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

TONG GENG*

JOB MARKET PAPER†

January 22, 2018

Abstract

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.

*Department of Economics, Columbia University. Email: tg2430@columbia.edu.

†I am deeply grateful to Bentley MacLeod, Jonah Rockoff, and Miguel Urquiola for their comments and encouragement. I also thank Peter Bergman, Michael Best, George Bulman, Alex Eble, Michael Gilraine, Rucker Johnson, Wojciech Kopczuk, Robert McMillan, Randall Reback, Evan Riehl, Miikka Rokkanen, Norbert Schady, Elizabeth Setren, Mandy Shen, Ludger Woessmann, and participants of the Columbia Applied Microeconomics Colloquium for helpful discussions. All errors are my own.

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. 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 [Athey and Stern, 1998, Almond and Mazumder, 2013].

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]

[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.

[2] NYC is the largest school district in the United States, and this paper uses administrative data that include

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 learners) 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 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 pos-

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.

[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.

[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.

itive 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 [Jacob and Rockoff, 2012, Bloom et al., 2015, Mbiti et al., 2016] and a larger literature on organizational complementarities in other settings [Milgrom and Roberts, 1995, Brynjolfsson and Milgrom, 2013]. 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.

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

[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.

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.

2.1 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 et al., 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 [McCombs et al., 2009]. 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 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

[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.

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.

Figure 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: $index_{ist} = score_{ist} - cutoff_{gst}$, 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.

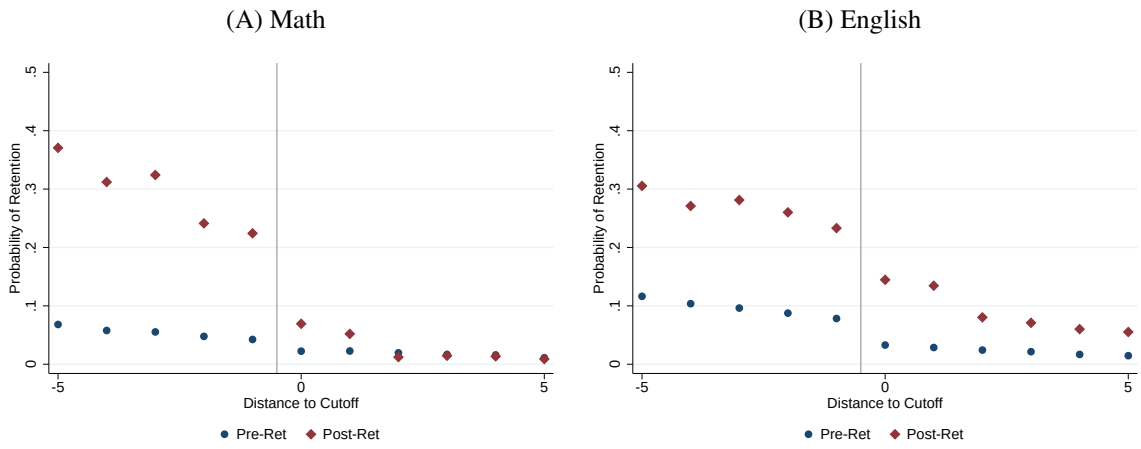
Figure 2 converts Figure 1 into a time series and shows that the increase in retention risks occurred in 2005, when the policy was implemented. Each point restricts the observations to the students in Figure 1 and represents the probability of retention conditional on failing the test in each subject-grade-year 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]

[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.

[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, 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.

Figure 1: The Probability of Retention for Eligible Students

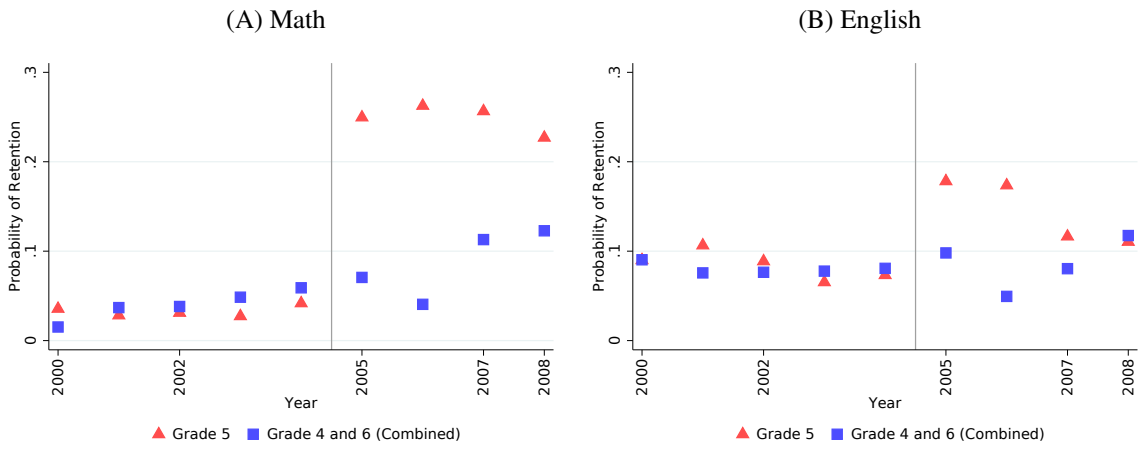


Notes: Both panels are restricted to the years prior to the accountability scheme (prior to 2007). Each point represents the probability of being retained at each value of the index. The index is defined as the difference between a student's spring test score and the cutoff in each subject. Students 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.

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 pressure from peers and teachers [Byrnes, 1989, Andrew, 2014]. 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 [Jacob and Lefgren, 2004, Geng and Rockoff, 2016, Ozek, 2015, Eren et al., 2017].

Figure 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 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.

2.2 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 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

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

[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`.

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-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]

[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.

[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.

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 importance for educational equity and many studies [e.g., Neal and Schanzenbach, 2010, Ladd and Lauen, 2010, Deming et al., 2016] 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, 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 [Reback, 2008, Neal and Schanzenbach, 2010, Macartney et al., 2015] and varying accountability pressure on students' test scores [Reback et al., 2014, Deming et al., 2016], 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.

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

[20] See Figlio et al. [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 [Rockoff and Turner, 2010, Rouse et al., 2013, Chiang, 2009].

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

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 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 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 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.

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*.

4.1 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). Ac-

[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.

Table 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.

cording 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:

$$A_{ist} = \beta_0 + \gamma' X_{it} + \delta_t + \beta_1 Risk_{ist} + \beta_2 Risk_{ist} * RetPol_{it} + \epsilon_{ist} \quad (1)$$

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 et al., 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.

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 et al. [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 et al. [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 account-

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

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

ability scheme. There are potentially two issues associated with the year 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.

4.2 Graphical Evidence and Inference

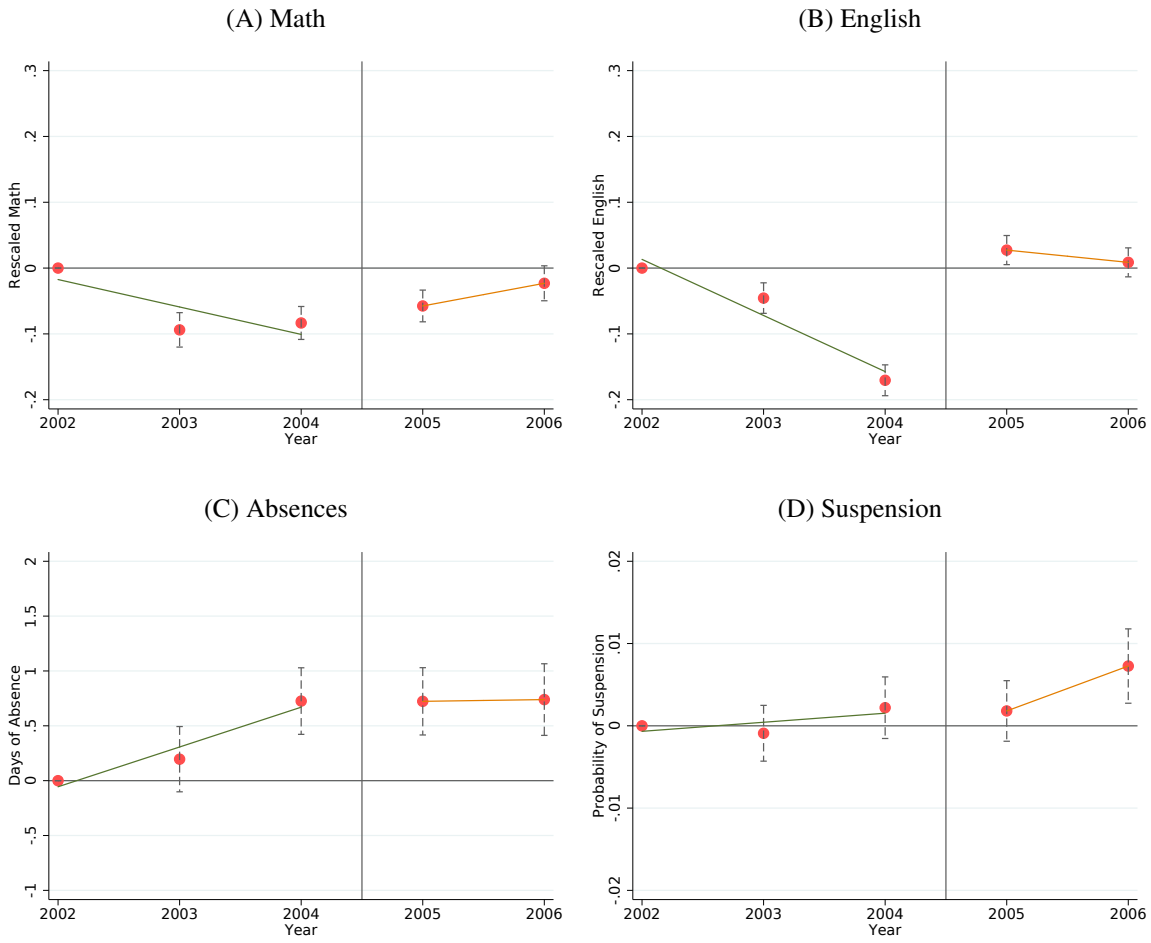
Figure 3 plots coefficients with 95% confidence intervals from an event-study version of Equation 1, which calculates β_2 for each year. Panels A and B present the results for math and English test scores and show a clear difference in the pre-treatment trends for the treatment and control groups, which 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 4 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 A7 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%.

