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

Tong Geng
Columbia University

Jonah E. Rockoff
Columbia Business School

Feb 2017*

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

*Geng: tg2430@columbia.edu Rockoff: jonah.rockoff@columbia.edu We would like to thank Marco Manacorda, Miikka Rokkanen, Marty West, Miguel Urquiola, and seminar participants at Columbia University, the Federal Reserve Bank of Cleveland, and Universidad Torcuato di Tella, for helpful comments and suggestions. This project was funded in part by the Smith Richardson Foundation (Grant #20088571).

1 Introduction

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

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

Grade retention is highly controversial, with critics arguing that it imposes negative academic and psychological effects on low performing students (Anderson et al. (2005)) and advocates contending that the policy can be academically helpful and psychologically en-

[1] Of course, public school systems may also have to deal with students whose knowledge or preparation far exceeds that of their age cohort. To the best of our knowledge, there is little, if any, research in economics on promoting students ahead of their cohort. Some research on "Gifted and Talented" programs finds that the marginal students admitted to the program did not see improvements in achievement (Bui et al. (2014)). The practice of "redshirting," i.e., holding children out of school for an additional year before they start kindergarten (see Deming and Dynarski (2008)), is also similar in many ways to retention, but is beyond the scope of this paper. Similarly, we do not address the larger literature on the effects of tracking students by age or ability, e.g., Duflo et al. (2011).

couraging (Wu et al. (2010)).[2] Addressing this controversy with empirical research is also difficult, as it necessitates understanding the likely counterfactual experience of retained students who almost certainly are experiencing severe difficulties in school. For this reason, a number of researchers have turned to the use of regression discontinuity, comparing the outcomes of students who just fail or just pass high-stakes academic assessments that determine grade retention (e.g., Jacob and Lefgren (2004, 2009), Manacorda (2012), Mariano and Martorell (2013), Özek (2015), Schwerdt et al. (2015)). These studies conclude that retention leads to increased academic achievement in the short-run, particularly for students held back in elementary school, but also find evidence of short-run increases in disciplinary incidents and long-run decreases in educational attainment, particularly for students held back at later grade levels.[3]

An additional hurdle in evaluating the effects of grade retention is that many outcomes are not easily comparable between retained and promoted students. For instance, students who are retained typically take different exams than those who are promoted, making it difficult to compare their relative academic performance.[4] Examinations of longer-run outcomes (e.g.,

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

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

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

school completion or wages) avoid this type of measurement problem, but face other issues related to interpretation.[5]

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

We find robust evidence that overall parental satisfaction with school quality rises significantly in the three years after students are retained. We do not see similar impacts for students' overall satisfaction with school quality, but we do find significant positive effects on

[5] For example, suppose that students (randomly) retained had higher years of schooling but lower grades completed than those who were (randomly) promoted. Whether this represents a positive or negative net effect on human capital is unclear. Similarly, if one found that students (randomly) retained had slightly lower earnings as those that were (randomly) promoted early in their work careers, does this mean retention (which entails greater educational costs) is not a cost-effective policy? If returns to experience and job-tenure are concave (e.g., Topel 1991), then the early-career and lifetime earnings gaps may have opposite signs, given that promoted students, who finished school a year earlier, are likely to have one more year of early career labor market experience.

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

students' feelings of personal safety in school in the years following retention. We also examine more conventional outcomes such as test scores, student absences, student suspension, and receiving special education. As in prior research, we find retention has large positive effects on students' test scores relative to their (younger) same-grade peers. We see little impacts on student absences and student suspension but we discern a positive effect on the likelihood of receiving special education. Overall, our results indicate that grade retention has positive impacts on the educational experience of students who comply with the test-based policies in place in New York City, as indicated by their parents' opinions. Whether parental opinion is a good barometer of educational quality is beyond the scope of this paper. Yet we would note that society relies on parents to make myriad decisions related to their children's education, and these views are therefore important to examine.

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

2 Data Description and Policy Background

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

The second database includes responses of parents and students to survey questions collected by the NYCDOE between 2007 and 2010.[7] Starting in the spring of 2007, the NYCDOE has distributed annual surveys to all students from grade 6 to 12 and all parents in public schools. Survey results count for 10-15 percent of a school's score in its annual Progress Report, the main school accountability tool used by NYCDOE (see Rockoff and Turner 2010 for details). The surveys have roughly 20 questions, are translated into nine languages and, of great importance for our study, ask the same questions of all parents and students regardless of grade level.[8] In our sample around 80 percent of students and 50 percent of parents responded. Survey questions differed slightly between years but the vast majority of questions remained the same throughout and response rates were relatively high compared to other school surveys (Nathanson et al. 2013).

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

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

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

grade starting in 2005, and seventh grade starting in 2006.[9]

Figure 1 describes the chronological order of testing, survey administration, and the steps in the process leading to retention decisions. Tests and survey administration are completed by April, but test results are not typically reported until close to the end of the school year. If a student fails either test, the school principal and teachers review the student's academic portfolio and decide whether to promote the student or require the student to attend summer school. At the end of summer school, students are given another opportunity to pass the tests and, after another review, final retention/promotion decisions are made. We do not have any information on portfolio review, summer school, or make-up testing.

We focus on students tested in school years 2004-2005 through 2007-2008, as we can measure grade retention and link subsequent survey responses for these students. Descriptive statistics are shown in Table 1 for our sample, which excludes observations when students would not have been subject to the test-based retention policy due to receipt of special education services or classification as an English Language Learner (ELL). Before focusing on the set of students scoring close to the cutoffs, we present some descriptive statistics on the broader sample (Table 1 Column 1). This sample's average test scores are around 0.2 standard deviations higher than the district mean due to our dropping ELL and special education students exempt from test-based retention policies. Six percent of the students in the sample failed their English exam, eight percent failed their math exam, two percent were retained the year after taking the test, and 5.7% of the student sample was retained at least once.[10] Among our sample of students subject to the retention policy in a particular year, 5.5% are later exempt due to a future change in their ELL or special education status; in

[9] Eighth grade was also subject to the new policy starting in 2009, but we do not examine these tests due to a data limitation. Our administrative records on student enrollment end in grade eight, we cannot observe promoted 8th graders who do not respond to surveys in 9th grade, whereas we observe all retained 8th graders regardless of survey response. We therefore do not examine retention in 8th grade in our main tables, although our results are robust to their inclusion (see Appendix Tables A.1, A.2, and A.3).

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

order to prevent potential bias due to endogenous selection into these categories, we examine outcomes in future years regardless of this future classification. Last, but not least, Table 1 shows that the vast majority of students in our NYC sample are poor, as indicated by receipt of free or reduced price lunch (86 percent), that these students are absent an average of 12 days during the year, and that three percent were suspended from school for misbehavior at least once during the year.

Our regression discontinuity design is based on student-year observations with English and/or math test scores close to the cutoff. We show summary statistics separately for students whose (lowest) score is just below the cutoff (Column 2) from those whose scores are exactly at or just above the cutoff (Column 3).[11] On average, 13 percent of students scoring just below a cutoff are retained, compared with around 0.01% for students scoring just above the cutoff. Thus, failing an exam seems to be a necessary, but clearly not sufficient, condition for retention. Compared to the entire sample, it is not surprising that students scoring near the cutoff are far more likely to be from poor households, from disadvantaged minority groups, have higher absences from school, and are more likely to have been suspended. There are also much smaller but still statistically significant differences in observables between those who score just above versus those who score just below the cutoff. Students just above the cutoff are less likely to receive free/reduced lunch (93% vs. 95%), had fewer absences (14.7 vs. 17) and were less likely to be suspended from school (4.0% vs. 5.4%). In Columns 4 and 5 of Table 1, we further split the students scoring just below a cutoff by whether or not they were actually retained. Again, there are small but significant differences, showing that retained students are clearly not a random subset among the students who barely fail these exams. Those who are eventually promoted have roughly 0.40 standard deviations higher average test scores, fewer absences (16 vs. 20), lower suspension rates (5.0% vs 7.0%), and are slightly less likely to receive free/reduced lunch (94% vs. 96%).

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

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

The lower rows of Table 1 provide information on the values of our survey measures *in the year that students were tested*. Compared to the sample as a whole, students with scores near the cutoff have considerably worse survey outcomes and, even within the sample near the cutoff, students who passed both exams have better parental survey outcomes (about 0.1 standard deviations) and better sense of personal safety (0.07 standard deviations). We also see consistent differences if we compare students who were promoted vs. those who were retained among students who failed at least one of their exams. Thus, it is again evident that there is positive selection of students for promotion among those who fail the exams, reinforcing the need for a credible identification strategy that addresses potential selection on unobservables.

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

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

3 Empirical Strategy

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

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

In our data we essentially have 25 quasi-experiments — five test years and five tested grades — and we combine them for our analyses. As noted before, some grades/years are affected by a more intensive retention policy regime, and the exact cutoff score for failing

[14] Appendix Figure A.1 plots the discontinuity of retention at the cutoff in each grade-year cell. The post-policy retention rates are much larger than the pre-policy retention rates. This pattern supports our empirical strategy.

the tests varies by grade and year. To accommodate these factors, we allow each test grade in each year to have its own control function but impose the same retention jump at the cutoff within each policy regime and a single Local Average Treatment Effect (LATE). Later we will explore allowing the LATE to differ by policy regime.

We use two-stage least squares for estimation:

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

$$Y_{i,t,l} = \sigma * \widehat{r_{i,t}} + G_{t,g}(index_{i,t}) + FE + \eta_{i,t,l} \tag{2}$$

Each observation is represented by student i, test year t, and the number of years l ("lag") between when the test was taken and when the outcome $Y_{i,t,l}$ is measured. $r_{i,t}$ is an indicator of retention in year t ($\widehat{r_{i,t}}$ is the predicted value from equation 1) and does not vary between lags; $policy_{t,g}$ indicates whether individual i is enrolled at time t in a grade g affected by the new policy regime; and $G_{t,g}(index_{i,t})$ is a grade-year specific cubic function of index. We include grade × year fixed effects to account for different cutoffs between years and grades and also outcome grade fixed effects to separate out grade effects on survey responses. Our outcomes $Y_{i,t,l}$ include normalized test scores, absences, suspension from school, students' overall satisfaction, personal sense of safety, and perception of the environment, and parents' overall satisfaction, and perception of school safety within three years after test year t, i.e., $l \in (0, 1, 2, 3)$.

We stack each observation in the administrative datasets up to four times to match with both current outcomes and future outcomes within three years following the initial retention decision. Appendix Table A.4 provides an illustrative example. Given that we observe the same student multiple times in the data, we cluster standard errors at student level. [15] We choose the index range of [-5,5] as our main bandwidth and check other bandwidths for robustness. We also present an analysis that includes 8th grade tests and the spring 2009 tests as a robustness check.

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

To support the validity of a regression discontinuity design (RDD), it is important that scores are not manipulated around the cutoff. There is little reason to believe such manipulation takes places, as the math and English tests are developed and graded externally to the school district, and Figure 3 shows that the density of observations at each index runs smoothly across the cutoff.[17] Further evidence that there is no manipulation is provided in Figure 4, which shows that the percentage of female students and students who receive free/reduced price lunch are also smooth through the cutoff. Appendix B provides additional continuity graphs and regression analyses of other covariates, including attrition rates and

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

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

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

survey response rates.

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

4 Main Results

4.1 First Stage Results

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

4.2 The Effects of Retention on Non-Survey Outcomes

The main contribution of our paper is to examine how retention affects subjective measures such as parental satisfaction about educational quality. However, in order to provide comparisons with earlier literature and some context for interpreting the survey evidence, we first present effects of retention on test scores, school absences, suspensions, subsequent grade repetition, and subsequent receipt of special education services. Graphical evidence is shown in Figures 5 and 6, which plot residuals from a regression of each outcome on grade × year fixed effects against our index variable, while Table 3 presents point estimates from regressions based on Equations 1 and 2.

Figure 5 shows outcomes in the year of the test ($l = 0$) and is therefore akin to a placebo, since retention decisions are made after these measures are taken. Consequently, there is no visual evidence of a significant jump at the cutoff, and the estimates in Row 1 of Table 3 confirm this conclusion. Figure 6 shows future outcomes, combining all data within three years after the test ($l \in (1, 2, 3)$) for simplicity. Figure 6a and 6b show that test scores relative to *same-grade* peers are dramatically higher for students who just fall below the cutoff on at least one exam; estimated effects of retention in Table 3 are 0.55 and 0.63 standard deviations for English and math, respectively.[18] These results are consistent with Jacob and Lefgren (2004), Schwerdt et al. (2015), and Mariano and Martorell (2013). Of course, retained students take different tests and, as discussed in the introduction, we cannot interpret these results as an improvement in academic achievement without further assumptions. Any effect on absences and suspensions over the following three years (Figures 6c and 6d) is difficult to discern graphically, and regression estimates in Table 3 suggest being retained has little impact on aggregate absences and some (marginally significant) effect on suspension over the subsequent three years.[19] Figure 6e shows that students just below the cutoff are more likely to receive special education within three years after the test;

[18] An important issue to consider is that failing an exam can have a discontinuous effect on educational experience outside of retention, such as having to attend summer school. If, for example, summer school leads to improved achievement, regardless of retention outcomes, then our interpretation of the two-stage least squares estimates may be incorrect. To shed some light on this issue, we present some admittedly suggestive evidence in Appendix Figure A.2, which takes average future test scores for students with index values below zero and plots them separately for retained and non-retained students. We can see that the future scores of non-retained students are quite continuous through the cutoff, while those of retained students are discontinuously higher. While retention within the set of students below the cutoff is obviously endogenous, we believe this graph is reassuring that our RD estimates are driven through the effects of retention, rather than other experiences related to having failed an exam.

