386)	(4.938)
Retention $[l = 3]$	5.192	5.078
	(3.737)	(4.842)
Observations	189,807	395,442
R-squared.	0.044	0.011

Note: Each column reports coefficients from a single regression with grade × year control functions of student's index score and survey grade fixed effects. [placebo] stands for coefficients on current outcome and $[l = 1, 2, 3]$ stands for coefficients on next-year, two-year-later, and three-year-later outcomes, respectively. We include students who were in 8th grade or tested in school year 2009-2010 in addition to our main RD sample.

Table A9: Effects of Retention by Policies with Additional Grade and Year

Variable	Parent satisfied	Student feels safe
Post-Policy Retention	-1.448	-7.126*
[placebo]	(4.975)	(3.856)
Pre-Policy Retention	-10.45	13.47
[placebo]	(20.36)	(19.88)
Post-Policy Retention	5.525**	5.560**
[future]	(2.378)	(2.621)
Pre-Policy Retention	-10.15	1.372
[future]	(13.02)	(13.28)
Observations	189,807	395,442
R-squared.	0.036	0.008

Note: Each column reports coefficients from a single regression with grade × year control functions of student's index score and survey grade fixed effects. [placebo] stands for coefficients on current outcome and [future] stands for coefficients on the average outcome in the next three years after tests. We include students who were in 8th grade or tested in school year 2009-2010 in addition to our main RD sample.

Table A10: An Example of Data Stacking

ID	Test Year	Test Grade	Index	Failing a test	Retention	Survey Year	Survey Grade
1	2007	5	-3	1	1	2007	5
1	2007	5	-3	1	1	2008	5
1	2007	5	-3	1	1	2009	6
1	2007	5	-3	1	1	2010	7

Notes: In this example, a student with identification number 1 was in 5th grade in 2007, took the tests that year, failed the English exam by 3 points, passed the math exam, and was retained. This record is matched to his/her survey response in 2007, which was collected before this student knew his/her test scores and the retention decision, and also matched to survey responses in 2008, 2009, and 2010. Since the test year is the same, his test scores, and therefore the running variable, do not change. His survey grade reflects his grade when he took the survey each year. Because he was retained in 2007, his survey grade is the same in 2008 as in 2007.

Table A11: Effects on Test Scores between Policies

Variable	ELA	Math
Post-Policy Retention	-0.00566	0.00378
[placebo]	(0.0403)	(0.0391)
Pre-Policy Retention	-0.0109	-0.0376
[placebo]	(0.161)	(0.123)
Post-Policy Retention	0.537***	0.625***
[future]	(0.0495)	(0.0582)
Pre-Policy Retention	0.672***	0.674***
[future]	(0.171)	(0.182)
Observations	939,661	939,962
R-squared.	0.186	0.212

Note: Each column reports coefficients from a single regression with grade × year control functions of student's index score. [placebo] stands for coefficients on current outcome and [future] stands for coefficients on the average outcome in the next three years after tests. Test Scores are normalized within each grade × year to have a mean of zero and a standard deviation of one. ELA stands for the English Language Arts exam. Absences are capped at 50 days per year and suspension is an indicator for being suspended at least once during the school year.

Table A12: Effects on Survey Responses by Bandwidths

Bandwidth	(-4,4)	(-6,6)	(-7,7)	(-8,8)	(-9,9)	(-10,10)
Variable	Parent satisfied	Parent satisfied	Parent satisfied	Parent satisfied	Parent satisfied	Parent satisfied
Retention	0.809	0.731	-1.352	-2.113	-1.778	-1.045
[placebo]	(6.047)	(4.387)	(4.045)	(3.741)	(3.512)	(3.358)
Retention	7.474***	6.499***	4.678**	5.726***	4.669***	4.107***
[future]	(2.750)	(2.112)	(1.889)	(1.728)	(1.586)	(1.489)
Observations	130,540	199,393	237,672	279,677	325,712	375,863
R-squared	0.041	0.040	0.041	0.039	0.039	0.040
Variable	Student feels safe	Student feels safe	Student feels safe	Student feels safe	Student feels safe	Student feels safe
Retention	-6.382	-5.976*	-5.822*	-4.093	-2.631	-1.732
[placebo]	(4.526)	(3.531)	(3.262)	(3.034)	(2.878)	(2.751)
Retention	6.144**	5.832**	4.512**	4.735**	4.492**	4.758***
[future]	(3.075)	(2.344)	(2.126)	(1.955)	(1.809)	(1.696)
Observations	248,425	369,328	432,809	499,647	570,109	643,937
R-squared	0.008	0.008	0.009	0.010	0.011	0.012