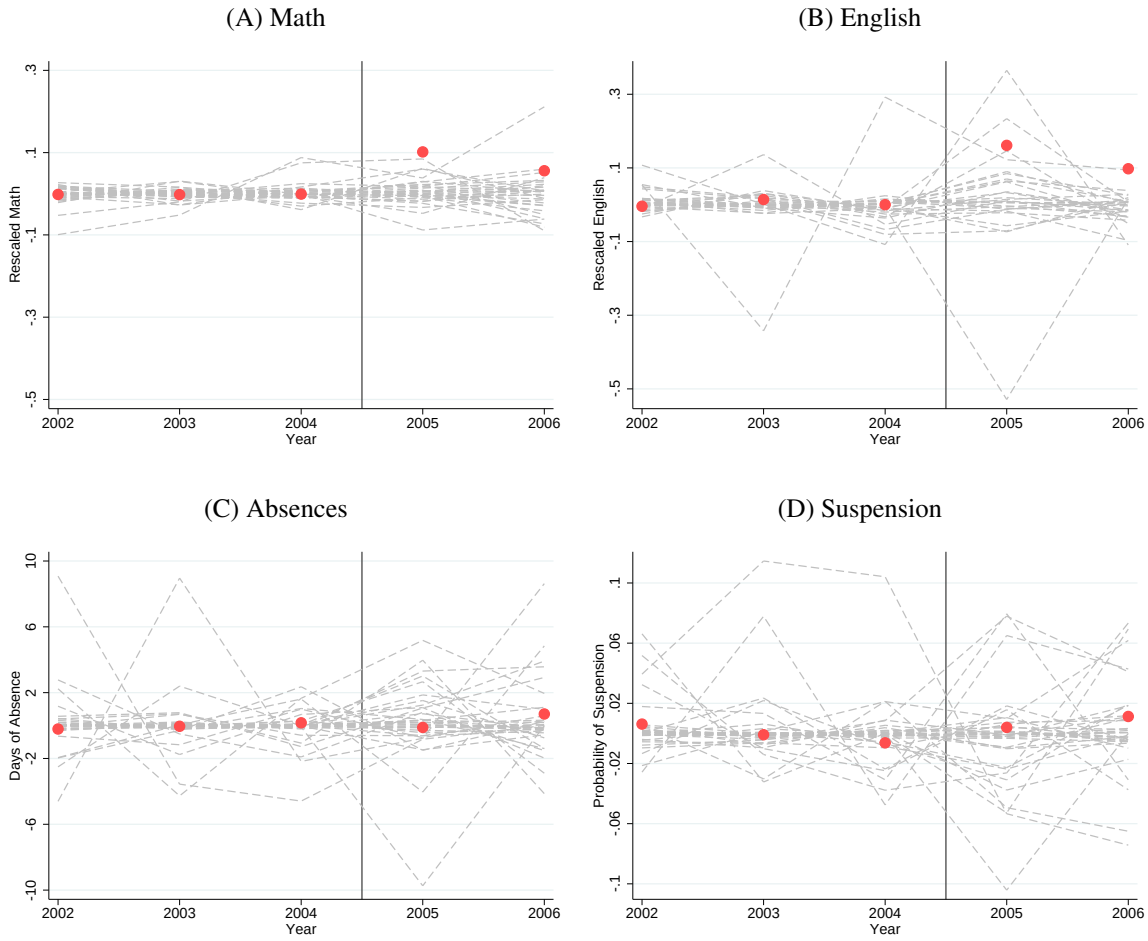
One concern is whether the policy only affected at-risk students, since teachers' efforts and Saturday schools may have benefited other students. To explore this possibility, I examine the distributional effects on eligible students, using exempt students as a con-

Figure 3: 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. 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 4: 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.

trol group.[26] Since these two groups of students might be incomparable, such evidence is suggestive.[27] Appendix Figure A8 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 have an overall improvement in all students' test scores.

Appendix Figure A9 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 A10 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.

[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.

[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 et al. [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.

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-third students.

5.1 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:

$$A_{ist} = \beta_0 + \phi F_{gr}(A_{it'}) + \gamma' X_{it} + \theta_{gt} + \beta_1 Low_{ist'} + \beta_2 Low_{ist'} * Act_{it} + \epsilon_{ist} \qquad (2)$$

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 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 A11 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

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

[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.

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 A12 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.

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:

$$A_{it} = \beta_0 + \phi F_{gr}(A_{it'}) + \gamma' X_{it} + \theta_{gt} + \underbrace{\epsilon_{it}}_{\widetilde{A_{it}}} \tag{3}$$

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.

5.2 Graphical Evidence

Figure 5 presents an event-study version of Equation 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

[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.

B, D, and F present the results for English-lowest-third students. There is an upward trend in the pre-treatment period but no clear jump in the year of the policy.

Appendix Figure A13 presents the distributional effects and supports Figure 5. 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 A14 uses the same specification and shows no discernible discontinuity for teachers' experience levels and absences in the year of the accountability scheme.

Appendix Figure A15 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 5. However, it appears that the accountability scheme induced little improvement in these lowest-third students' test scores, absences, and suspension rates.

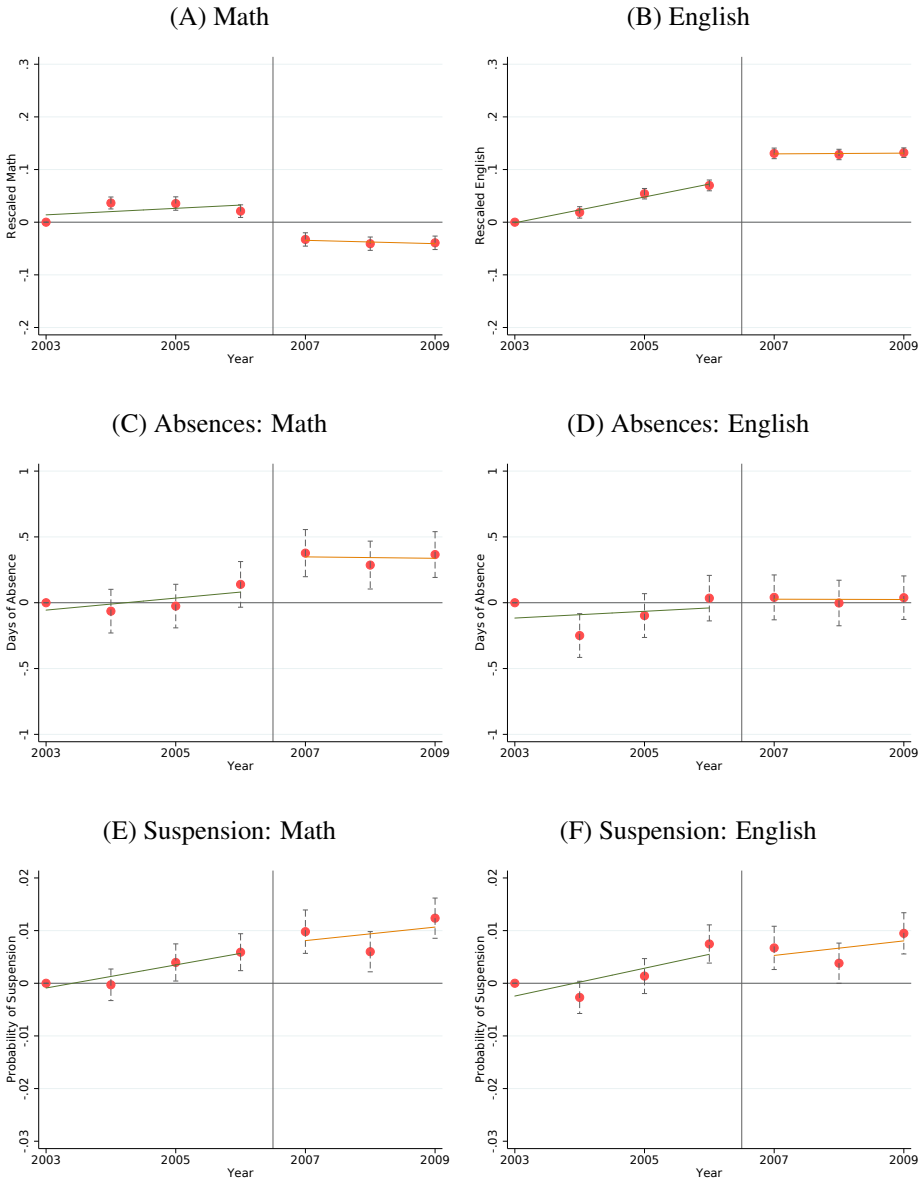
5.3 Regression Results

Table 2 presents the point estimates for general education students. The results are generated by Equation 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 A16 plots

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

Figure 5: 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 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.

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

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.

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.

6.1 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:

$$A_{ist} = \beta_0 + \phi F_{gr}(A_{it'}) + \gamma' X_{it} + \theta_{gt} + \beta_1 Low_{ist'} + \beta_2 Low_{ist'} * G5_{it} + \beta_3 Low_{ist'} * RetPol_{it} + \beta_4 Low_{ist'} * Act_{it} + \beta_5 Low_{ist'} * RetPol_{it} * Act_{it} + \epsilon_{ist} \quad (4)$$

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.

6.2 Graphical Evidence

Figure 6 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.

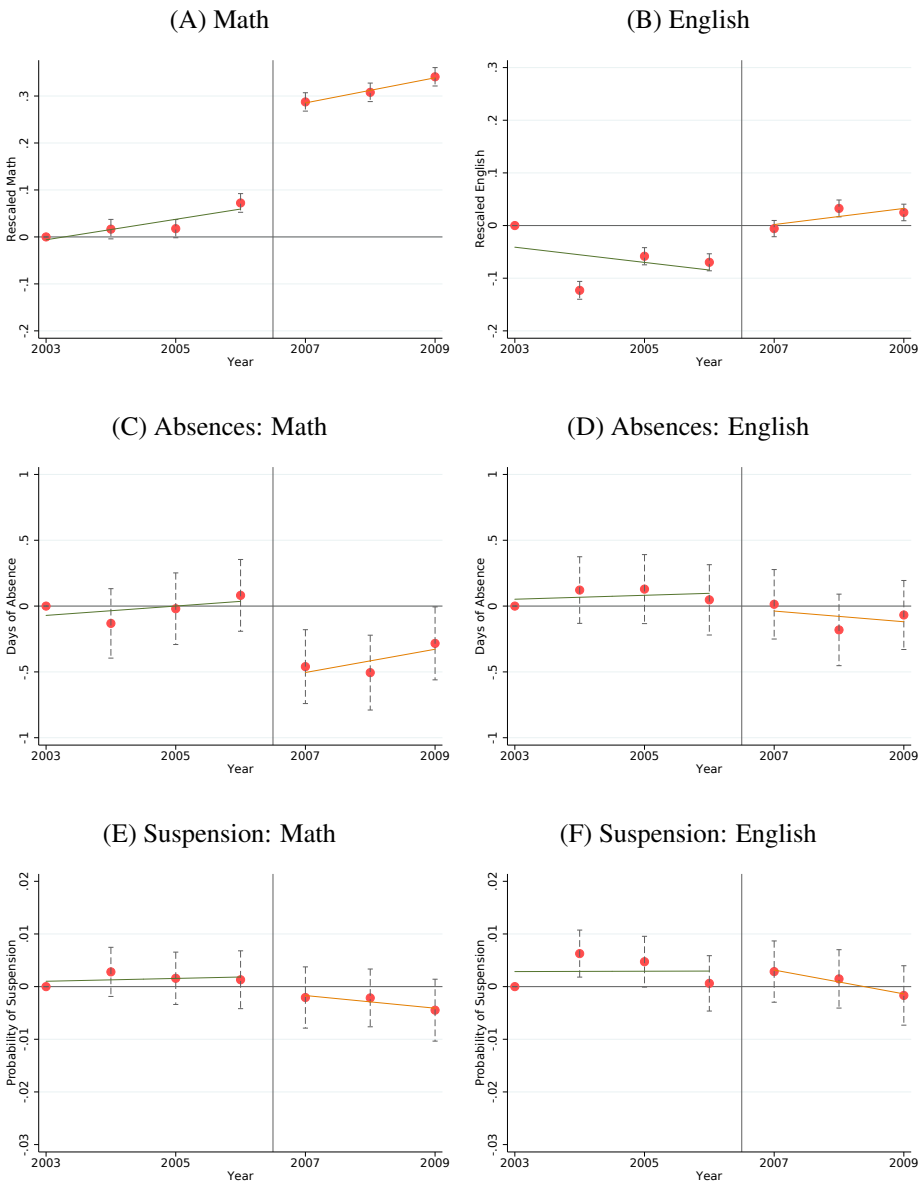
Figure 7 presents the placebo test, using students exempt from the retention policy. Because of the smaller sample size, the overall patterns are jumpier and noisier. Panels A, C, and E present the results for math-lowest-third exempt students and show no evidence of any effects in the year when the 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 A17 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 6 are divided into three periods: years prior to both policies, years with only the retention policy, and years with both policies.