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

Table 3 indicates an point estimate of 5.7% (more than twice the sample average after the tests). Although classification of special education follows some absolute standard, parents may see retention as a signal and react by seeking additional assistance through special education. This reaction seems more natural since special education exempts students from the retention policy in NYC. [20]

Students who barely pass exams and avoid retention in a given year may have significantly higher probabilities of failing and/or being retained in the future. The tendency for students who initially act as the "control group" to be given the "treatment" of grade retention at a later date can dampen our estimated effects of retention at time t for lags greater than 1. In Column 1 of Table 4, we present estimates of the effect of retention on grade level at one, two, and three years after the exam (l = 1, 2, 3). The (mechanical) coefficient of 1.00 at $l = 1$ fades slightly to 0.96 at $l = 2$ and slides further to 0.90 at $l = 3$, suggesting that 10 percent of students who would have been retained had they not barely passed their exams are still retained at some point within three years. This "fade-out" of the first-stage effects of failure on retention in NYCDOE is somewhat smaller than what Schwerdt et al. (2015) document in the state of Florida, where approximately 17 percent of the "marginally promoted" students are retained within three years and 25 percent retained within five years. Not surprisingly then, the effects of retention on academic performance relative to same grade peers is largest at $l = 1$ (0.66 in English, 0.79 in math), and declines through $l = 3$ (.36 in English, 0.39 in math). This pattern is also unsurprising given the wider literature documenting "fade-out" of the impacts of academic interventions on standardized test scores (e.g., Chetty et al. (2011), Cascio and Staiger (2012)), but it is notable that retention has substantial positive effects on test scores – relative to same-grade peers – several years later. [21]

[20] We suspect that parents may react to failing exams instead of retention and plot the average probability of receiving special education separately by whether the students below the cutoff are actually retained or not in Appendix Figure A.2. Although retention is endegenous, we are assured by this figure that our result is driven by actual retention.

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

4.3 The Effects of Retention on Survey Outcomes

In this section, we turn to our main outcomes of interest from parent and student surveys. As in the previous section, we provide graphical evidence first by regressing each survey outcome on grade × year fixed effects and outcome grade fixed effects and plot the average residual at each index around the cutoff. Before we present figures, it is worth emphasizing that only students above 6th grade respond to surveys and results on student surveys does not necessarily apply to students in elementary schools. Figure 7 shows results pooling surveys taken in the three years after the test.[22] Panel A of Figure 7 shows a clear jump in parental satisfaction at the cutoff; parents whose children barely passed the tests are less satisfied than parents whose children barely failed. It is also interesting to note that, while there is a weak positive relationship between satisfaction and index above the cutoff, the relationship below the cutoff is strongly negative, which mirrors the "first-stage" relationship of index with retention. We also see a smaller and slightly less clear jump at the cutoff in students' sense of safety, while students' overall satisfaction, students' views about the school environment, and parents' beliefs about school safety appear fairly continuous through the cutoff.

Before moving to our regression results, we provide two more pieces of graphical evidence, focusing on parental satisfaction and students' sense of safety. First, we plot results separately for tests in grade-year cells with and without the more stringent retention policy (Figure 8). For both outcomes, we see clear discontinuities in survey outcomes in the post-policy grade-year cells, with more positive survey responses among students who just failed one of their exams, but no noticeable change at the cutoff in the pre-policy years.[23] The fact that we see clearer patterns in the policy years may simply be due to the first stage being dramatically stronger when the policy was in place. However, given the large (and

[22] Appendix Figure A.3 shows there is no evidence of "placebo" effects for surveys taken in the year of the test; recall that the surveys are administered after the tests but prior to scores being known or retention decisions being made.

[23] In Appendix Figure A.4, we replicate these plots using outcomes in the same year as the test and show that, regardless of the policy in place, these outcomes are smooth through the cutoff.

not uncontroversial) increase in retention brought about by the policy change, the evidence of positive effects on parental satisfaction is interesting. Second, we plot average outcomes for students below the cutoff separately by whether or not the student was actually retained (Figure 9). Retention is clearly endogenous, as we cannot separate students to the right of the cutoff by whether or not they would have been retained if the cutoff were higher. However, this plot is reassuring, albeit only suggestive, as it shows that the differences across the threshold seen in Figure 7 are driven by relatively high parental satisfaction and sense of safety among retained students; the outcomes for non-retained students below the cutoff appear to match in a very continuous manner with outcomes for students above the cutoff.

Regression results in Table 5 are largely consistent with what we observe in the figures described above. Retained students' parents are estimated to be 5.4 points (or 0.3 standard deviations) happier than promoted students' parents in the three years following the retention decision. We also find that retained students feel 5.4 points (0.25 standard deviations) safer than promoted ones in the years following the retention decision, while the effects on parental views on school safety, students' views on school environment, and students' overall satisfaction are statistically insignificant.[24] The difference between students' personal sense of safety and their views on the school environment (which include measures of school safety) is instructive. Retained students feel that personally they are more safe, even though their general views of safety at the school level and other measures of school environmental quality are unchanged.[25]

Rather than pooling up to three years, it is interesting to ask whether the effects of

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

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

retention grow or decay over time. Table 6 presents separate estimates of the effects of retention after one, two, and three years, focusing on parental satisfaction and students' personal safety. We find that retained students' parents are slightly less satisfied (although not statistically significant) with their child's education in the year after retention, but significantly happier (roughly 35 percent of a standard deviation) two and three years after retention. The fact that we see an immediate improvement in academic performance (at least relative to same-grade peers) but a delayed effect on parental satisfaction is interesting. On one hand, it may be that satisfaction from academic improvement is wiped out by negative aspects of retention (e.g., stigma). On the other hand, since surveys are administered before test results are known each year, it may be that parents do not know how much their child has improved (at least relative to his/her new same-grade peers) one year after retention.

One might naturally wonder whether improvements in test scores relative to same-grade peers might possibly explain the positive effects on parental satisfaction that we find over three years. This is an important issue; if parents simply value their child's performance rank relative to same-grade peers, then retention may simply re-order students so that parents of those retained are happier but parents of low-achieving students in the younger cohort are made less happy. [26] A purely ordinal interpretation would essentially mean that retention is a zero-sum game.[27] To investigate this question a bit further, we ran cross sectional regressions of parental satisfaction on on students' test scores and find that a one standard deviation increase in both mathematics and ELA test scores increases parental satisfaction by about 1.3 points. This does not represent a causal estimate of the impact of test score performance on satisfaction, and there are several reasons to think such coefficients might be biased upward. Nevertheless, if we apply this coefficient, test scores would explain less than 25 percent of the effect of retention on parental satisfaction. We cannot rule out that improvements in performance relative to a new, younger peer group explains some of our

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

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

results, nor that parents do not care about ordinal rank, but it is unlikely that these factors are the main drivers of our findings.

The pattern of effects on students' personal sense of safety in Table 6 reveal a different pattern, with the largest effect in the first year after retention (about half a standard deviation) and positive but gradually declining effects over the next two years. One interpretation is that students feel much safer when enrolled alongside younger peers, and that this age advantage grows less important over time. We explore the explanatory power of a relative-age effect by running a cross-sectional regression of students' personal sense of safety on students' relative age for students who have never been retained; this yields a coefficient that implies that the oldest child in a class responds only about 0.75 points (0.04 standard deviations) higher on average than the youngest child to the questions about personal safety. Thus, the marginal students who are retained due to test failure would have to be much more sensitive than the typical student to their age position for this to explain the effects we find on personal safety.

As mentioned above, we are also interested in whether the effects of retention differ between the pre- and post-policy retention regimes. Estimated effects of retention on parental satisfaction and students' safety that are allowed to differ between policy regimes (Table 7) show that our main findings are driven by the grade-year cells in the new policy regime. The point estimates of the old policy regime for parental satisfaction and students' safety are both negative but statistically insignificant and very imprecisely estimated. This is not terribly surprising since the first stage power under the old policy regime is rather weak. However, similar estimation with test scores as the outcome (Appendix Table A.5) shows that the effects of retention on academic performance relative to same-grade peers is remarkably similar in the pre- and post-policy periods. If the effect of retention on parental satisfaction is simply due to academic improvement, we should see similar effects between the two policy regimes but we do not. [28]

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

4.4 Robustness Checks

In this subsection, we present four robustness checks that further support our main findings. First, we re-analyze the effects of retention on parental satisfaction and students' personal safety while widening our bandwidths in one point increments from [-4,4] through [-10,10] (see Appendix Table A.6). Our point estimates for effects on parental satisfaction are quite insensitive to bandwidth. Indeed, the only noticeable change is that the coefficient for a "placebo effect" of retention on students' current (i.e., pre-retention) sense of personal safety goes closer towards zero as the bandwidth widens, while the estimated effect on students' future sense of personal safety remain quite stable. Thus, we do not find any evidence that the choice of bandwidth is driving our results.

Second, we note that Lee and Card (2008) suggests RDDs should cluster errors at the running variable level to minimize specification errors and this practice has become somewhat standard. However, we see little reason to perform this clustering practice because our outcomes are not subject to any common shocks at the index level and Kolesár and Rothe (2016) suggests that this practice may do more hard than good. Nevertheless, we implement a two-way clustering at both the individual and index level in order to make sure that this does not have a major impact on our statistical inference. Reassuringly, we find that the two-way clustered standard errors are quite similar to clustering at the student level (see Appendix Tables A.7, A.8, and A.9).

Third, recall that we do not examine retention in grade 8 or in the school year 2009-2010 in our main results. We omit these observations because we are only able to observe the retention decision and future outcomes of students in 8th grade (and/or tested in school year 2009-2010) if they stayed in the New York City public school system in the next year and they (or their parents) responded to the surveys. In other words, we cannot distinguish between a student who was promoted to grade 9 and left NYC schools from a student who was promoted but did not respond to the NYC survey, nor can we measure retention for students who were tested in 2011 who did not respond to surveys. The addition of these observations to regressions of parental satisfaction and students' sense of personal safety (see Appendix Table A.1, A.2, and A.3) do not significantly alter our main findings.

Last, but not least, retention may induce students to trasfer to another school and, as we investigated before, to seek assistance through special education. We include the number of years a student has spent in a school, the type of the school, or an indicator of receiving special education as additional covariates in our estimation. These results are shown in Table A.10 and similar to our previous estimates. [29]

5 Regression Discontinuity Extrapolation

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

In addition to standard RDD assumptions, this technique requires a Conditional Independence Assumption (CIA) and Common Support (CS). CIA requires the potential outcomes to be mean-independent of the running variable after conditioning on other pre-determined covariates; CS requires treatment status to vary conditional on these covariates. Following Angrist and Rokkanen (2015), we test the CIA assumption by regressing our survey outcomes on predetermined covariates (e.g., two-year prior test scores), and then examining the relationship between residuals of this regression and our running variable on each side of the cutoff.[30] We focus on the grades and years under the new policy regime to maximize the power of first stage and explore LATEs on students' personal sense of safety and parental satisfaction.

Results for tests of the CIA assumption (Appendix Table A.11 and Appendix Figure A.5)

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

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

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

We indirectly test CS by checking the distribution of pre-determined covariates at each index score. Appendix Figure A.6 shows a box plot of two-year prior standardized mathematics scores at each index score. The extensive coverage at each index score supports CS.

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

$$\frac{E(Y_{1i} - Y_{0i}|x_i, r_i)}{E(W_{1i} - W_{0i}|x_i, r_i)} = \frac{E(Y_{1i}|x_i, r_i) - E(Y_{0i}|x_i, r_i)}{E(W_{1i}|x_i, r_i) - E(W_{0i}|x_i, r_i)} \tag{3}$$

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

$$\begin{aligned} E(y_i|x_i, r_i < 0) &= x_i'\beta_1 \\ E(y_i|x_i, r_i \geqq 0) &= x_i'\beta_0 \\ E(w_i|x_i, r_i < 0) &= x_i'\delta_1 \\ E(w_i|x_i, r_i \geqq 0) &= x_i'\delta_0 \end{aligned} \tag{4}$$

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

$$\frac{(\beta_1 - \beta_0)'E(x_i|r_i = c)}{(\delta_1 - \delta_0)'E(x_i|r_i = c)}. \tag{5}$$

Implicitly we assume the linear models for the conditional mean at each side of the cutoff are the same. In practice, we first use observations with $r_i < 0$ and regress y_i on x_i to estimate β_1 and, likewise, use observations with $r_i \geqq 0$ and regress y_i on x_i to estimate β_0. We apply an analogous procedure to estimate δ_0 and δ_1. Armed with these estimates and our predetermined covariates, we calculate the estimator based on Equation 5.[31] We compute the standard errors by bootstrapping non-parametrically with 500 replications. Our estimates, displayed in Figure 10, suggest that the impact of retention on parental satisfaction would be smallest (roughly 2 points) among students with scores well below the threshold, roughly constant (around 6 points, equivalent to our RDD estimate) in the range of index scores between -5 and +2, and then slope upward until reaching 11 points for students with an index score of 6. Our confidence intervals become quite wide for estimates farther away from the cutoff, but the evidence clearly suggests larger positive treatment effects on parental satisfaction for students who passed the exams by at least 3-4 index points. Of course, these effects apply only to students who would have been retained after the process of portfolio review, summer school, and re-testing, and our estimates of δ_0 and δ_1 suggest that a relatively small fraction of these students (16%) would have been retained. Interestingly, these results are not consistent with our prior beliefs, which were that the positive effects of retention would have been greatest among students scoring well below the cutoff; these "inframarginal" students have far higher retention rates which suggested to us that school officials and parents are more likely to agree that retention would be a beneficial educational intervention for the child.