Note: Each column reports coefficients from a single regression with grade × year control functions of student's index score and survey grade fixed effects. [placebo] stands for coefficients on current outcome and [future] stands for coefficients on the average outcome in the next three years after tests. First row stands for our choice of the bandwidth of the index score near the cutoff.

Table A13: Effects of Retention with Two-way Clustering

Variable	ELA	Math	Absences	Suspension	Parent satisfied	Student feels safe
Retention	-0.00597	0.00129	-0.409	-0.00369	0.235	-6.091**
[placebo]	(0.0401)	(0.0359)	(1.029)	(0.0138)	(3.169)	(2.854)
Retention	0.546***	0.629***	0.533	0.0481***	5.138**	6.133**
[future]	(0.0396)	(0.0498)	(1.325)	(0.0162)	(2.377)	(3.023)
Observations	939,661	939,962	945,555	945,555	163,594	307,465
R-squared.	0.190	0.214	0.035	0.021	0.042	0.008

Note: Each column reports coefficients from a single regression with grade × year control functions of student's index score (and survey grade fixed effects for last two columns). [placebo] stands for coefficients on current outcome and [future] stands for coefficients on the average outcome in the next three years after tests. Test Scores are normalized within each grade × year to have a mean of zero and a standard deviation of one. ELA stands for the English Language Arts exam. Absences are capped at 50 days per year and suspension is an indicator for being suspended at least once during the school year. Some estimates in this table are different from previous ones because two-way clustering is only implementable under ivreg2 in Stata and we use ivregress in previous analysis. Some anecdotes suggest ivreg2 has some algorithm issues that may cause the discrepancies. In most cases in our analysis, the two estimates from ivreg2 and ivregress are very close.

Table A14: Persistent Effects of Retention with Two-way Clustering

Variable	ELA	Math	Parent satisfied	Student feels safe
Retention	-0.00597	0.00129	0.235	-6.091**
[placebo]	(0.0401)	(0.0359)	(3.169)	(2.854)
Retention $[l = 1]$	0.664***	0.788***	-1.229	11.72
	(0.0594)	(0.0657)	(3.602)	(7.198)
Retention $[l = 2]$	0.447***	0.497***	4.761	9.672
	(0.0405)	(0.0691)	(3.239)	(6.188)
Retention $[l = 3]$	0.362***	0.386***	4.314	7.559
	(0.0782)	(0.0832)	(3.081)	(5.979)
Observations	939,661	939,962	163,594	307,481
R-squared.	0.194	0.215	0.044	0.012

Note: Each column reports coefficients from a single regression with grade × year control functions of student's index score (and survey grade fixed effects for last two columns). [placebo] stands for coefficients on current outcome and $[l = 1, 2, 3]$ stands for coefficients on next-year, two-year-later, and three-year-later outcomes, respectively. Test Scores are normalized within each grade × year to have a mean of zero and a standard deviation of one. ELA stands for the English Language Arts exam. Some estimates in this table are different from previous ones because two-way clustering is only implementable under ivreg2 in Stata and we use ivregress in previous analysis. Some anecdotes suggest ivreg2 has some algorithm issues that may cause the discrepancies. In most cases in our analysis, the two estimates from ivreg2 and ivregress are very close.