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

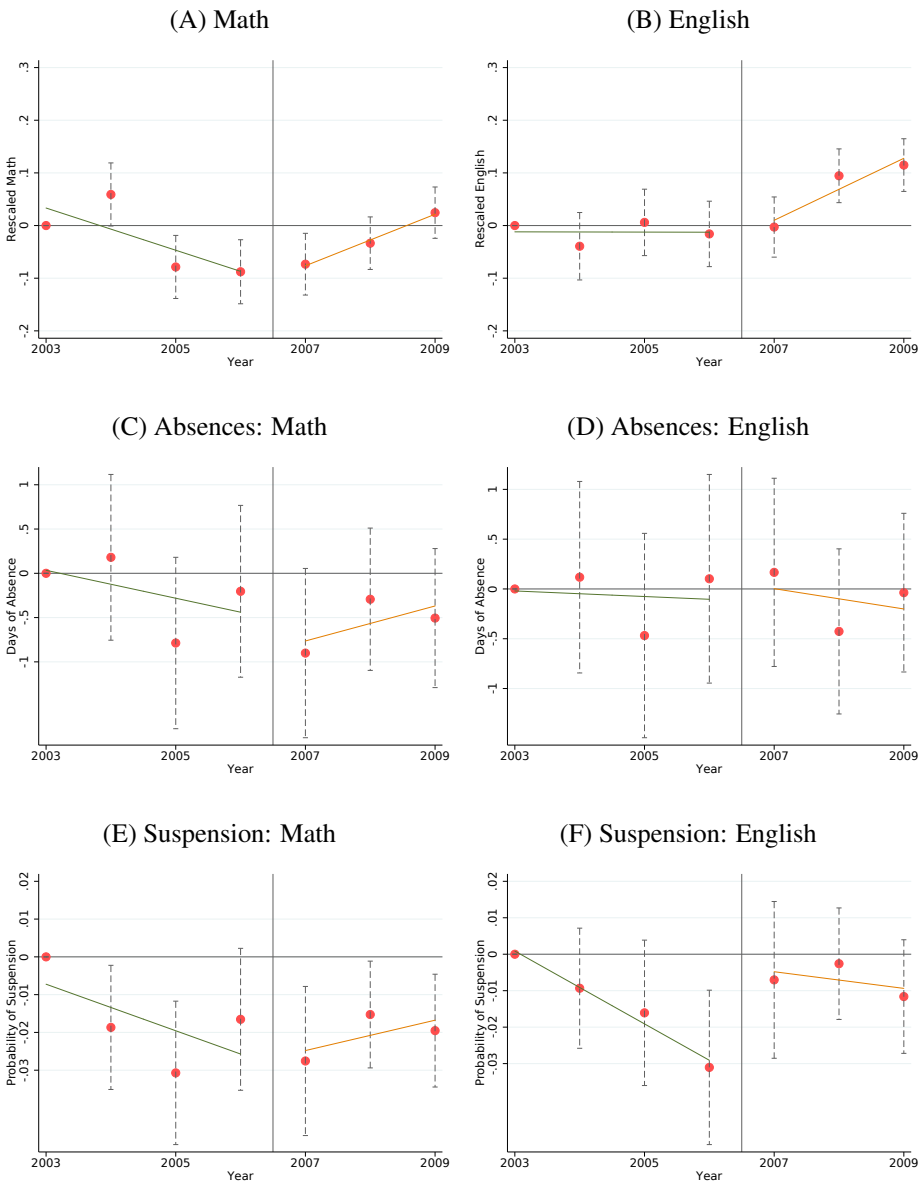
[35]The increase in suspension rates is driven by both a sharp increase in 5th grade and a drop in 6th grade, but the exact cause is unclear.

Figure 6: 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 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 7: 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 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.

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 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 A18 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.

6.3 Regression Results

Table 3 presents the point estimates based on Equation 4 for students subject to the retention policy. Panel A shows that math-lowest-third students experienced 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,

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

[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).

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.

6.4 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.

[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.

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

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

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 shows no effects on all outcomes.[43]

6.5 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

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

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, 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.

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, ab-

sences, 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 et al., 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.

References

A. Abadie, A. Diamond, and J. Hainmueller. Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American Statistical Association*, 105(490):493–505, 2010.

A. Abadie, A. Diamond, and J. Hainmueller. Synth: Stata module to implement synthetic control methods for comparative case studies, 2014.

A. Abadie, A. Diamond, and J. Hainmueller. Comparative politics and the synthetic control method. *American Journal of Political Science*, 59(2):495–510, 2015.

D. Almond and B. Mazumder. Fetal origins and parental responses. *Annu. Rev. Econ.*, 5 (1):37–56, 2013.

M. Andrew. The scarring effects of primary-grade retention? A study of cumulative advantage in the educational career. *Social Forces*, page 74, 2014.

S. Athey and S. Stern. An empirical framework for testing theories about complimentarity in organizational design. Technical report, National Bureau of Economic Research, 1998.

J. R. Behrman, S. W. Parker, P. E. Todd, and K. I. Wolpin. Aligning learning incentives of students and teachers: Results from a social experiment in Mexican high schools. *Journal of Political Economy*, 123(2):325–364, 2015.

N. Bloom, R. Lemos, R. Sadun, and J. Van Reenen. Does management matter in schools? *The Economic Journal*, 125(584):647–674, 2015.

E. Brynjolfsson and P. Milgrom. Complementarity in organizations. *The Handbook of Organizational Economics*, pages 11–55, 2013.

D. A. Byrnes. Attitudes of students, parents and educators toward repeating a grade. *Flunking grades: The policies and effects of retention*, 1989.

H. Chiang. How accountability pressure on failing schools affects student achievement. *Journal of Public Economics*, 93(9):1045–1057, 2009.

D. J. Deming, S. Cohodes, J. Jennings, and C. Jencks. School accountability, postsecondary attainment, and earnings. *Review of Economics and Statistics*, 98(5):848–862, 2016.

A. Dixit. Incentives and organizations in the public sector: An interpretative review. *Journal of Human Resources*, pages 696–727, 2002.

O. Eren, B. Depew, and S. Barnes. Test-based promotion policies, dropping out, and juvenile crime. *Journal of Public Economics*, 2017.

D. Figlio, S. Loeb, et al. School accountability. *Handbook of the Economics of Education*, 3(8):383–417, 2011.

R. G. Fryer. Teacher incentives and student achievement: Evidence from New York City public schools. *Journal of Labor Economics*, 31(2):373–407, 2013.

T. Geng and J. E. Rockoff. Does repeating a grade make students (and parents) happier? Regression Discontinuity Evidence from New York City. Technical report, Columbia University, 2016.

B. Holmstrom and P. Milgrom. The firm as an incentive system. *The American Economic Review*, pages 972–991, 1994.

B. A. Jacob and L. Lefgren. Remedial education and student achievement: A regression-discontinuity analysis. *The Review of Economics and Statistics*, 86(1):226–244, 2004.

B. A. Jacob and J. E. Rockoff. Organizing schools to improve student achievement: Start times, grade configurations, and teacher assignments. *The Education Digest*, 77(8):28, 2012.

R. C. Johnson and C. K. Jackson. Reducing inequality through dynamic complementarity: Evidence from Head Start and public school spending. Technical report, National Bureau of Economic Research, 2017.

M. F. Koppensteiner. Automatic grade promotion and student performance: Evidence from Brazil. *Journal of Development Economics*, 107:277–290, 2014.

H. F. Ladd and D. L. Lauen. Status versus growth: The distributional effects of school accountability policies. *Journal of Policy Analysis and Management*, 29(3):426–450, 2010.

H. Macartney, R. McMillan, and U. Petronijevic. Incentive design in education: An empirical analysis. Technical report, National Bureau of Economic Research, 2015.

O. Malamud, C. Pop-Eleches, and M. Urquiola. Interactions between family and school environments: Evidence on dynamic complementarities? Technical report, National Bureau of Economic Research, 2016.

I. Mbiti, K. Muralidharan, M. Romero, Y. Schipper, C. Manda, and R. Rajani. Inputs, incentives, and complementarities in primary education: Experimental evidence from Tanzania. *Unpublished Paper*, 2016.

J. S. McCombs, S. N. Kirby, and L. T. Mariano. *Ending social promotion without leaving children behind: The case of New York City*. Rand Corporation, 2009.

P. Milgrom and J. Roberts. Complementarities and fit strategy, structure, and organizational change in manufacturing. *Journal of Accounting and Economics*, 19(2):179–208, 1995.

D. Neal and D. W. Schanzenbach. Left behind by design: Proficiency counts and test-based accountability. *The Review of Economics and Statistics*, 92(2):263–283, 2010.

U. Ozek. Hold back to move forward? Early grade retention and student misbehavior. *Education Finance and Policy*, 10(3):350–377, Summer 2015.

R. Reback. Teaching to the rating: School accountability and the distribution of student achievement. *Journal of Public Economics*, 92(5):1394–1415, 2008.

R. Reback, J. Rockoff, and H. L. Schwartz. Under pressure: Job security, resource allocation, and productivity in schools under No Child Left Behind. *American Economic Journal: Economic Policy*, 6(3):207–241, 2014.

S. G. Rivkin, E. A. Hanushek, and J. F. Kain. Teachers, schools, and academic achievement. *Econometrica*, 73(2):417–458, 2005.

J. Rockoff and L. J. Turner. Short-run impacts of accountability on school quality. *American Economic Journal: Economic Policy*, 2(4):119–147, 2010.