6 Conclusion

We examine variation in grade retention stemming from policies in New York City public schools which create discontinuities in the relationship between retention probability and

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

test scores. Merging administrative data on student enrollment and testing with self-reports by students and parents about the quality of their educational experience, we contribute to the literature on the effects of retention by examining outcomes which, unlike test scores, can be easily compared across students in different grade levels. We find that students retained in NYC as a result of the district's more stringent test-based retention policy saw significant improvements in parental satisfaction with the quality of their child's education and students' personal sense of safety. We provide evidence that suggests these effects are driven by factors beyond attending summer school, changes in age relative to classmates, or changes in performance on high stakes tests relative to same-grade peers. However, there are many additional ways in which retention can alter a student's school experience, and we lack the data to examine these other various channels.

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

References

G. E. Anderson, A. D. Whipple, and S. R. Jimerson. Student ratings of stressful experiences at home and school: Loss of a parent and grade retention as superlative stressors. *Journal of Applied School Psychology*, 21(1):1–20, Jul 2005.

J. Angrist and M. Rokkanen. Wanna get away? regression discontinuity estimation of exam school effects away from the cutoff. *Journal of the American Statistical Association*, 110 (512):1331–1344, 2015.

S. A. Bui, S. G. Craig, and S. A. Imberman. Is gifted education a bright idea? assessing the impact of gifted and talented programs on students. *American Economic Journal: Economic Policy*, 6(3):30–62, 2014.

C. Campanile. Pols give school reform an f, Jun 22 2004. Copyright - (Copyright 2004, The New York Post. All Rights Reserved; Last updated - 2012-01-26.

D. Card, A. Mas, E. Moretti, and E. Saez. Inequality at work: The effect of peer salaries on job satisfaction. *The American Economic Review*, 102(6):2981–3003, 2012.

E. U. Cascio and D. O. Staiger. Knowledge, tests, and fadeout in educational interventions. Technical report, National Bureau of Economic Research, 2012.

R. Chetty, J. N. Friedman, N. Hilger, E. Saez, D. W. Schanzenbach, and D. Yagan. How does your kindergarten classroom affect your earnings? evidence from project star*. *Quarterly Journal of Economics*, 126(4), 2011.

A. Crego, D. Gershwin, G. S. Ikemoto, J. S. McCombs, V.-N. Le, S. N. Kirby, J. A. Marsh, L. T. Mariano, S. Naftel, C. M. Setodji, and N. Xia. Ending social promotion without leaving children behind: The case of new york city. Technical report, Santa Monica, CA: RAND Corporation, 2009.

T. S. Dee, W. Dobbie, B. A. Jacob, and J. Rockoff. The causes and consequences of test score manipulation: Evidence from the new york regents examinations. Technical report, National Bureau of Economic Research, 2016.

D. Deming and S. Dynarski. The lengthening of childhood. *Journal of Economic Perspectives*, 22(3):71–92, 2008.

M. Dobbs. Ready for fourth grade? not so fast, new york says; policy against social promotion draws fire from some groups, Jul 07 2004. Copyright - Copyright The Washington Post Company Jul 7, 2004; People - Bloomberg, Michael; Last updated - 2010-08-05.

E. Duflo, P. Dupas, and M. Kremer. Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in kenya. *The American Economic Review*, 101:1739–1774, 2011.

E. R. Eide and M. H. Showalter. The effect of grade retention on educational and labor market outcomes. *Economics of Education Review*, 20:563–576, 2001.

J. I. Garcia-Pereza, M. Hidalgo-Hidalgoa, and J. A. Robles-Zurita. Does grade retention affect students' achievement? some evidence from spain. *Applied Economics*, 46(12): 1373–92, 2014.

E. Gootman. Test policy for 3rd graders is met by more resistance, Feb 11 2004. Copyright - Copyright New York Times Company Feb 11, 2004; Document feature - photographs; People - Bloomberg, Michael; Last updated - 2010-06-29; CODEN - NYTIAO.

D. M. Herszenhorn. Stricter standards in new york may hold 15,000 in 3rd grade, Jan 09 2004. Copyright - Copyright New York Times Company Jan 9, 2004; Last updated - 2010-06-29; CODEN - NYTIAO.

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

B. A. Jacob and L. Lefgren. The effect of grade retention on high school completion. *American Economic Journal: Applied Economics*, 1(3):33–58, 2009.

P. Kline. Oaxaca-blinder as a reweighting estimator. *American Economic Review: Papers and Proceedings*, 101:532–537, May 2011.

M. Kolesár and C. Rothe. Inference in regression discontinuity designs with a discrete running variable. March 2016.

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

V. Lavy, M. D. Paserman, and A. Schlosser. Inside the black box of ability peer effects: Evidence from variation in the proportion of low achievers in the classroom. *The Economic Journal*, 122(559):208–237, 2012.

D. S. Lee and D. Card. Regression discontinuity inference with specification error. *Journal of Econometrics*, 142:655–674, 2008.

M. Manacorda. The cost of grade retention. *The Review of Economics and Statistics*, 94(2): 596–606, 2012.

L. T. Mariano and P. Martorell. The academic effects of summer instruction and retention in new york city. *Educational Evaluation and Policy Analysis*, 35(1):96–117, March 2013.

R. Murphy and F. Weinhardt. Top of the class: The importance of ordinal rank. 2014.

L. Nathanson, M. McCormick, and J. J. Kemple. Strengthening assessments of school climate: Lessons from the new york city school survey. Brief, Research Alliance for NYC Schools, June 2013.

J. Rockoff and C. Speroni. Reliability, consistency, and validity of the nyc doe environmental surveys: A preliminary analysis. Nov 2008.

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

G. Schwerdt, M. R. West, and M. A. Winters. The effects of test-based retention on student outcomes over time: Regression discontinuity evidence from florida. National Bureau of Economic Research working papers, Aug 2015.

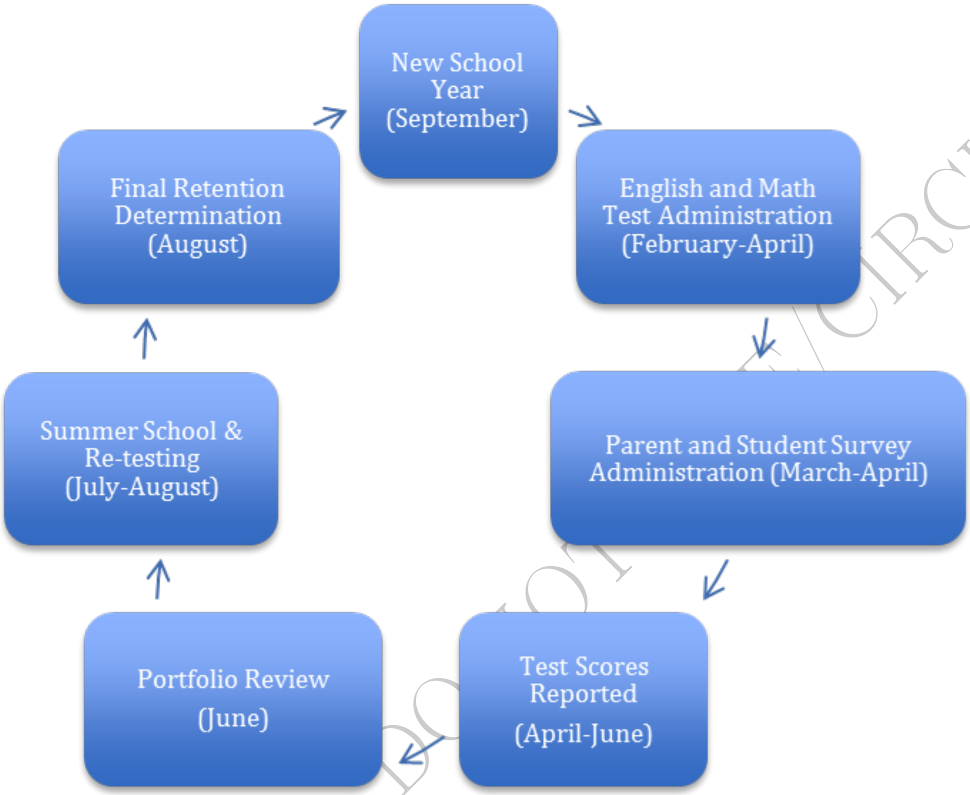
R. Topel. Specific capital, mobility, and wages: Wages rise with job seniority. *Journal of Political Economy*, pages 145–176, 1991.

W. Wu, S. G. West, and J. N. Hughes. Effect of grade retention in first grade on psychosocial outcomes. *Journal of Educational Psychology*, 102(1):135–152, Feb 2010.

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

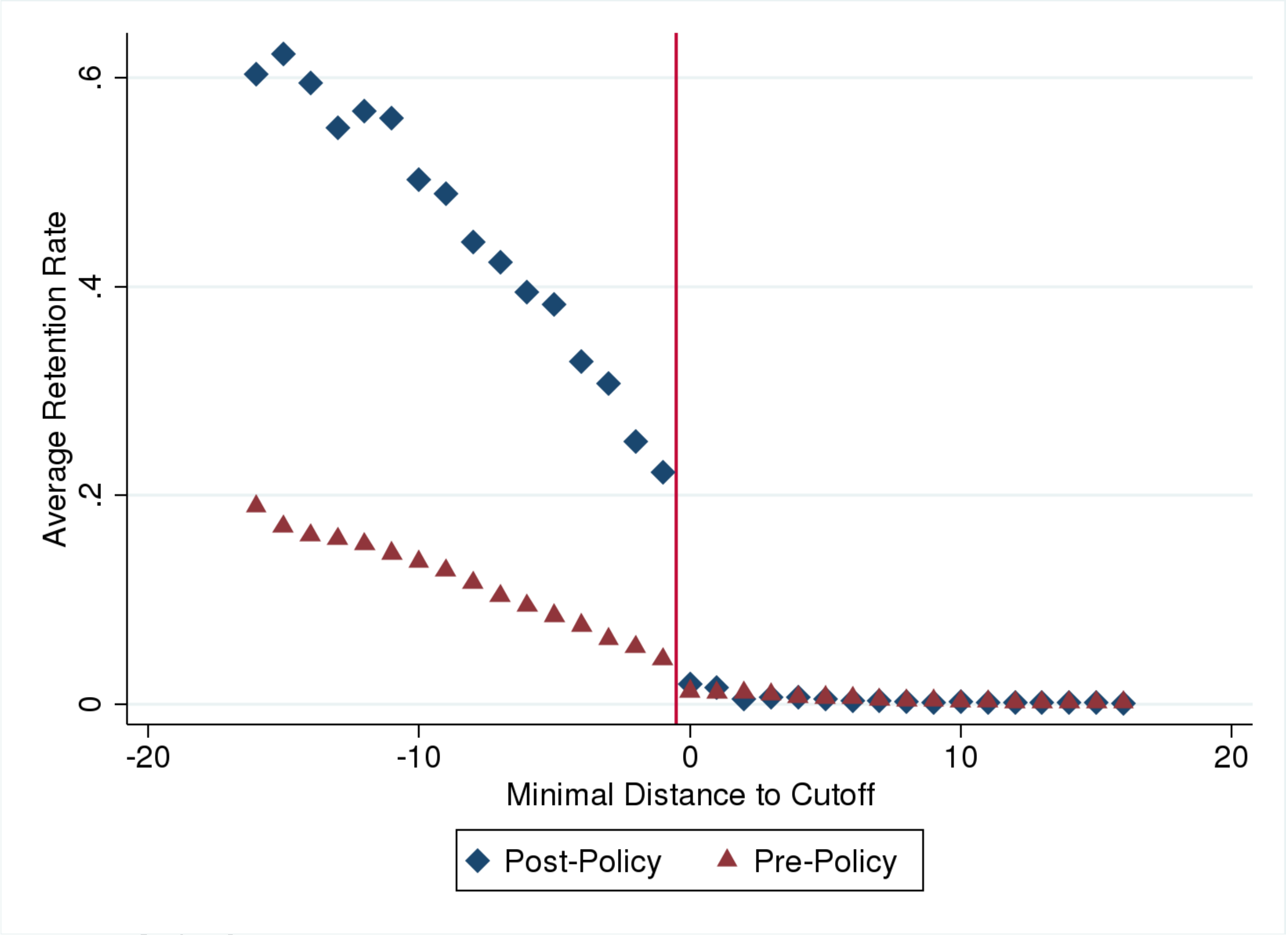
7 Figures

Figure 1: Timing and Process for Testing, Surveys, and Promotion Decisions


New School Year (September)
English and Math Test Administration (February-April)
Parent and Student Survey Administration (March-April)
Test Scores Reported (April-June)
Portfolio Review (June)
Summer School & Re-testing (July-August)
Final Retention Determination (August)