Table A15: Effects of Retention by Policies with Two-way Clustering

Variable	ELA	Math	Parent satisfied	Student feels safe
Post-Policy	-0.00566	0.00378	-0.211	-5.961**
Retention [placebo]	(0.0417)	(0.0377)	(3.209)	(2.911)
Pre-Policy	-0.0109	-0.0376	10.63	-8.385
Retention [placebo]	(0.146)	(0.0896)	(19.19)	(14.57)
Post-Policy	0.537***	0.625***	5.495**	6.512**
Retention [future]	(0.0418)	(0.0524)	(2.403)	(3.107)
Pre-Policy	0.672***	0.674***	-7.617	-8.939
Retention [future]	(0.119)	(0.145)	(10.45)	(11.14)
Observations	939,661	939,962	163,594	307,481
R-squared.	0.186	0.212	0.038	0.006

Note: Each column reports coefficients from a single regression with grade × year control functions of student's index score (and survey grade fixed effects for last two columns). [placebo] stands for coefficients on current outcome and [future] stands for coefficients on the average outcome in the next three years after tests. Test Scores are normalized within each grade × year to have a mean of zero and a standard deviation of one. ELA stands for the English Language Arts exam. Some estimates in this table are different from previous ones because two-way clustering is only implementable under ivreg2 in Stata and we use ivregress in previous analysis. Some anecdotes suggest ivreg2 has some algorithm issues that may cause the discrepancies. In most cases in our analysis, the two estimates from ivreg2 and ivregress are very close.

Table A16: Effects on Survey Responses with Additional Covariates

Variable	Parent satisfied	Parent satisfied	Parent satisfied	Student feels safe	Student feels safe	Student feels safe
Retention [placebo]	0.217	0.220	0.235	-6.263	-6.251	-6.091
	(5.023)	(5.026)	(5.025)	(3.936)	(3.931)	(3.935)
Retention [future]	5.749*	7.170**	6.767**	7.881**	9.260**	9.234**
	(3.164)	(3.220)	(3.216)	(3.583)	(3.920)	(3.917)
Tenure at School	Yes	No	No	Yes	No	No
School Type	No	Yes	No	No	Yes	No
Special Education	No	No	Yes	No	No	Yes
Observations	136,852	122,322	122,439	239,732	213,162	213,372
R-squared	0.041	0.038	0.039	0.009	0.005	0.005

Note: Each column reports coefficients from a single regression with grade × year control functions of student's index score, survey grade fixed effects, and additional covariates as indicated in the table. [placebo] stands for coefficients on current outcome and [future] stands for coefficients on the average outcome in the next three years after tests. First row stands for our choice of the bandwidth of the index score near the cutoff.

Table A17: CIA Test

Variable	Parental Satisfaction		Students' Personal Safety	
	Right	Left	Right	Left
Index Score	0.00340	0.00472	0.225***	0.267***
	(0.0491)	(0.0921)	(0.0400)	(0.0736)
Observations	37,186	9,402	98,965	24,675
R-squared.	0.030	0.035	0.012	0.013

Note: Each column reports coefficients from regressing parental satisfaction and student safety on pre-determined covariates and the index score. Column two and four (three and five) restrict the sample to observations at the right (left) of cutoff.

Table A18: Continuity of Covariates Test

Variables	Female	Native	Hispanic	Stay	Parent Resp	Density
Failure	0.000710	-0.000397	-0.00509	0.000320	-0.00290	-0.0201
	(0.00526)	(0.000812)	(0.00514)	(0.00294)	(0.00491)	(0.0164)
Observations	437,420	437,420	437,420	342,828	437,420	11
R-squared	0.001	0.000	0.001	0.001	0.148	1
Variables	Asian	Black	White	Free lunch	Student Resp	
Failure	0.00421*	-0.00118	0.00244	-0.00292	-0.00691	
	(0.00217)	(0.00525)	(0.00265)	(0.00268)	(0.00486)	
Observations	437,420	437,420	437,420	424,060	437,420	
R-squared	0.004	0.008	0.006	0.003	0.191	

Note: Each column reports coefficients from a single regression with controls for student's index score. Stay stands for appearing in the datasets next year and Parent (Student) Resp stands for whether the parent (student) ever responded to the surveys.