S. Rose. Third grade reading policies. *Education Commission of the States (NJ3)*, 2012.

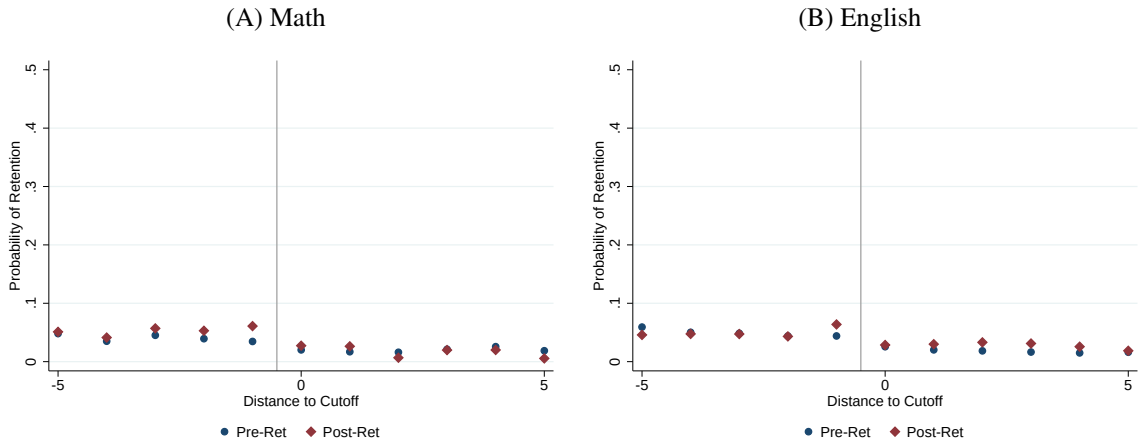
C. E. Rouse, J. Hannaway, D. Goldhaber, and D. Figlio. Feeling the Florida heat? How low-performing schools respond to voucher and accountability pressure. *American Economic Journal: Economic Policy*, pages 251–281, 2013.

P. E. Todd and K. I. Wolpin. On the specification and estimation of the production function for cognitive achievement. *The Economic Journal*, 113(485), 2003.

P. E. Todd and K. I. Wolpin. Estimating a coordination game within the classroom. *Manuscript, Univ. Pennsylvania*, 2012.

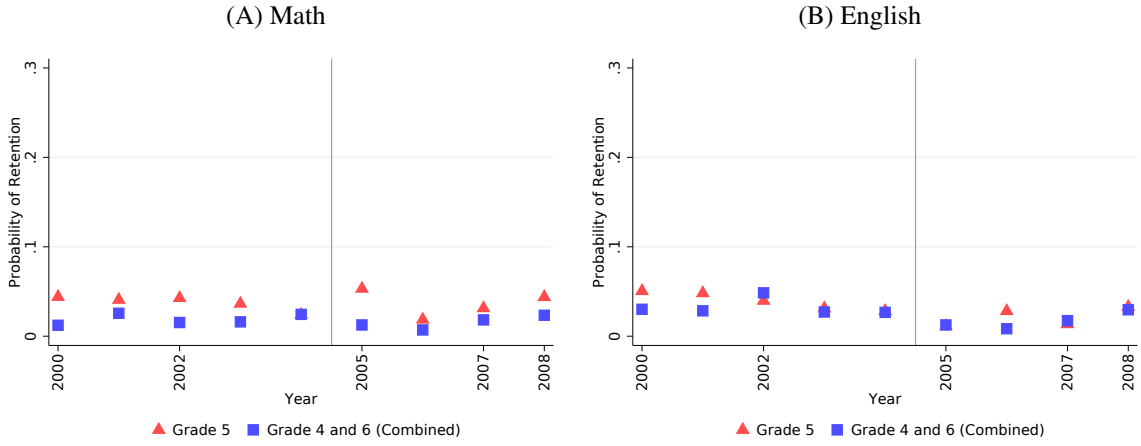
Appendix A Figures

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 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.

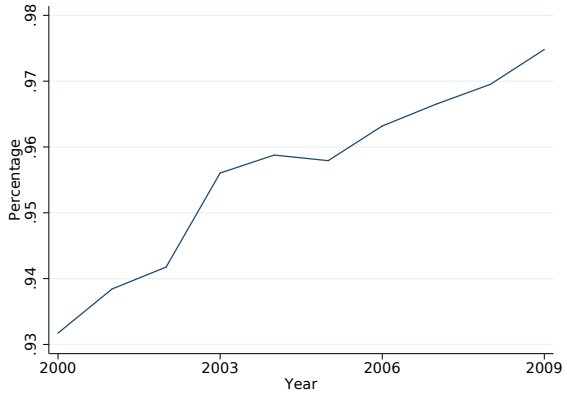
The maximum point values for each measure are indicated in the table below:

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
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			

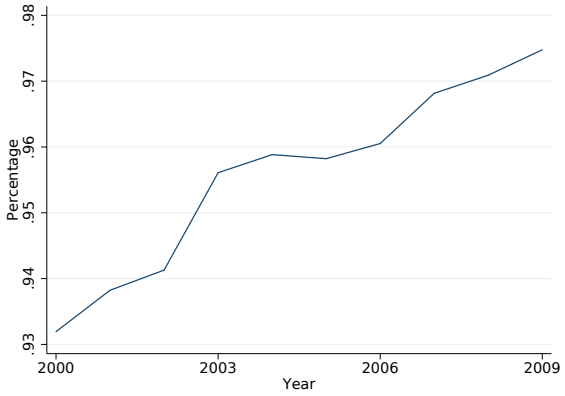
Notes: See the website of the New York City Department of Education for full documentation: `http://schools.nyc.gov/Accountability/tools/report/ProgressReport_2007-2013.htm`.

Figure A4: Selection

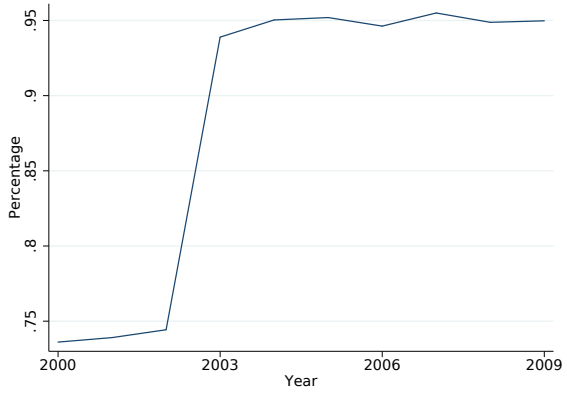
(A) Eligible: Math



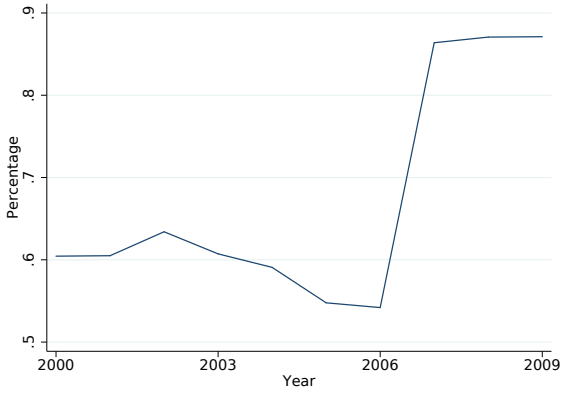
(B) Eligible: English



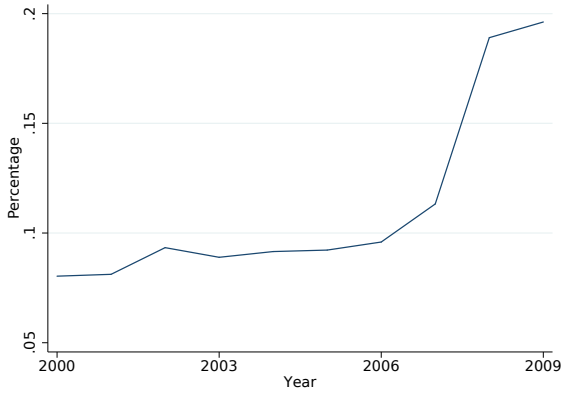
(C) Exempt: Math



(D) Exempt: English

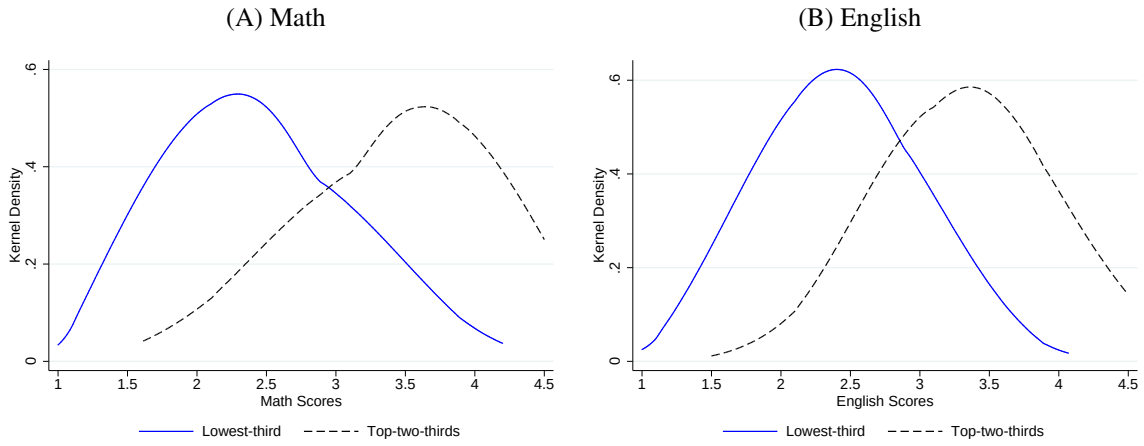


(E) Percentage of Exempt Students



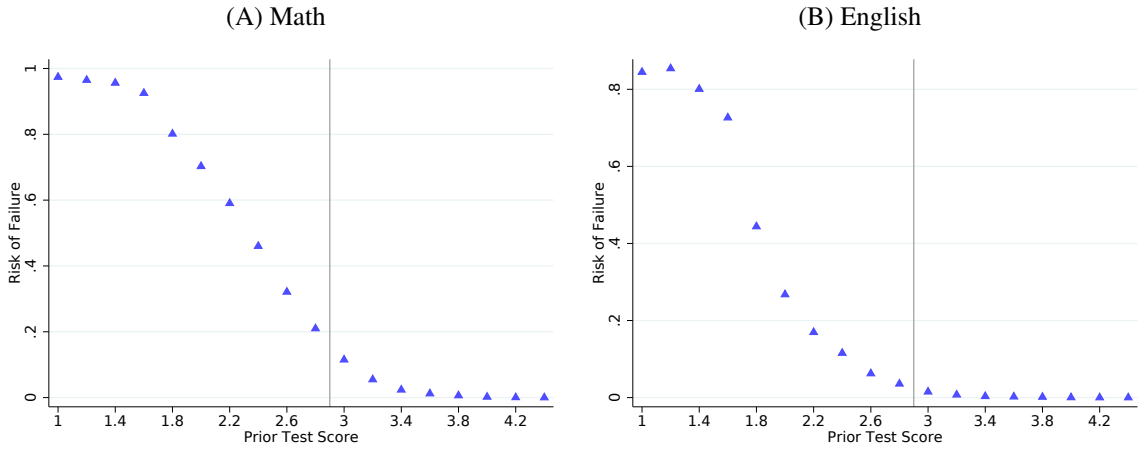
Notes: Panels A, B, C, and D plot the percentage of students who took the exam separately by subject and exemption status for the retention policy. Panel E plots the percentage of exempt students conditional on having both current and prior test scores.