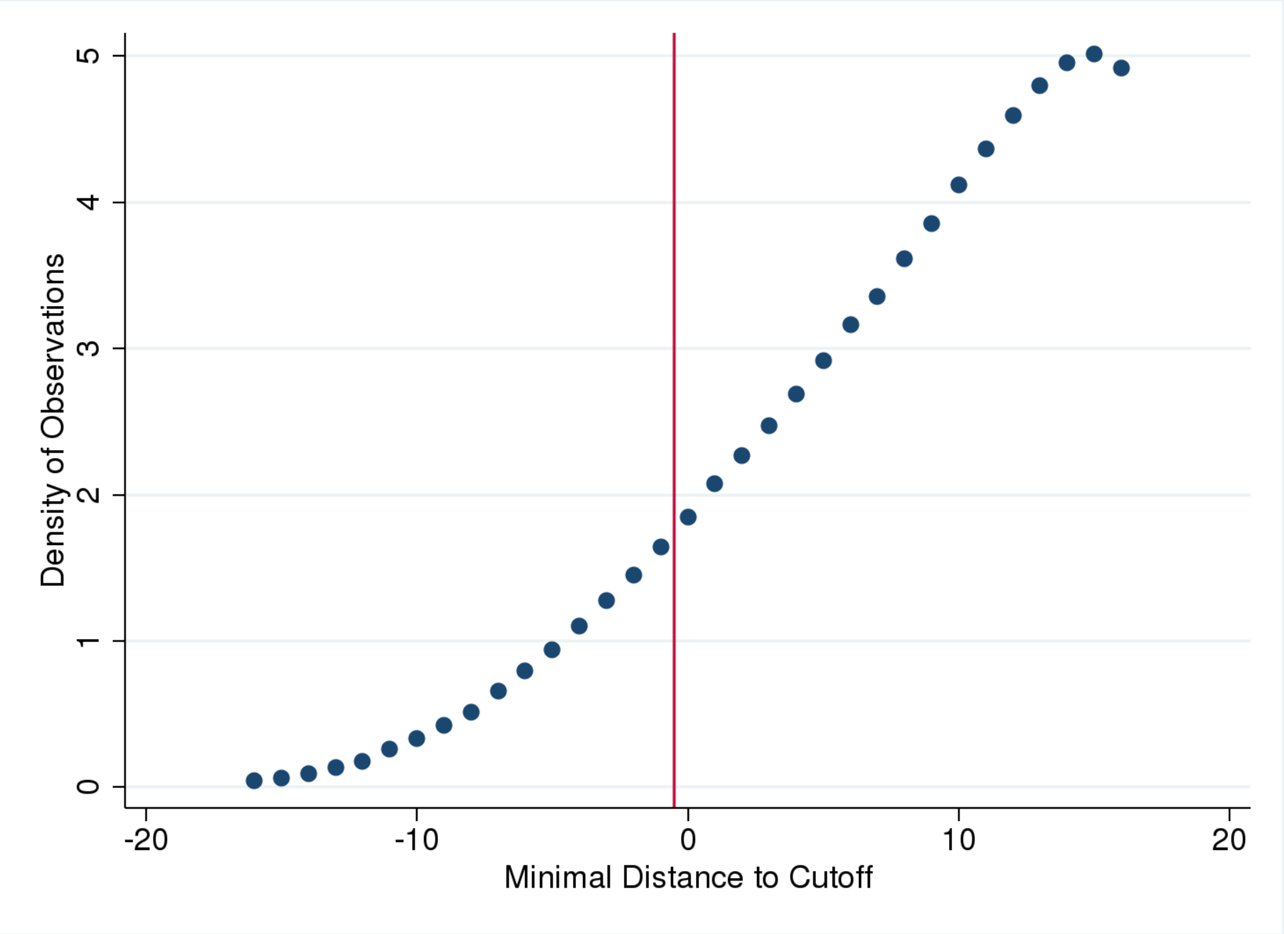
Note: Information on the timing and elements of promotion decisions is sourced largely from Crego et al. (2009).

Figure 2: Test-Score Based Retention Under Two Policy Regimes



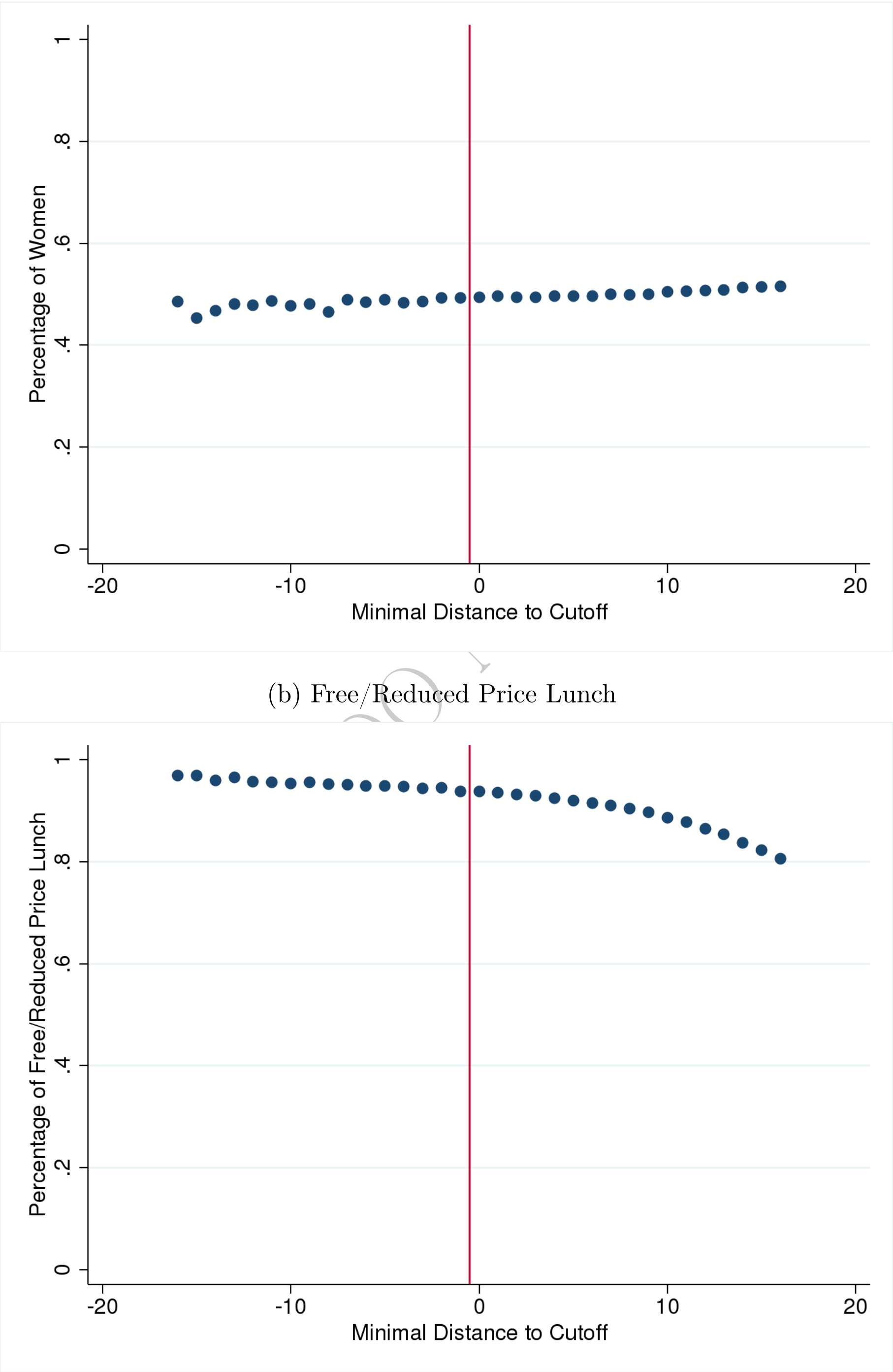
Note: This figure plots the average percentage of retained students at each index score, where a score of zero is equal to the cutoff for passing both exams. Retention rates are plotted separately by policy regime, where "post-policy" designates grade-year cells that have implemented a more stringent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

Figure 3: Density of Observations Across Cutoffs



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

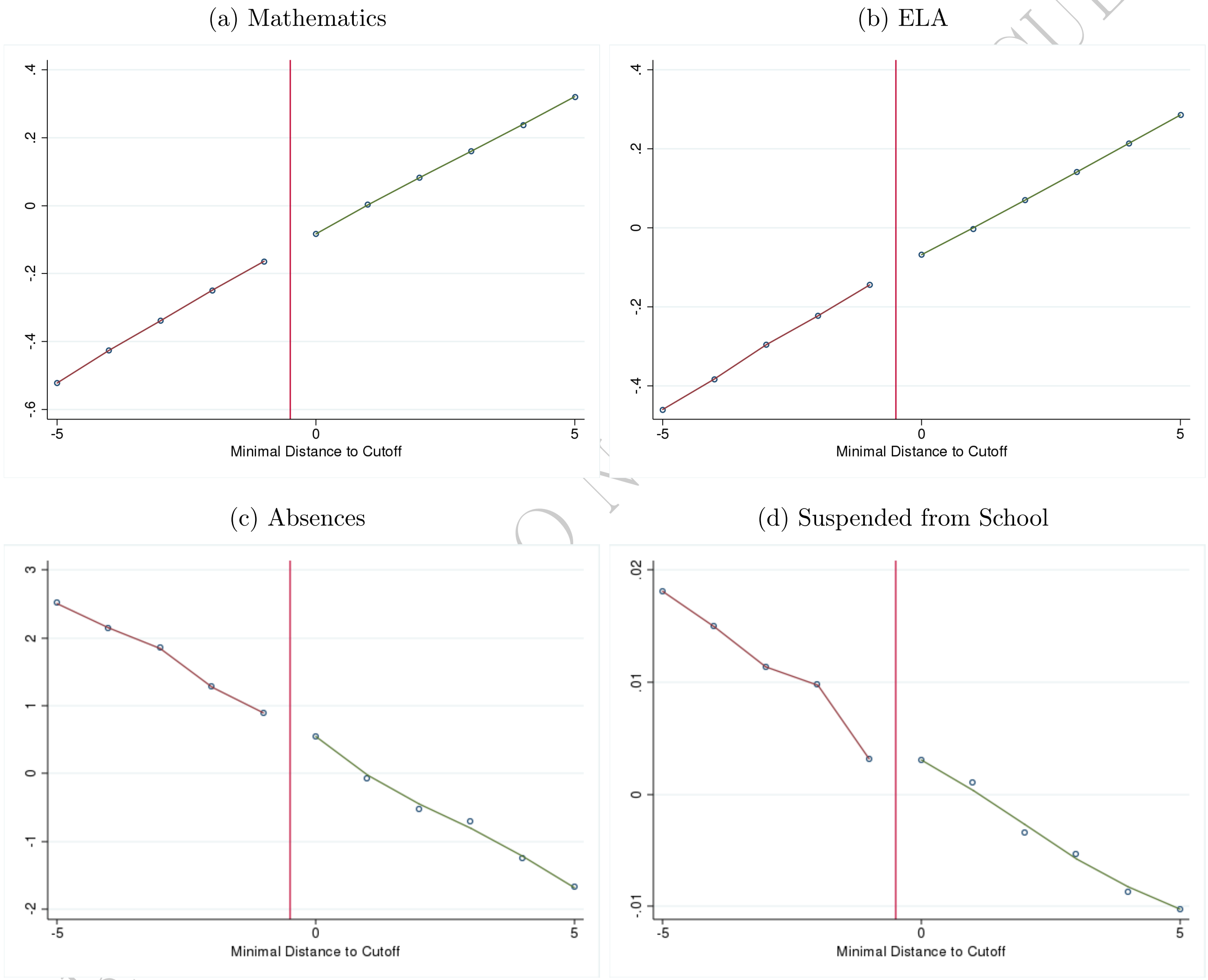
Figure 4: Continuity of Covariates Across Cutoffs


(a) Female

(b) Free/Reduced Price Lunch

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

Figure 5: Continuity of Current Test Scores, Absences, and Suspension


(a) Mathematics

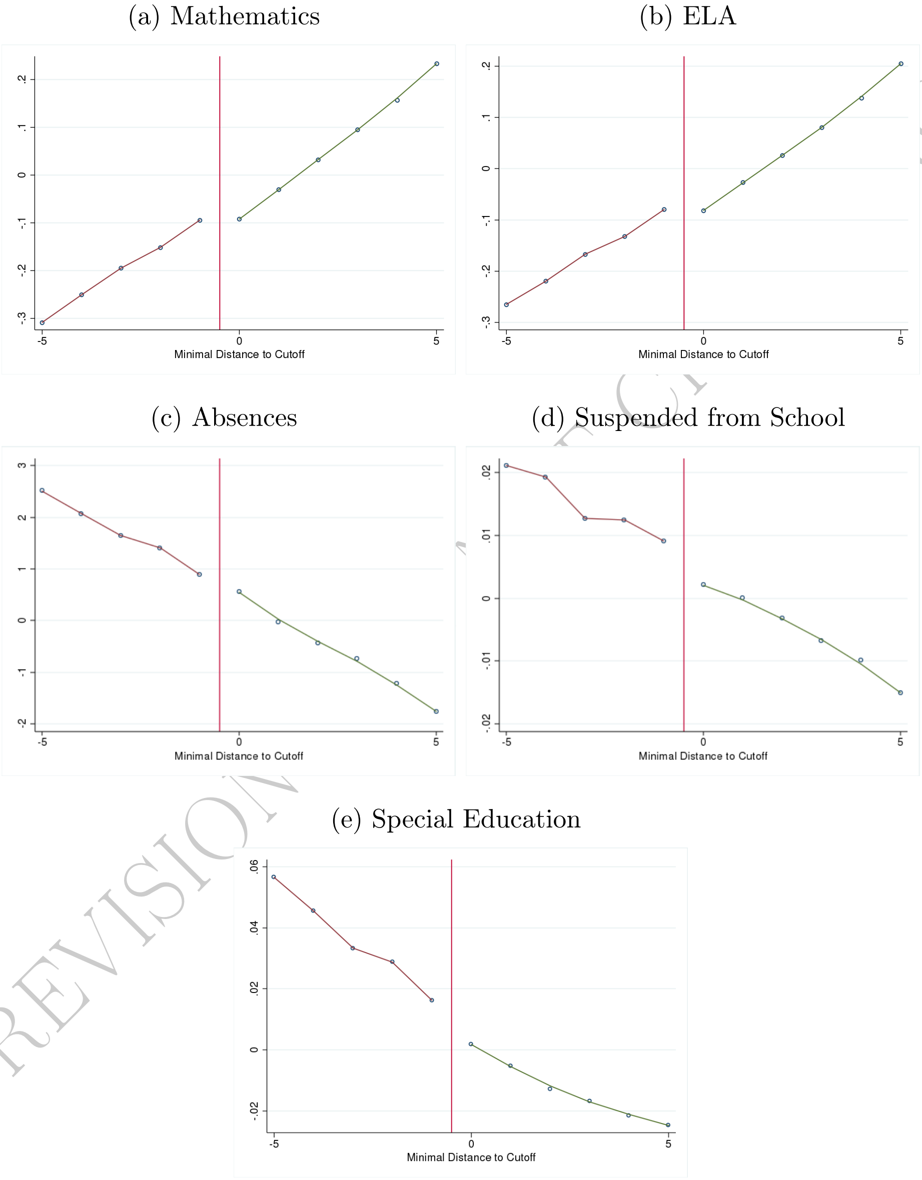
(b) ELA

(c) Absences

(d) Suspended from School

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

Figure 6: Visual Evidence on Future Test Scores, Absences, Suspension, and Speical Education