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



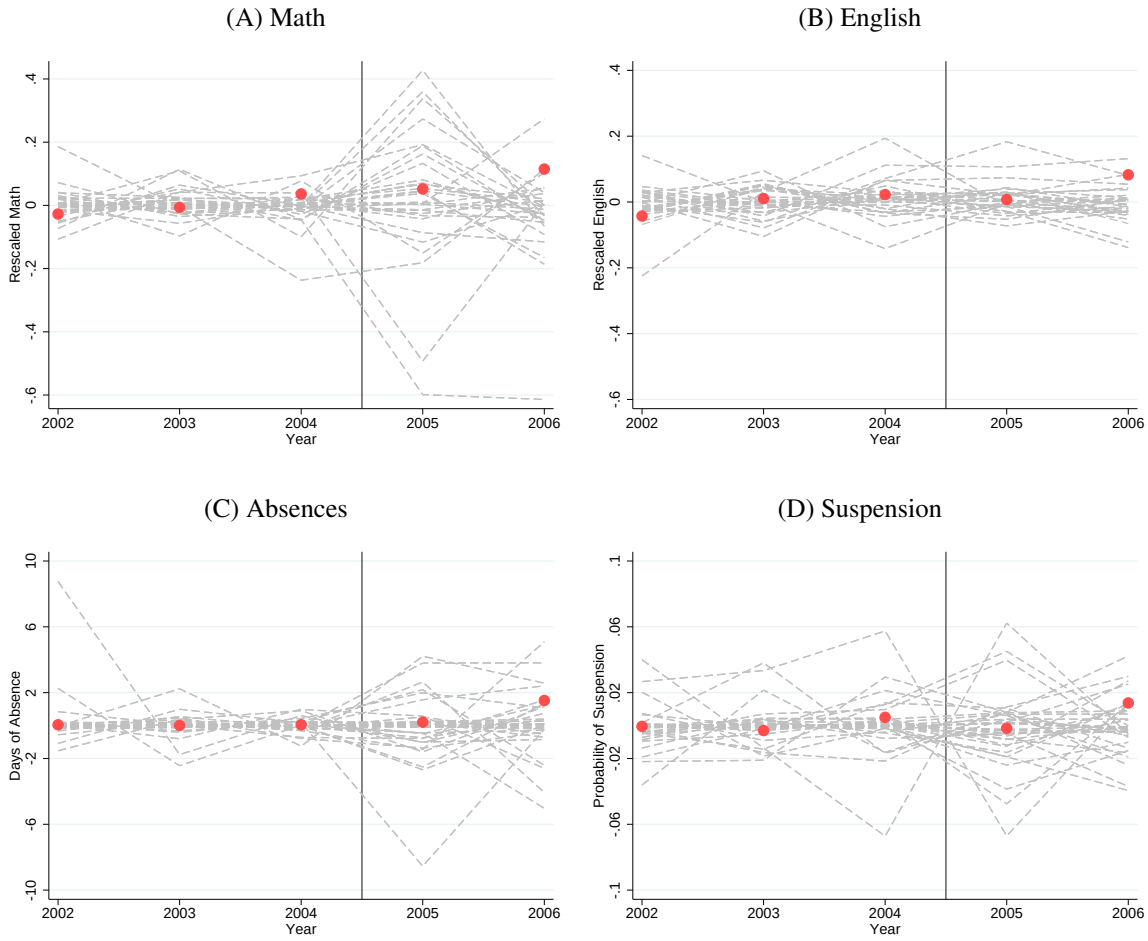
Notes: Each panel plots the kernel density of test scores separately for lowest-third and top-two-thirds students in each subject.

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.

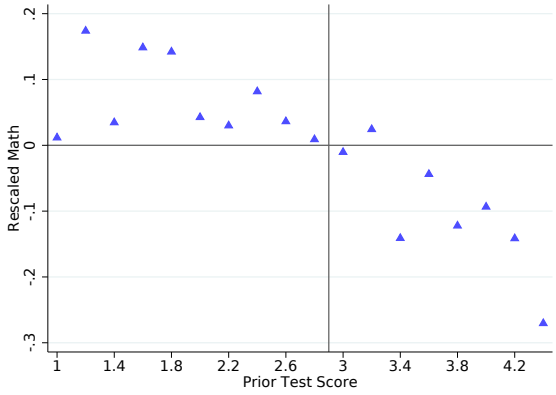
Figure A7: Effects of the Retention Policy (Synthetic Control): Placebo



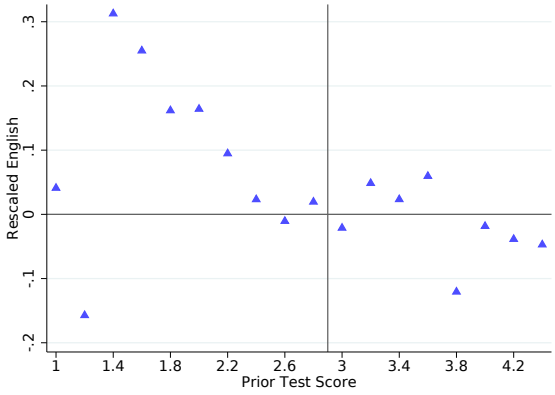
Notes: All panels are based on data from 2002 to 2006 and use grades not subject to the retention policy. The red line plots the difference between the treatment group and the synthetic control group in each year; the gray lines plot the difference between each member in the donor pool and its synthetic 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 A8: Distributional Effects of the Retention Policy

(A) Math

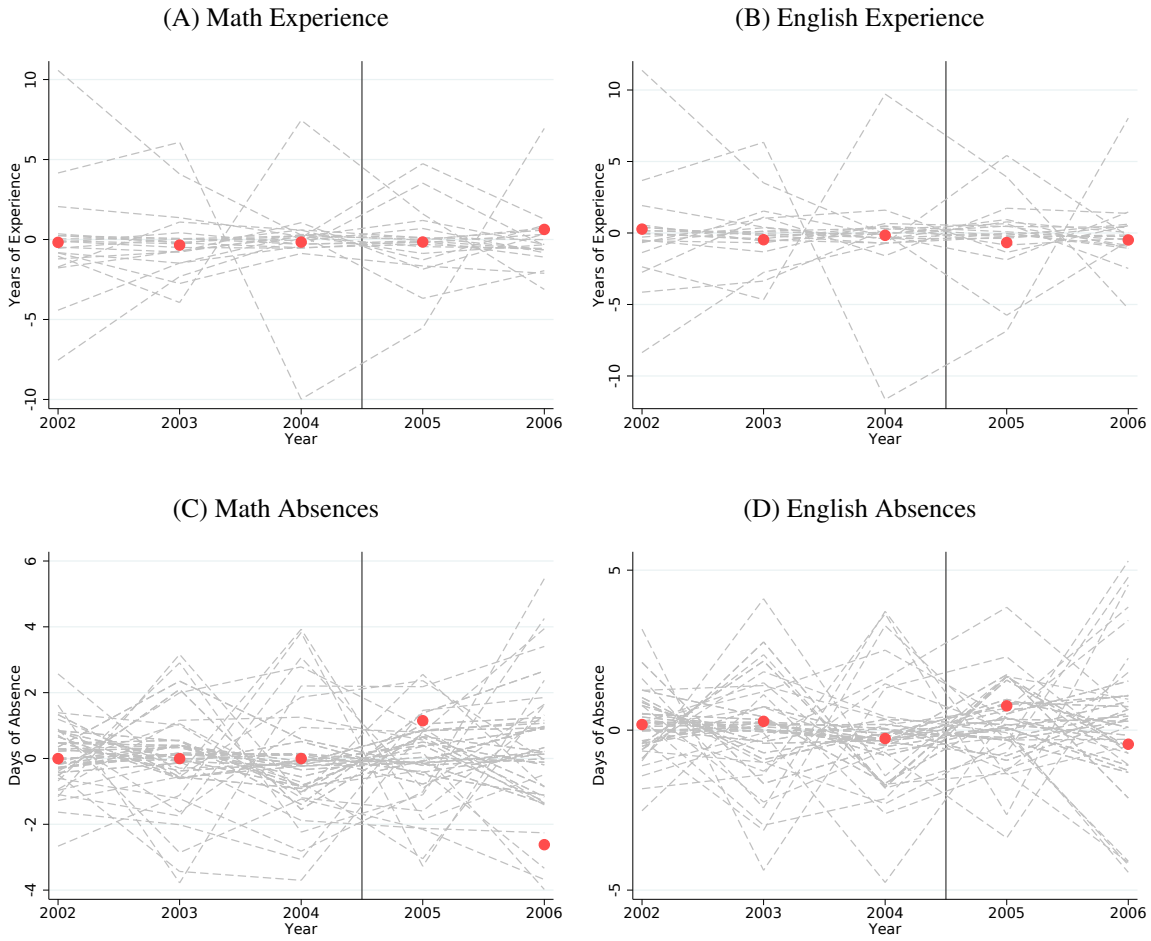


(B) English



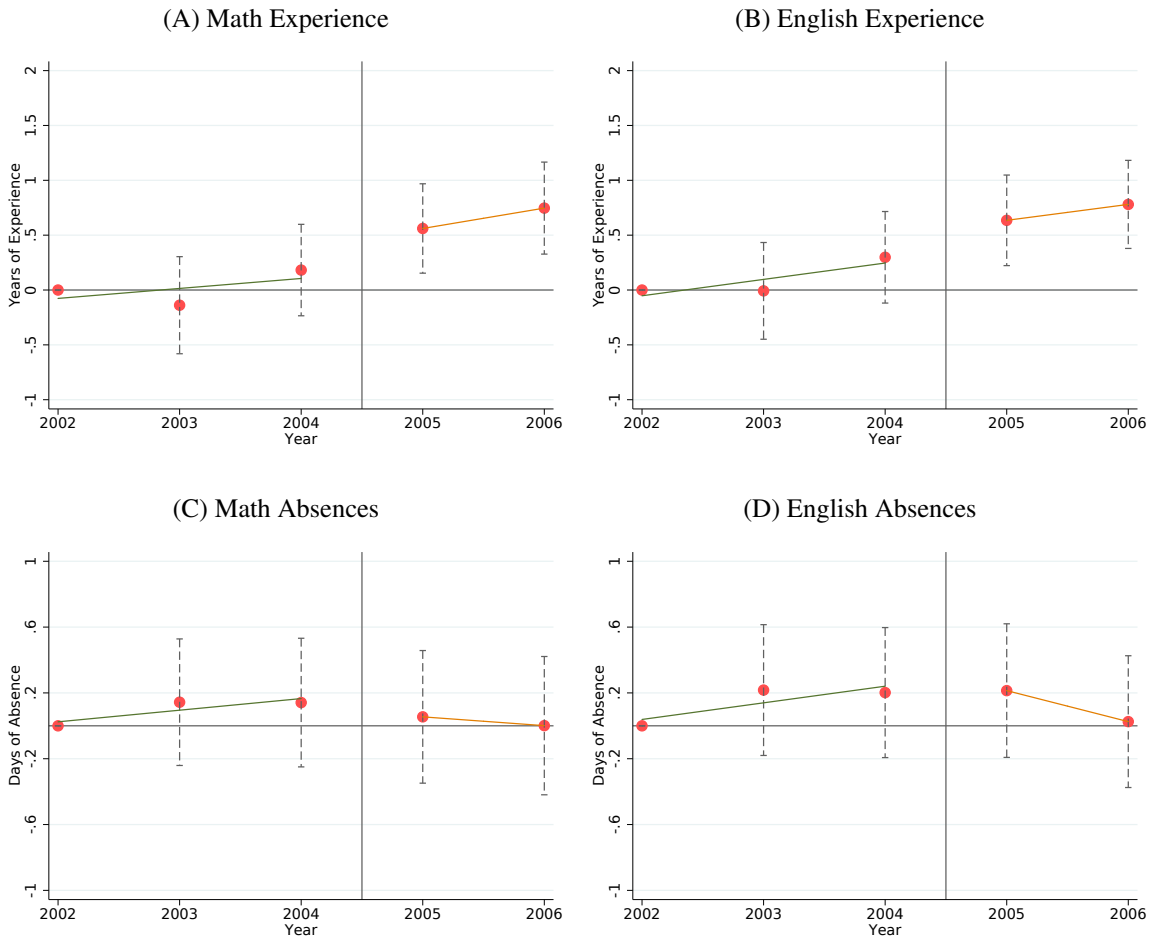
Notes: Both panels are based on students in the grade subject to the retention policy between 2003 and 2005 and divide prior test scores into bins of 0.2 points each. Each point represents a difference-in-difference estimate of the retention policy for each bin of students, using exempt students as a 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 A9: Effects of the Retention Policy on Teachers (Synthetic Control)