(a) Mathematics

(b) ELA

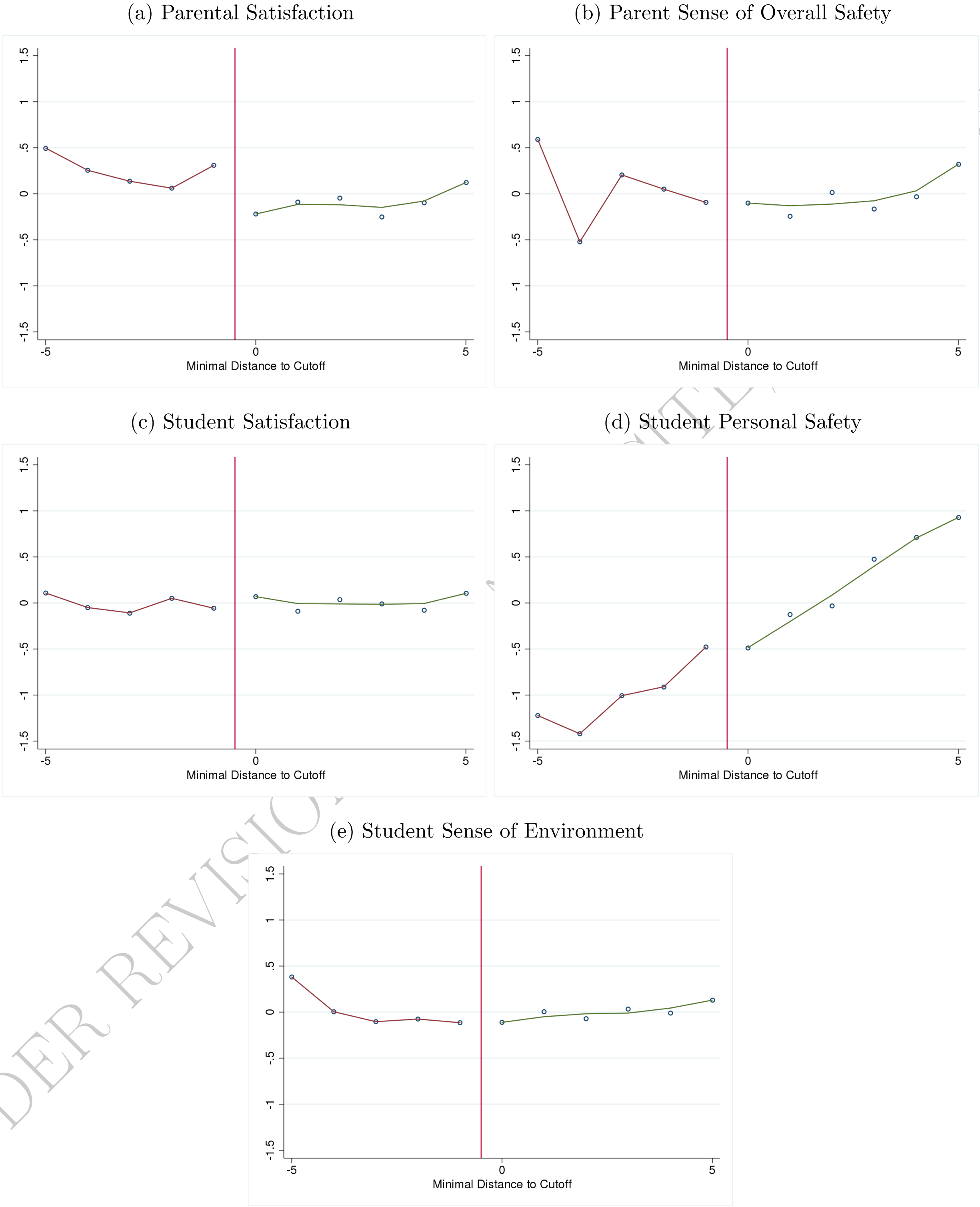
(c) Absences

(d) Suspended from School

(e) Special Education

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

Figure 7: Effects on Future Survey Responses



(a) Parental Satisfaction

(b) Parent Sense of Overall Safety

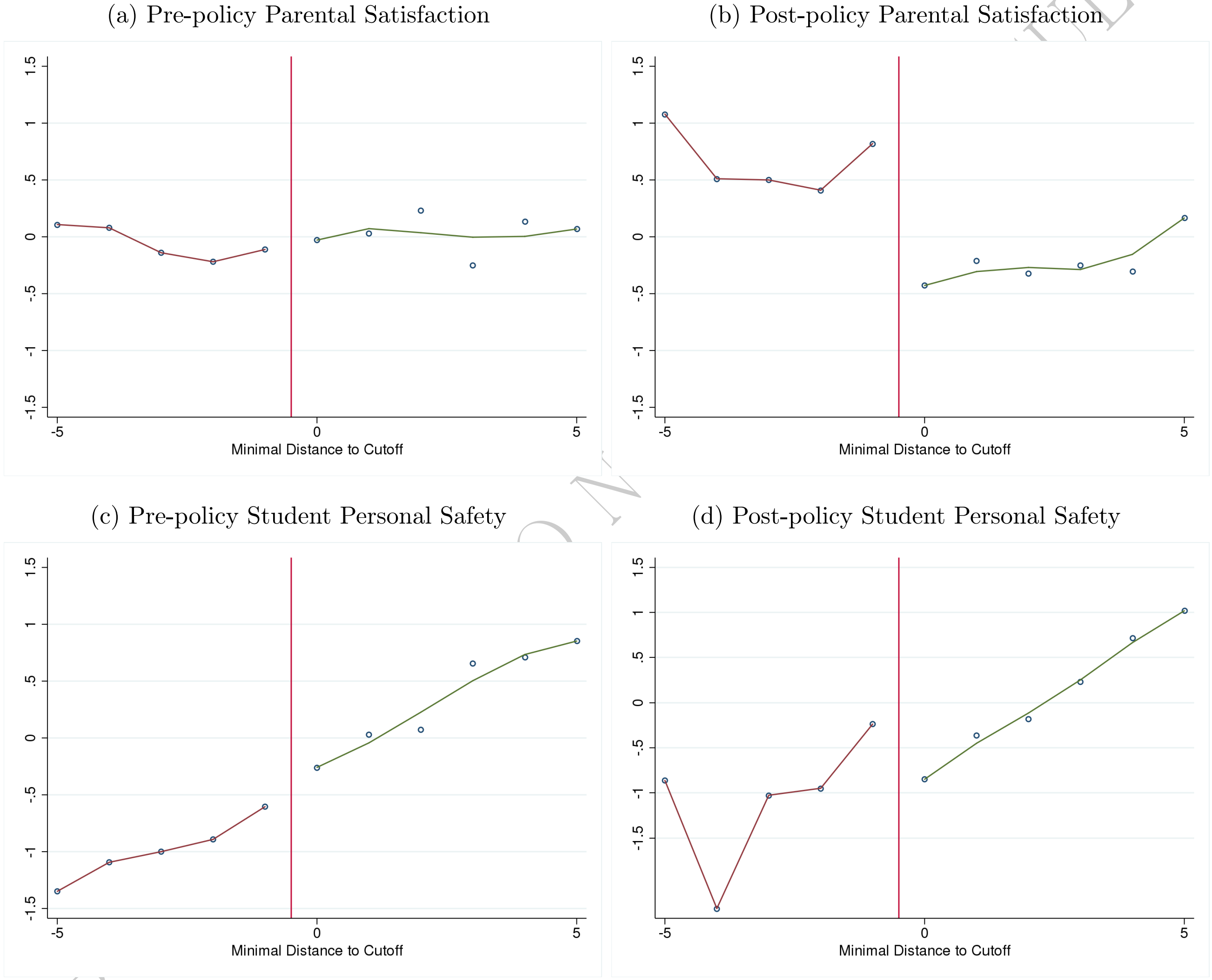
(c) Student Satisfaction

(d) Student Personal Safety

(e) Student Sense of Environment

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

Figure 8: Effects on Future Survey Responses by Policies



(a) Pre-policy Parental Satisfaction

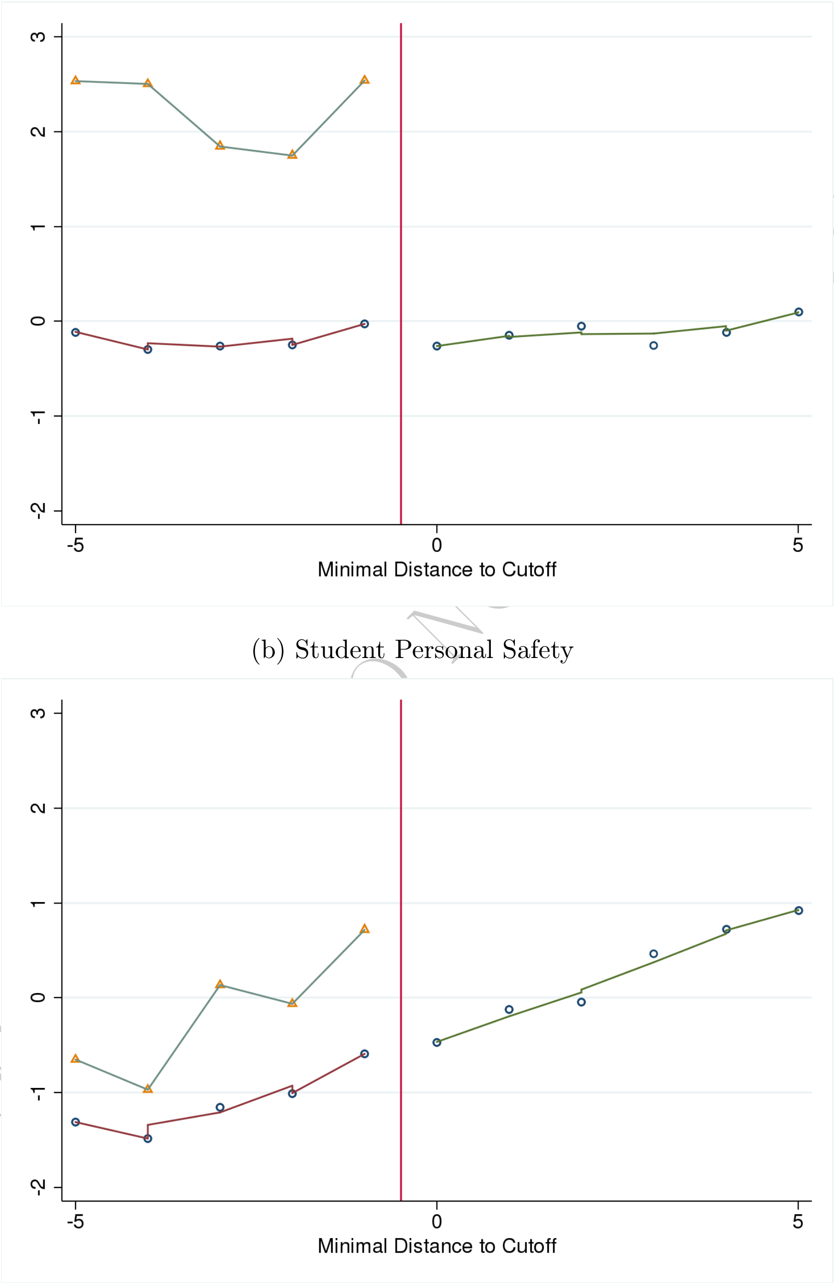
(b) Post-policy Parental Satisfaction

(c) Pre-policy Student Personal Safety

(d) Post-policy Student Personal Safety

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

Figure 9: Effects on Future Survey Responses by Actual Retention

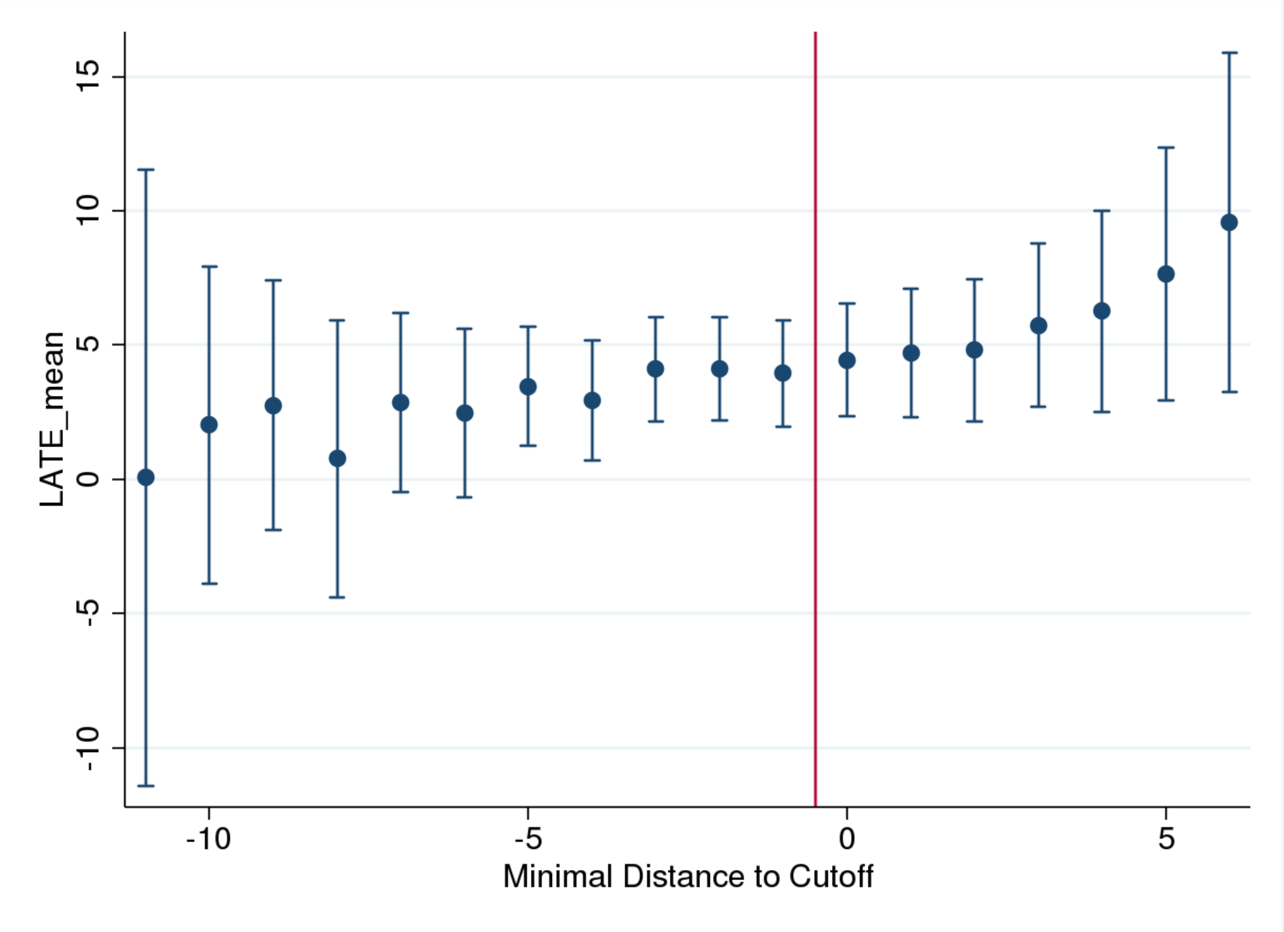


(a) Parental Satisfaction

(b) Student Personal Safety

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

Figure 10: CIA Estimates for Parental Satisfaction



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

8 Tables

Table 1: Summary Statistics

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

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

Table 2: First Stage Regression Results

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

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

Table 3: Effects on Test Scores, Absences, Suspension, and Special Ed from Schools

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

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

Table 4: Persistent Effects of Retention on Test Scores, Absences, and Suspension

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

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

Table 5: Effects on Survey Responses

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

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

Table 6: Persistent Effects on Survey Responses

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

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

Table 7: Effects on Parental Satisfaction and Students' Personal Safety between Policies

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

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

A Survey questions in each category

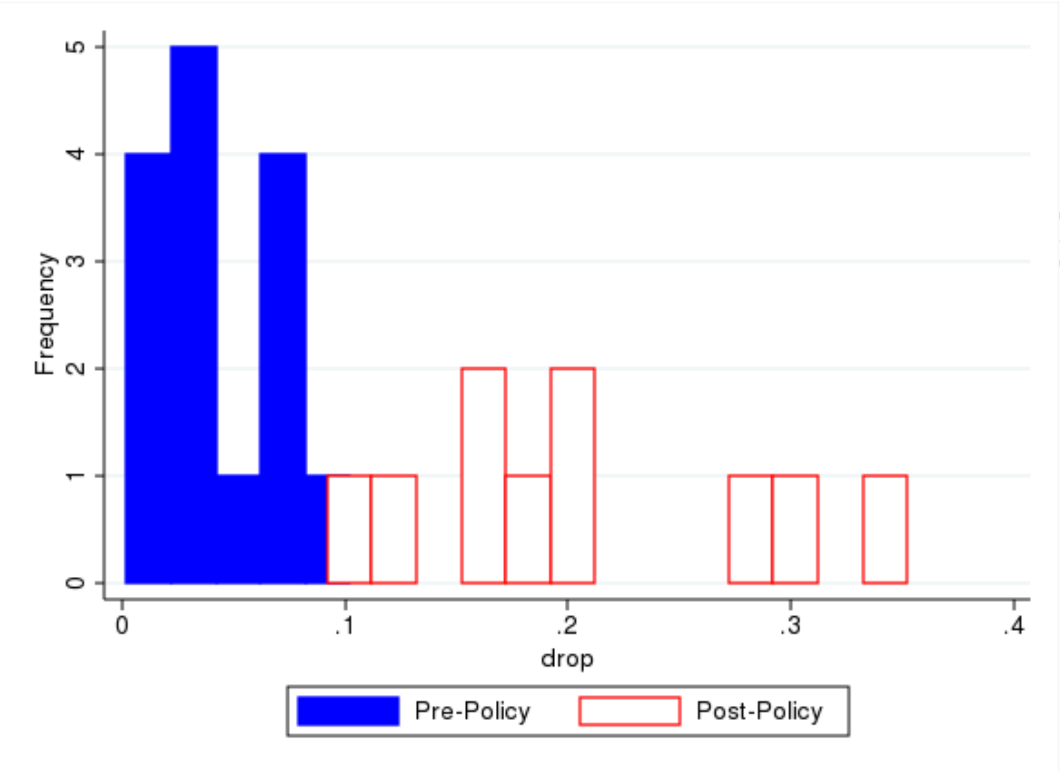
- Parents' overall satisfaction includes questions 2, 5, 9 and 13.
- Parents' sense of overall safety includes question 11.
- Students' overall satisfaction includes questions 2a, 3e, 3f, 3g, 6a, 6c-6g, 14a
- Students' sense of personal safety includes 13a, 13e, 13f, 13g
- Students' perception of environment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

B Continuity of Personal Characteristics

In order to validate our Regression Discontinuity Design, we test continuity of characteristics other than percent of women, percent of reduce/free-price lunch recipients, and density of observations (shown in Figures 3 and 4). Appendix Figure A.7 shows that the percentage of each ethnicity is continuous across the cutoff. Appendix Figure A.8 presents the percentage of students who stay at NYC public schools next year at each index and there is no discontinuity at the cutoff. Appendix Figure A.9 shows the percentage of students and parents who responded to surveys by index score and both rates are smooth through the cutoff. We also test continuity by regression analysis. These results are in Appendix Table A.12. This supports the notion that our results are not driven by any discontinuity of other student characteristics across the cutoff.

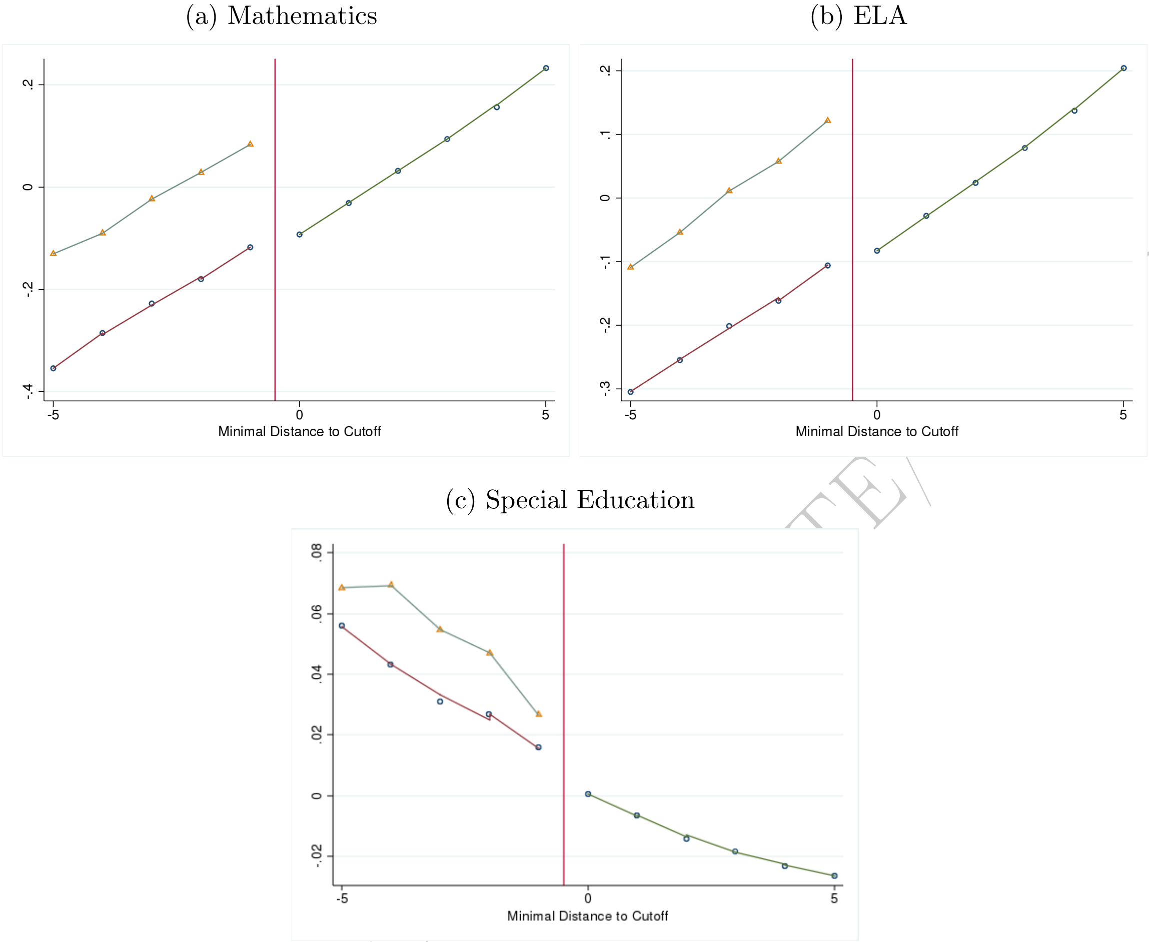
C Appendix Figures and Tables

Figure A.1: Frequency of Pre- and Post-Policy Retention Rates



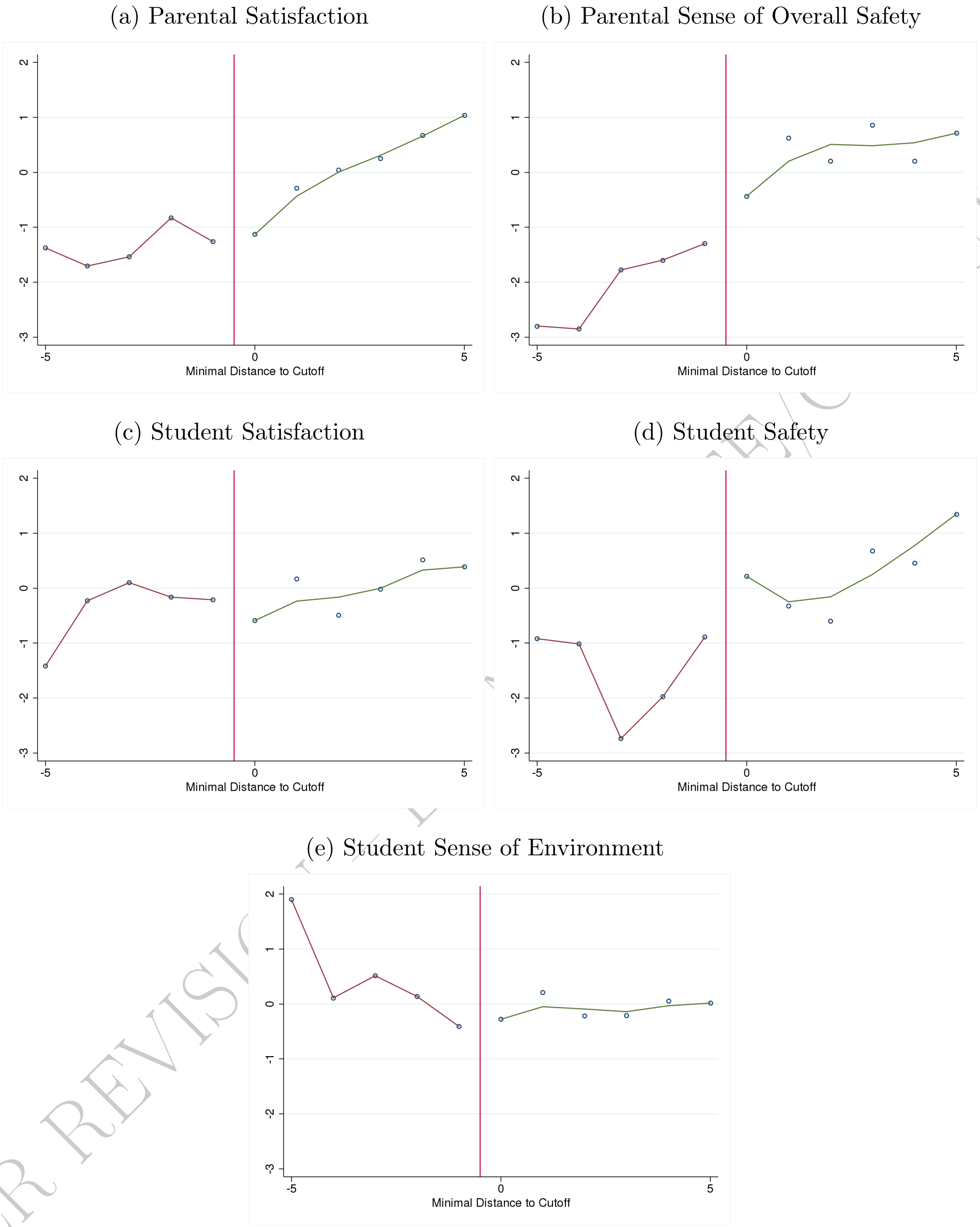
Note: The retention rates are the discontinuity in the probability of being retained at the cutoff. This figure plots the histogram of change in retention rate at the cutoff at each grade-year cell.

Figure A.2: Effects on Future Test Scores and Special Education



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

Figure A.3: Placebo Effects on Survey Responses



(a) Parental Satisfaction

(b) Parental Sense of Overall Safety

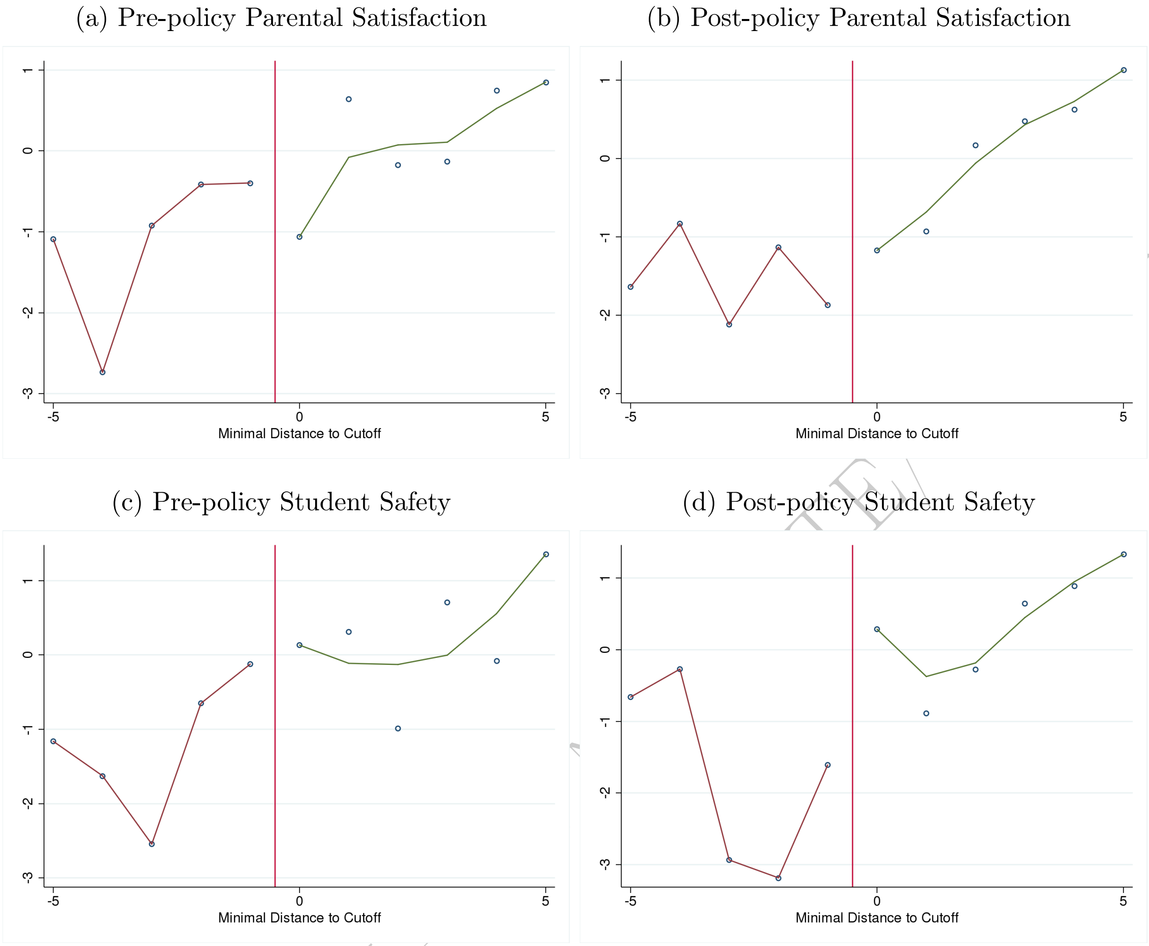
(c) Student Satisfaction

(d) Student Safety

(e) Student Sense of Environment

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

Figure A.4: Placebo Effects on Survey Responses by Policies



(a) Pre-policy Parental Satisfaction

(b) Post-policy Parental Satisfaction

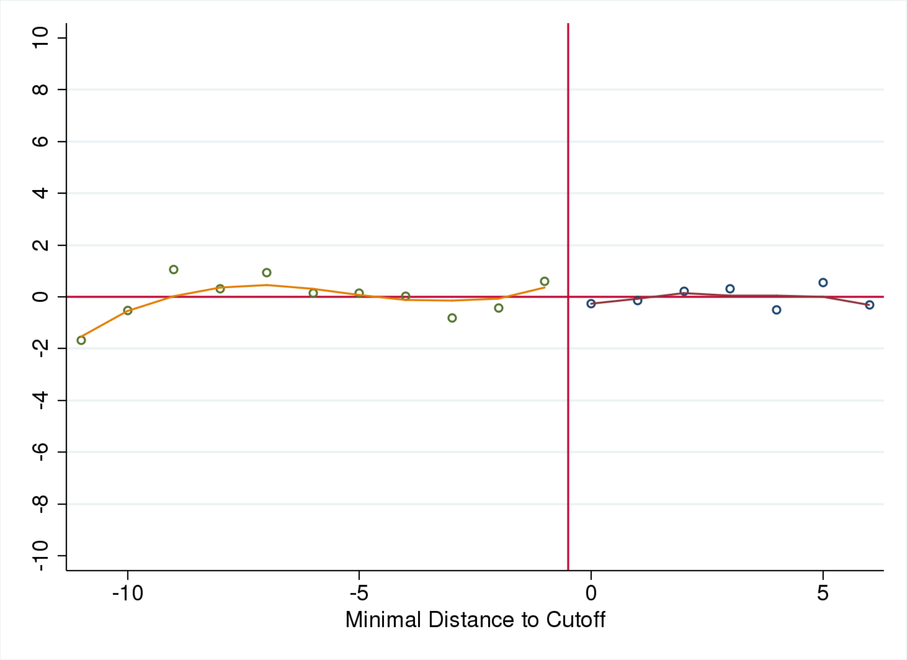
(c) Pre-policy Student Safety

(d) Post-policy Student Safety

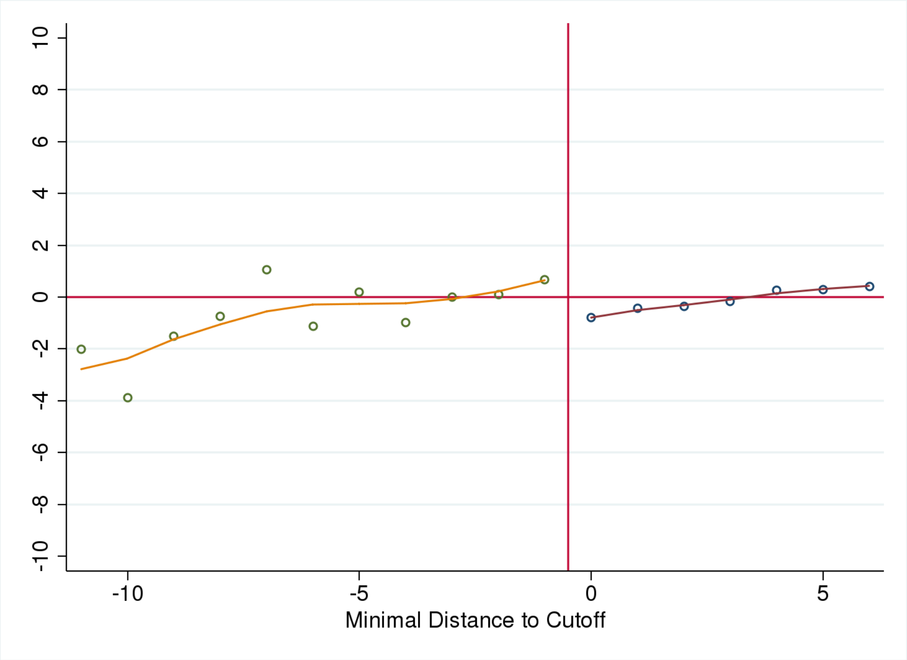
Note: These figures plot residuals from regressions of current (i.e. prior to retention) values of parental satisfaction, parental sense of overall safety, student satisfaction, student safety, and student sense of environment on fixed effects for test grade by test year. Plots are done separately by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