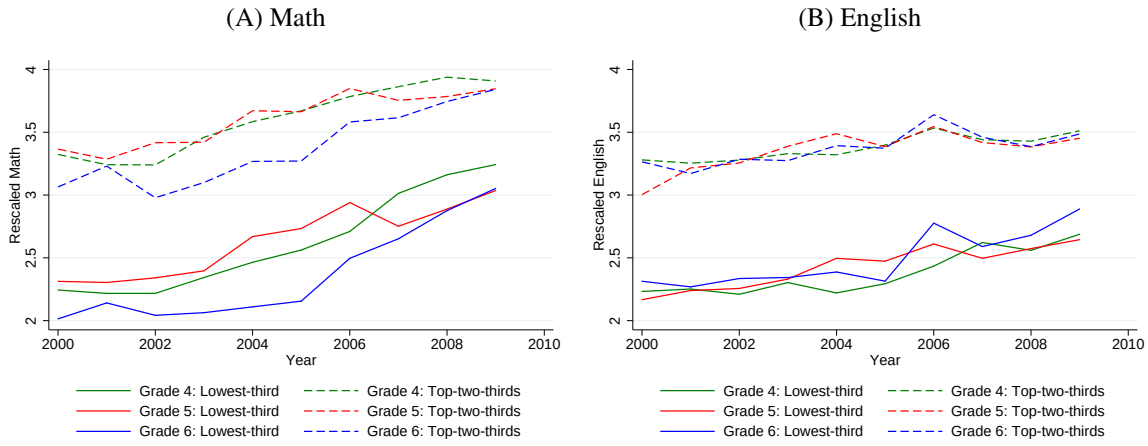
Notes: All panels are based on data from 2002 to 2006 and use the grade subject to the retention policy. The red line plots the difference between the treatment group and the synthetic control group in each year; the gray lines plot the difference between each member in the donor pool and its synthetic 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 A10: 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. 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.

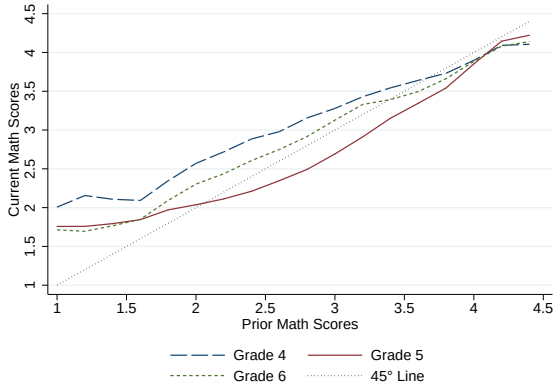
Figure A11: Trends in Prior Outcomes



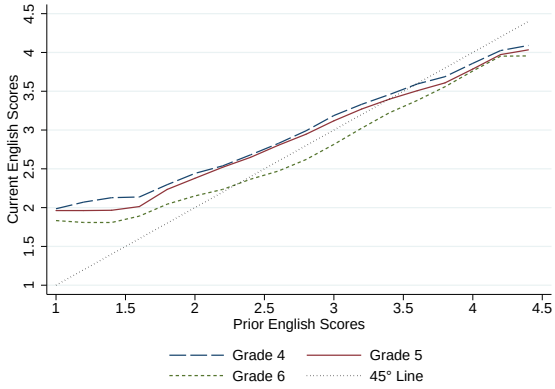
Notes: Both panels plot average prior outcomes in each year separately for lowest-third and top-two-thirds students.

Figure A12: Relationships between Current and Prior Outcomes

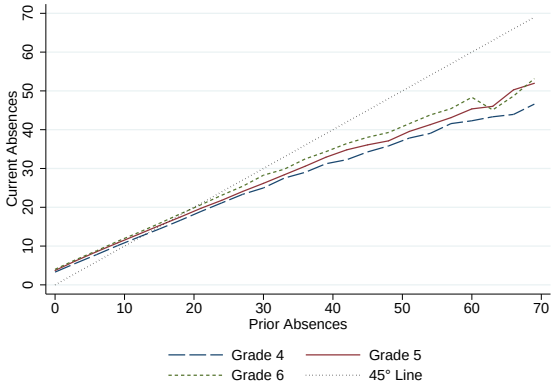
(A) Math



(B) English

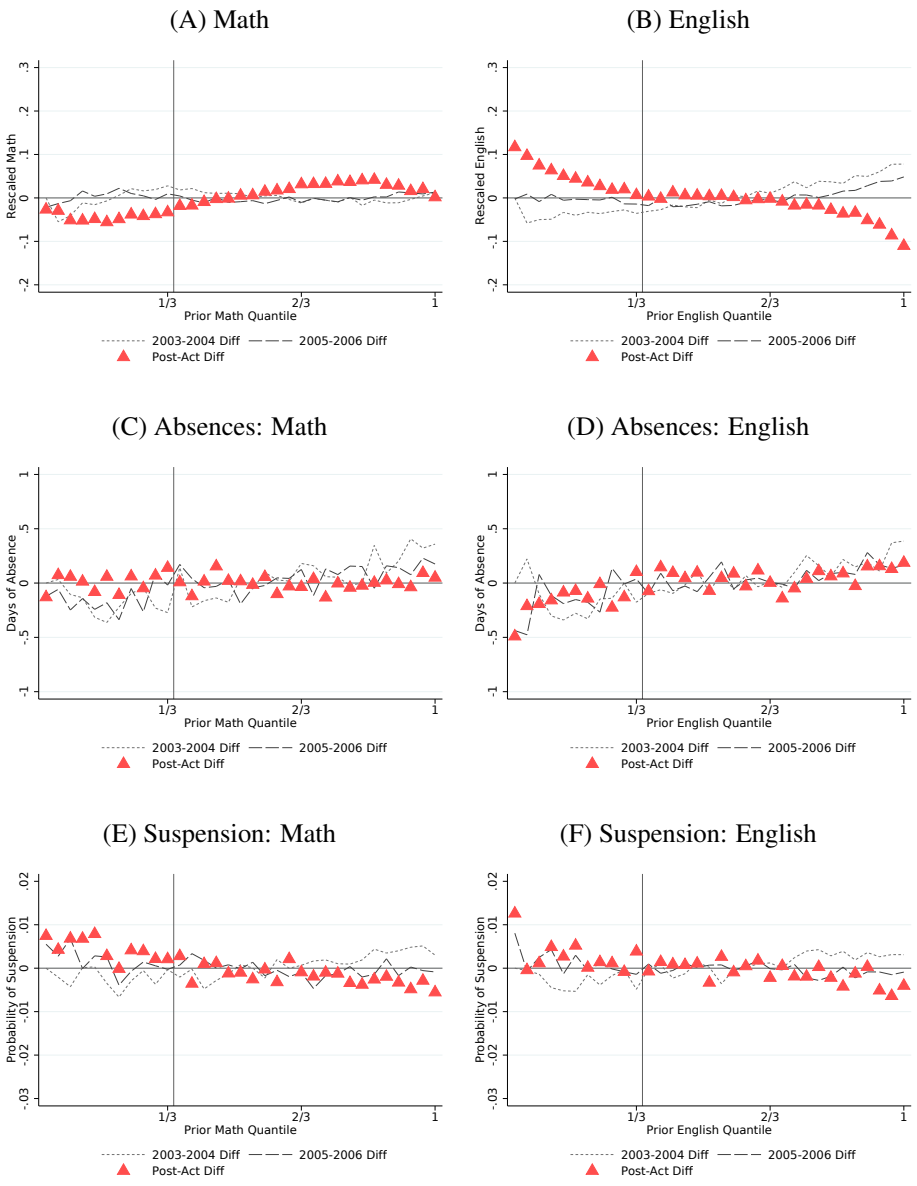


(C) Absences



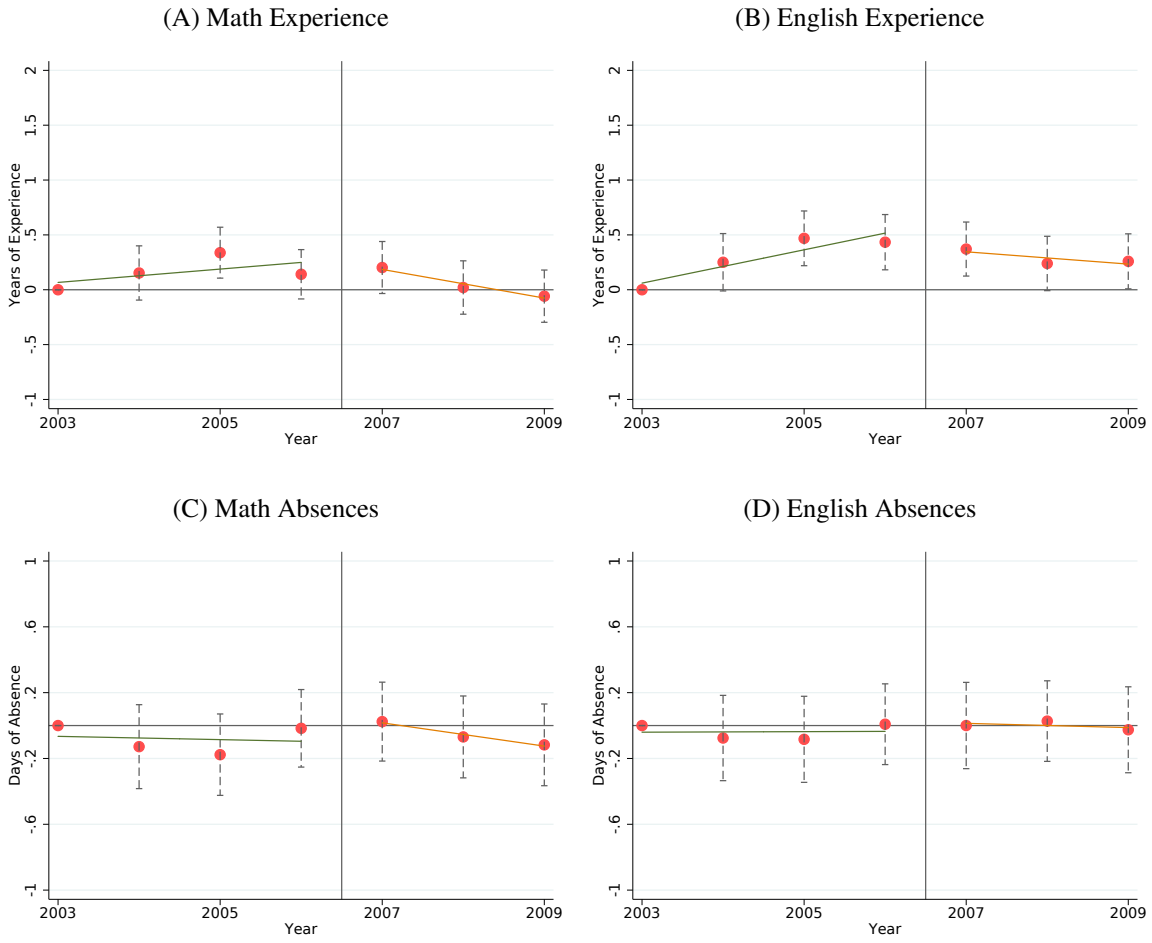
Notes: All panels are restricted to years when neither policies was in effect (prior to 2005).

Figure A13: Distributional Effects of the Accountability Scheme



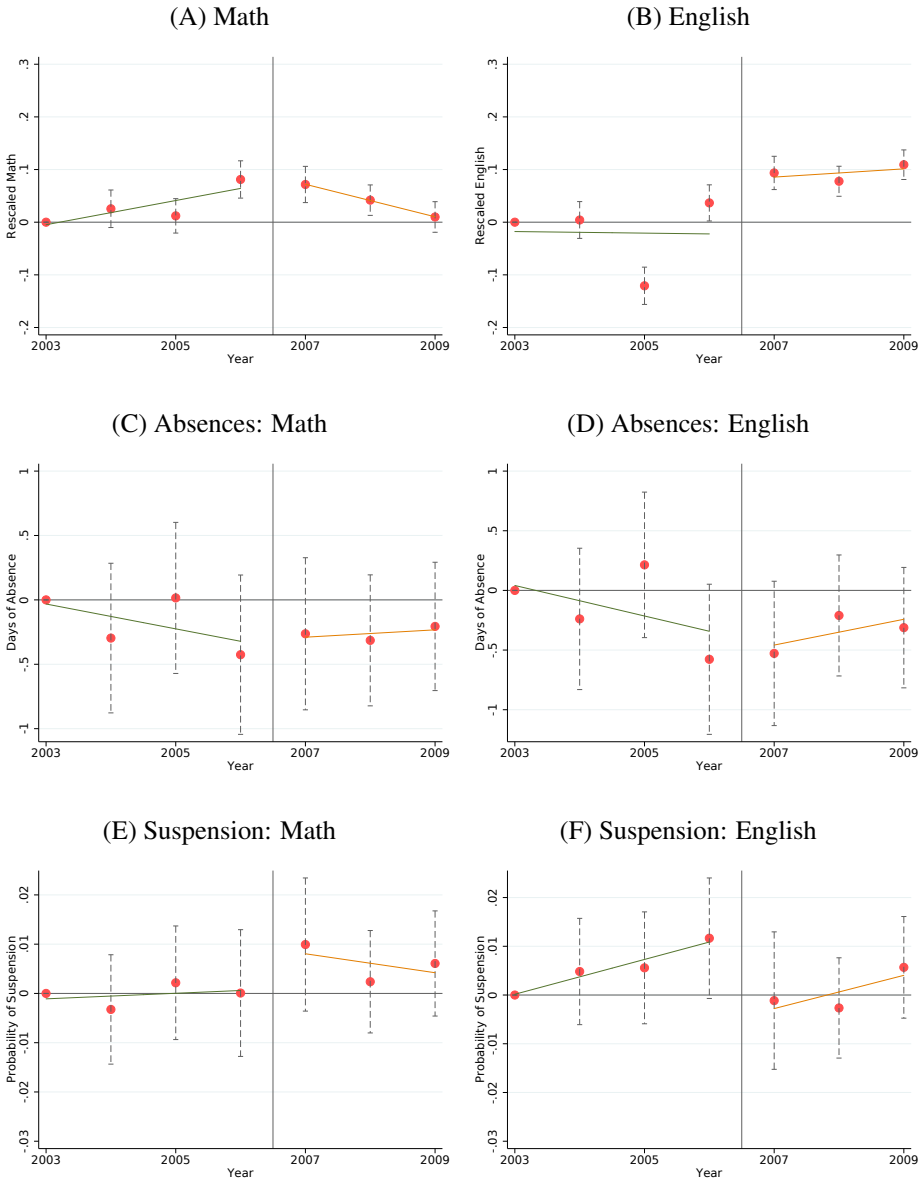
Notes: All panels are based on data from 2003 to 2009 and focus on students subject to the retention policy but in the grades not subject to the retention policy. This figure plots residuals obtained from regressing the outcomes on Equation 3. The x-axis in the left column uses prior math ranks; the 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 A14: 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 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.

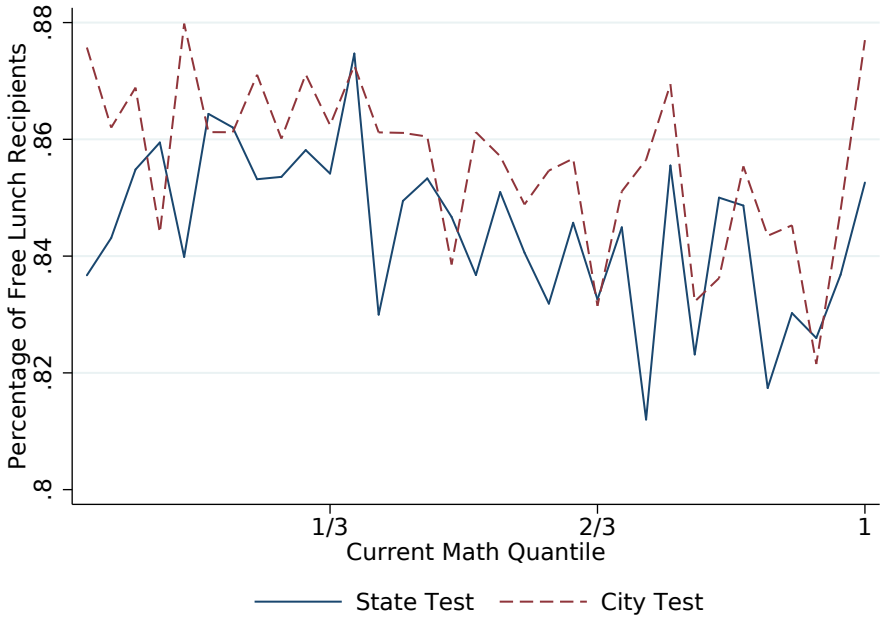
Figure A15: Effects of the Accountability Scheme: Special Ed/ELL



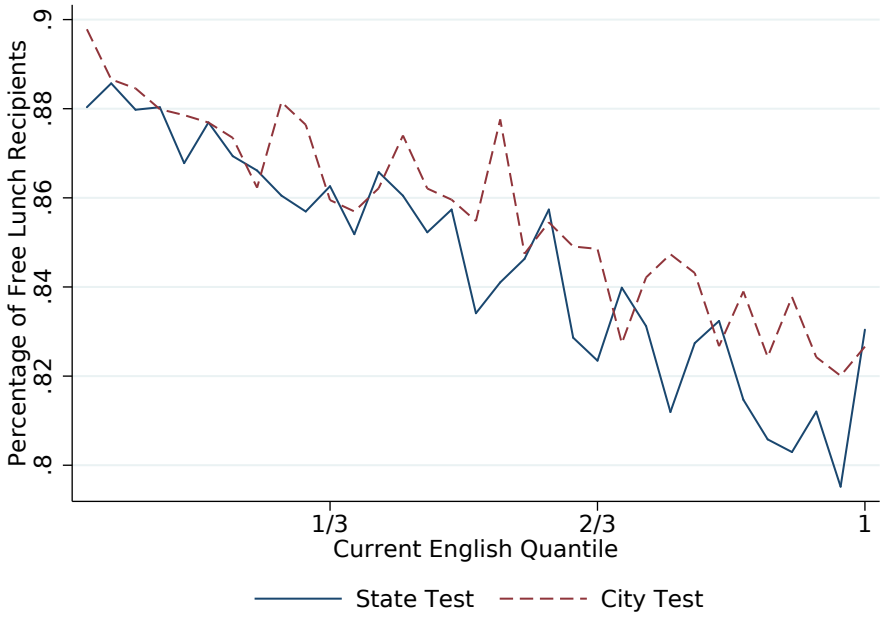
Notes: All panels are based on data from 2003 to 2009 and focus on special education/ELL students in 4th and 6th grades. This figure plots coefficients β_2 for each year from an event-study version of Equation 2. The dependent variables in Panels A and B are test scores in each subject; the dependent 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 A16: Distribution of Free Lunch Recipients: City vs. State Tests