Figure A.5: CIA Visual Test

(a) Parental Satisfaction

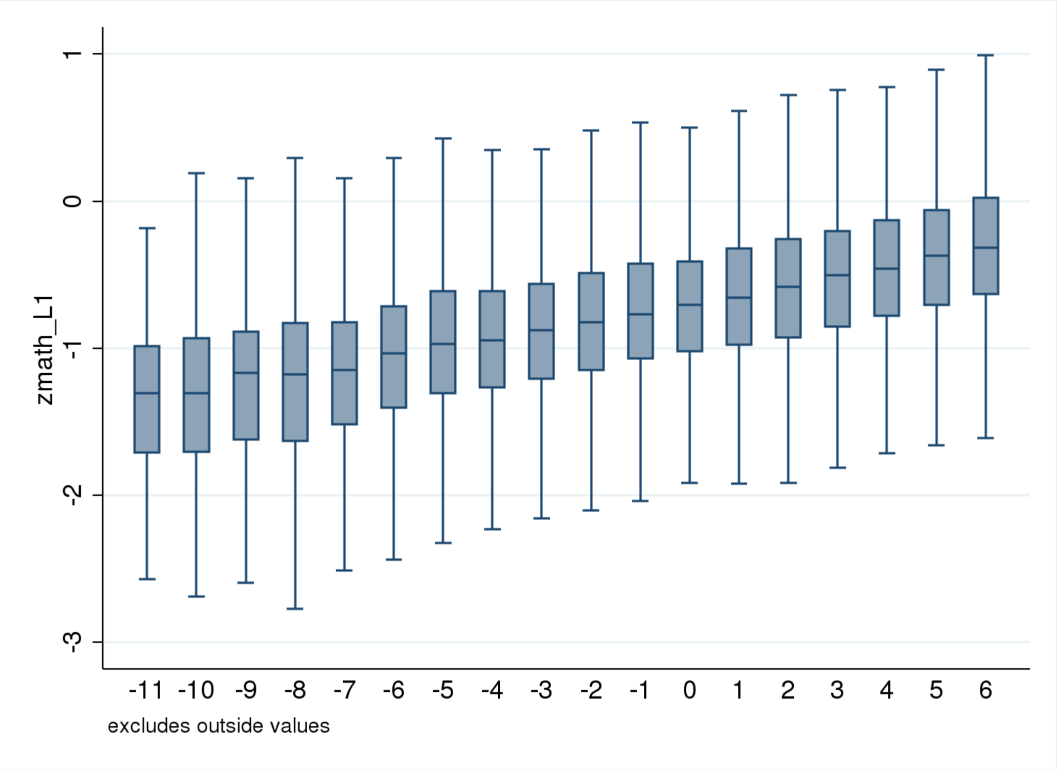


(b) Student Safety



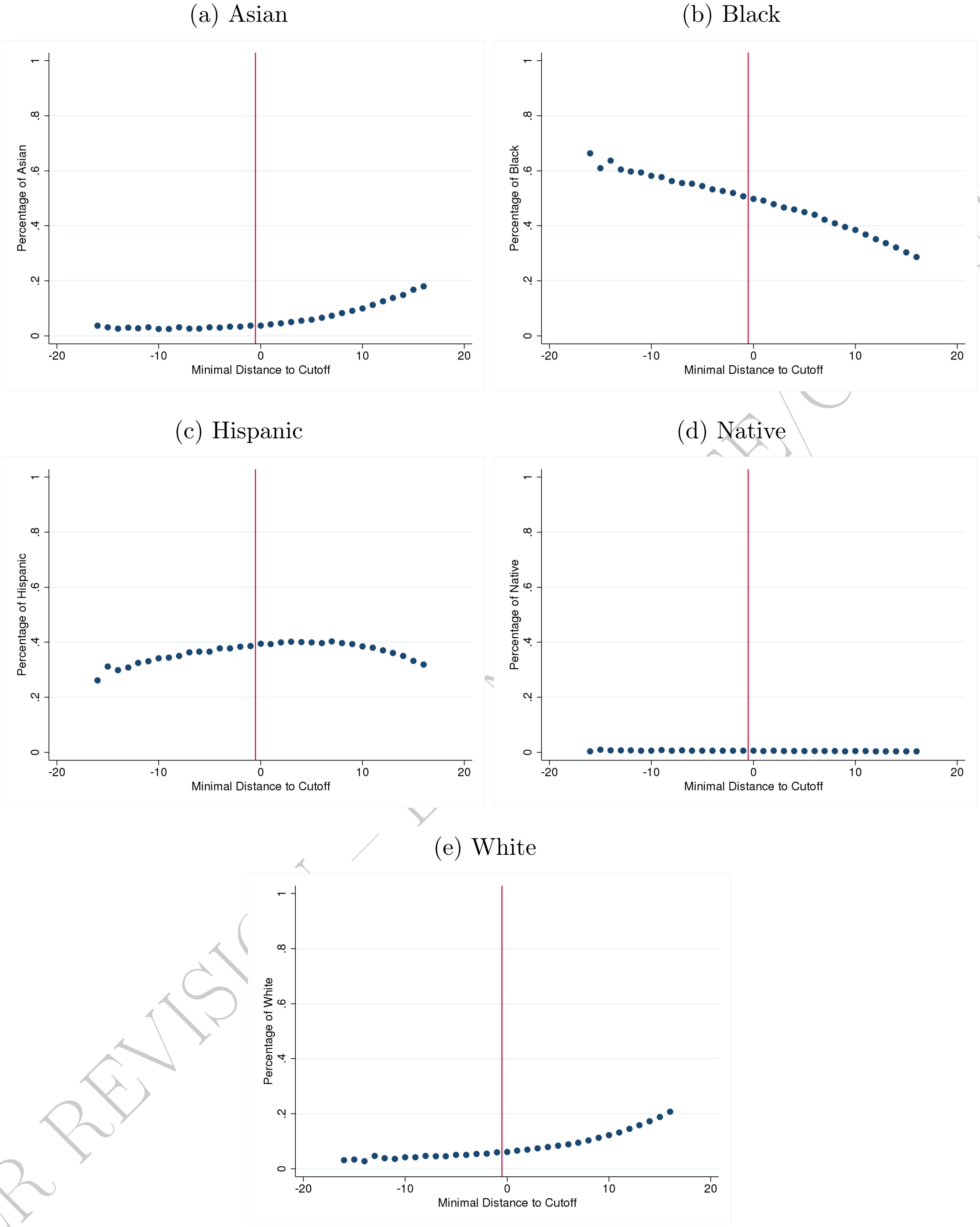
Note: These figures plot residuals from regressions on pre-determined covariates by index score for parental satisfaction and student safety. Comparing the LOWESS and the horizontal line at each side of the cutoff only supports CIA with respect to parental satisfaction outcome. Appendix Table A.11 presents regression results and suggests the same conclusion.

Figure A.6: Distribution of Two-year Prior Mathematics Score



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

Figure A.7: Continuity of Other Personal Characteristics



(a) Asian

(b) Black

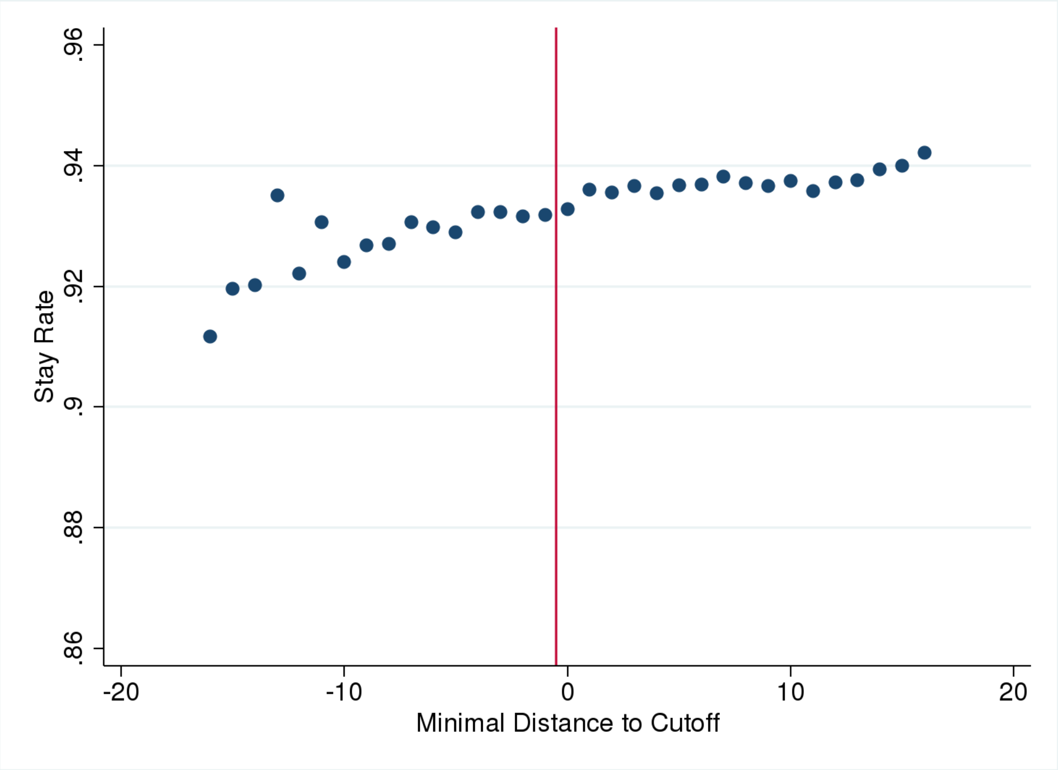
(c) Hispanic

(d) Native

(e) White

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

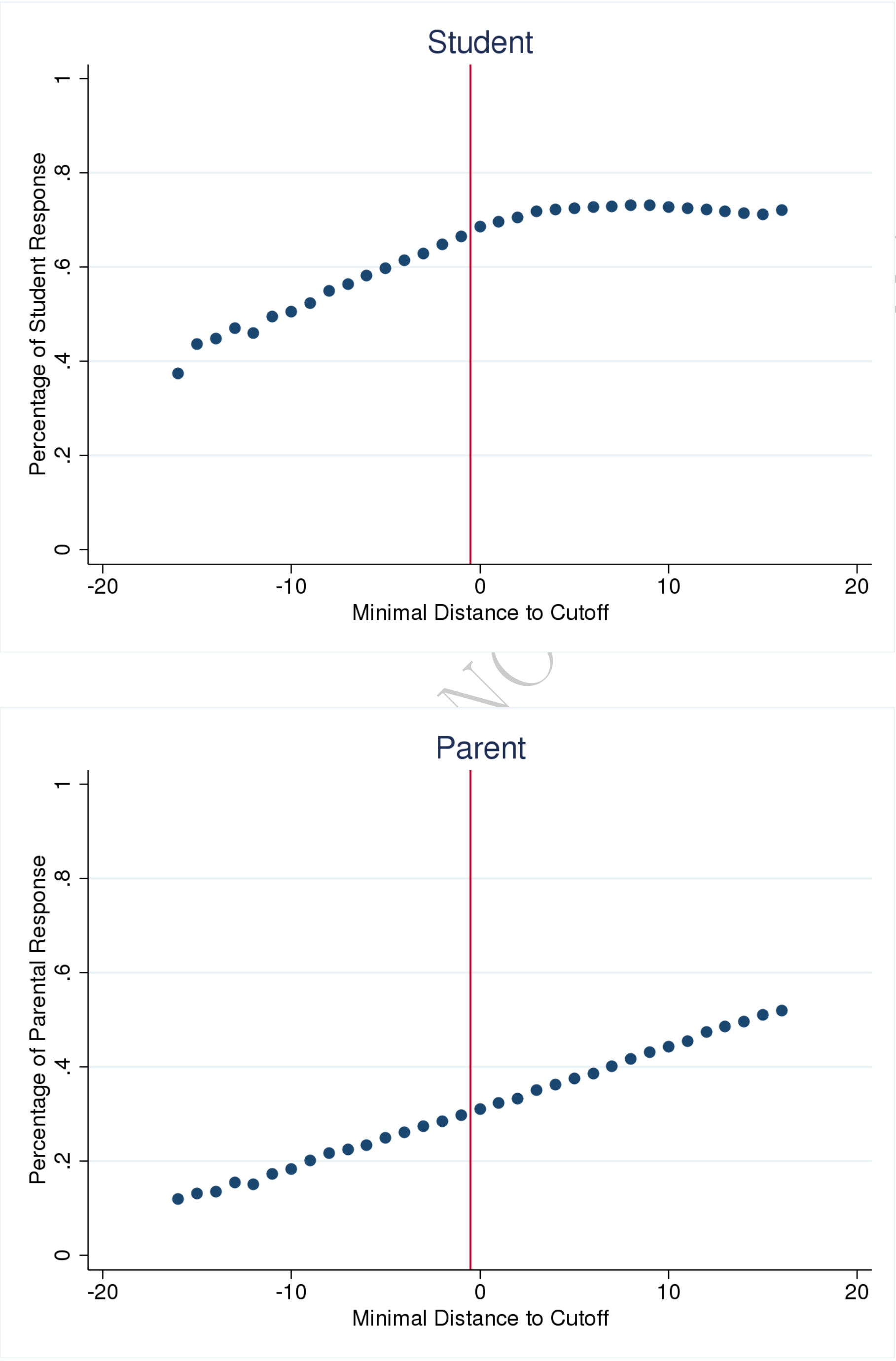
Figure A.8: Continuity of Attrition Rate



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

Figure A.9: Continuity of Response Rates against Indexes

(a) Student Response Rate



(b) Parent Response Rate

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

Table A.1: Effects of Retention with Additional Grade and Year

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

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

Table A.2: Persistent Effects of Retention with Additional Grade and Year

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

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

Table A.3: Effects of Retention by Policies with Additional Grade and Year

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

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

Table A.4: An Example of Data Stacking

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

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

Table A.5: Effects on Test Scores between Policies

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

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

Table A.6: Effects on Survey Responses by Bandwidths

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

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

Table A.7: Effects of Retention with Two-way Clustering

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

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

Table A.8: Persistent Effects of Retention with Two-way Clustering

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

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

Table A.9: Effects of Retention by Policies with Two-way Clustering

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

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

Table A.10: Effects on Survey Responses with Additional Covariates

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

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

Table A.11: CIA Test

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

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

Table A.12: Continuity of Covariates Test

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

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