(A) Math

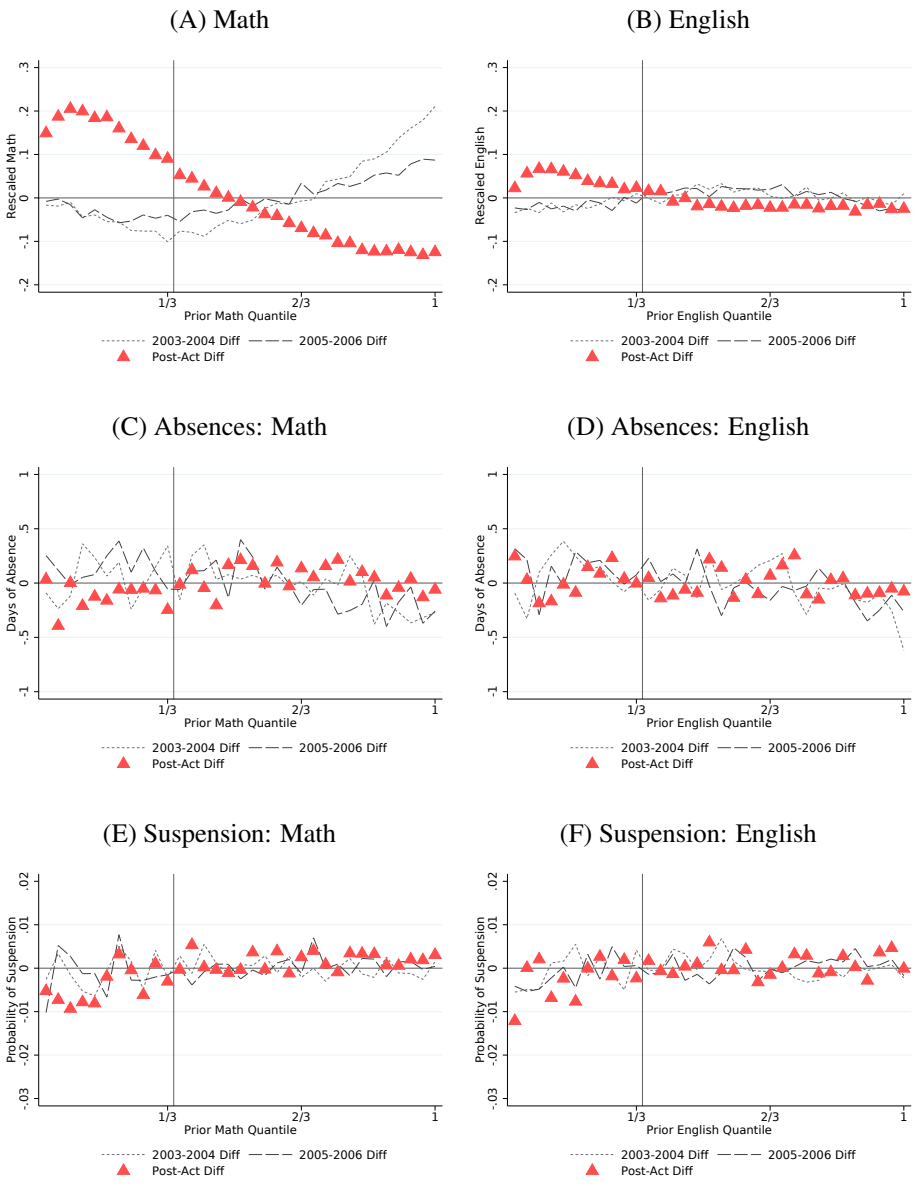


(B) English



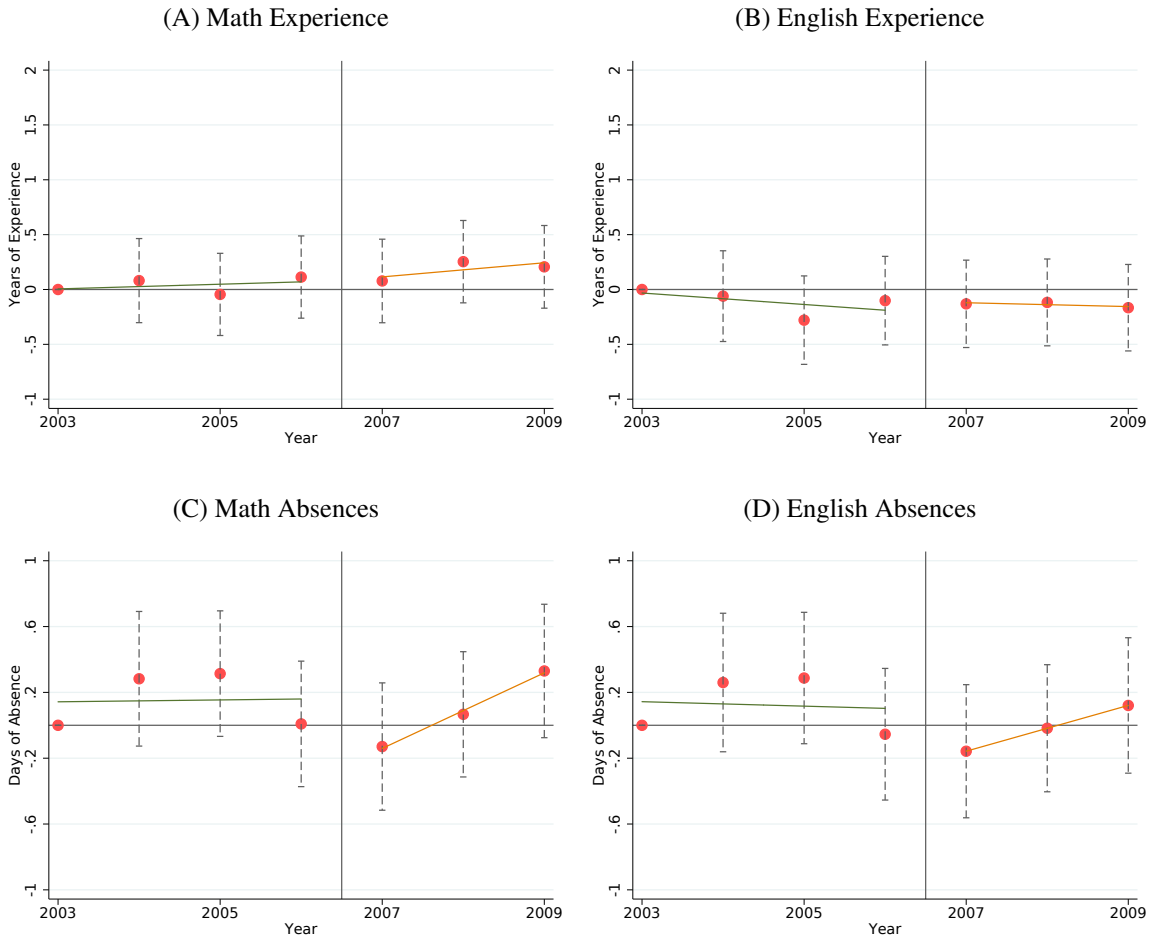
Notes: Both panels plot the percentage of free lunch recipients at each quantile immediately before (2005) and after adopting the state tests (2006) in grades 4 and 6. The solid line stands for the state tests (2006) and the dashed line stands for the city tests (2005).

Figure A17: 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 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 A18: 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 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.

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

Appendix C Probability of Retention

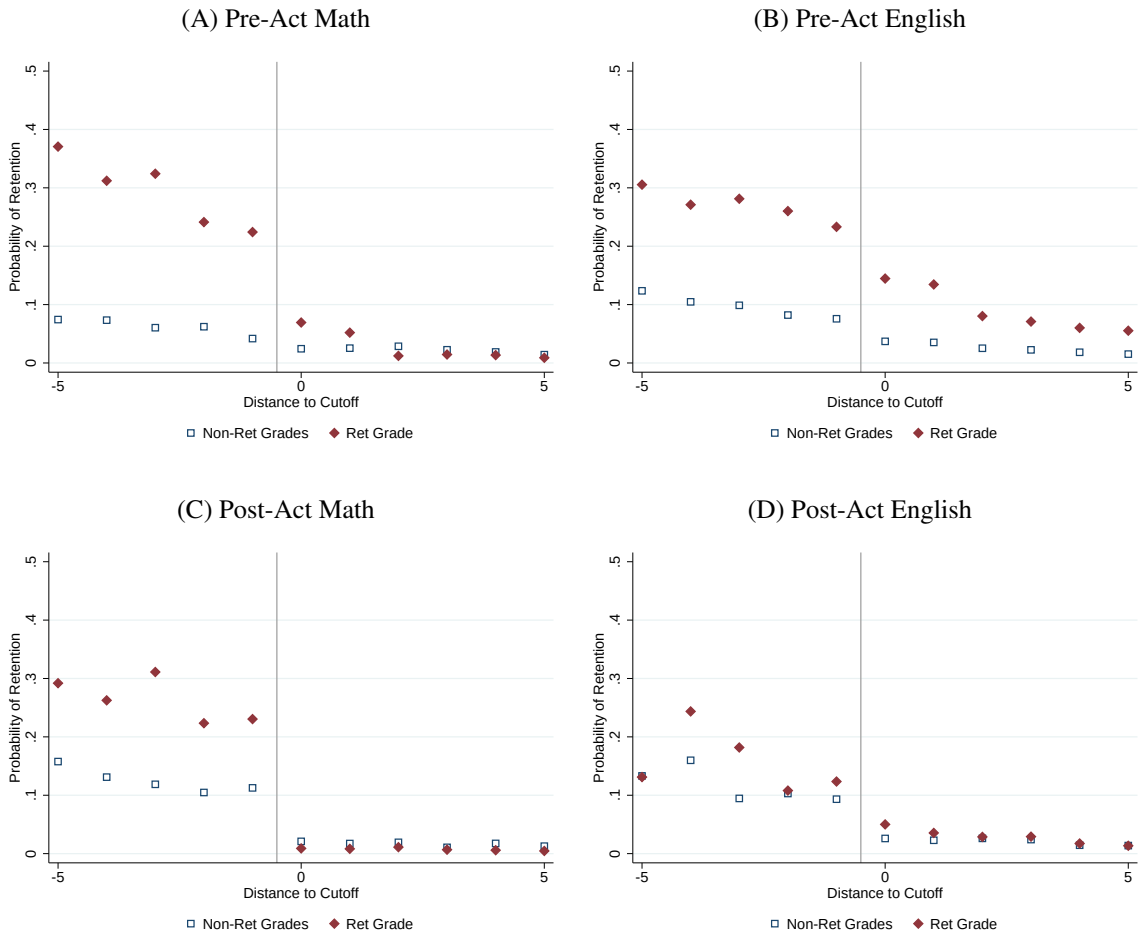
Figure 2 suggests that students' probability of retention may have changed after 2007, the year when the accountability scheme was implemented. Since retained students do not count toward the student progress scores in the accountability scheme, this change may be due to the accountability scheme.

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 A19 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.[44] 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.

[44]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.

Figure A19: 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.

Appendix D 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.

D.1 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:

$$\underset{s}{max}[\alpha A - C(s,t)] \tag{5}$$

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, $\alpha F_s(s,\beta) - C_s(s,\beta) = 0$.

D.2 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:

$$\Delta A(\Delta\alpha, \beta_0) \approx F_s(s(\alpha_0, \beta_0), \beta_0)\frac{ds}{d\alpha}\Delta\alpha \tag{6}$$

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.

D.3 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:

$$\Delta A(\alpha_0, \Delta\beta) \approx \underbrace{F_t(s(\alpha_0, \beta_0), \beta_0)\frac{dt}{d\beta}\Delta\beta}_{\text{Direct Effect}} + \underbrace{F_s(s(\alpha_0, \beta_0), \beta_0)\frac{ds}{d\beta}\Delta\beta}_{\text{Behavioral Response}} \tag{7}$$

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:

$$\frac{ds}{d\beta} = -\frac{\alpha F_{st} - C_{st}}{\alpha F_{ss}} \tag{8}$$

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.

D.4 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:

$$\Delta A(\Delta\alpha, \Delta\beta) \approx \underbrace{F_s(s(\alpha_0, \beta_0), \beta_0)\frac{ds}{d\alpha}\Delta\alpha}_{\text{Direct Effect from } \Delta\alpha} + \underbrace{F_t(s(\alpha_0, \beta_0), \beta_0)\frac{dt}{d\beta}\Delta\beta}_{\text{Direct Effect from } \Delta\beta} + \underbrace{F_s(s(\alpha_0, \beta_0), \beta_0)\frac{ds}{d\beta}\Delta\beta}_{\text{Behavioral Response to } \Delta\beta} \tag{9}$$

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:

$$\frac{\partial(ds/d\alpha)}{\partial\beta} = \frac{-\alpha F_{st} f_{ss} + \alpha F_s F_{sst}}{(\alpha F_{ss})^2} \tag{10}$$

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:

$$\frac{\partial(ds/d\beta)}{\partial\alpha} = \frac{\partial s}{\partial\alpha} \times \frac{-F_{ss}(F_{sts} - C_{sst}/\alpha + C_{st}/\alpha^2) - F_{sss}(F_{st} - C_{st}/\alpha)}{(F_{ss})^2} \tag{11}$$

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.[45] 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.

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