Retention Heterogeneity in New York City Schools*

Douglas Almond†, Ajin Lee,‡ and Amy Ellen Schwartz§

This version: November 2016

Abstract

Performance on proficiency exams can be a key determinant of whether students are retained or "held back" in their grade. In New York City, passing the statewide proficiency exam essentially guarantees promotion, while roughly 13% of those students who fail the exam are retained. Using regression discontinuity methods, we find that female students are 25% more likely to be retained in their grade due to exam failure than boys. Hispanic students are 60% more likely and Black students 120% more likely to be retained due to exam failure (relative to White students). Poverty and previous poor performance also increase the likelihood of retention, while being young for grade or short does not. We conclude that "patterned discretion" exists in how standardized test results are utilized.

*We thank the New York City Department of Education and Office of School Wellness Programs for providing the microdata. Financial support from the National Institutes of Health, Award #5 R01 HD070739 (Schwartz) and the National Science Foundation CAREER Award #SES-0847329 is gratefully acknowledged (Almond). We are grateful to Siddhartha Aneja and Meryle Weinstein for assistance with the data.

†Columbia University and NBER: da2152@columbia.edu

‡Columbia University: al3045@columbia.edu

§Maxwell School of Syracuse University: amyschwartz@syr.edu

1 Introduction

US school districts increasingly rely on standardized tests to evaluate teachers and students. Performance on "high stakes" tests can be a key determinant of whether students are retained or "held back" in their grade. Well-identified studies have found retention can be beneficial for short-term subsequent academic performance but possibly detrimental to longer-term outcomes that might be of greater importance [Jacob and Lefgren, 2004, 2009]. Reliance on such tests is controversial in the US. For example, New York State is grappling with a sharply increased opt-out rate in spring 2015 by students who declined to sit for the statewide proficiency exam [*New York Times*, 2015].

We depart from previous literature by considering heterogeneity in how performance on standardized tests maps into consequences for students. Despite benchmarking from a common test and cutoff score, substantial scope for discretion exists in how exam results are utilized. Failing the exam can merely "start a conversation" about retention, where more often than not the student is promoted to the next grade. The lack of deterministic link between exam performance and retention opens the door to other factors shaping the retention decision. At present, we have little sense of how non-test factors shape retention among students who scored the same.

We analyze longitudinal data on 250,000 New York City public school students scoring near the failure threshold. Passing the annual proficiency exam essentially guarantees promotion to grades 4-9, while roughly 13% of those students failing the exam are retained. Compliers in our application are those who are retained because they failed the proficiency exam. Because there is a large population of never takers (promoted despite exam failure), the compliant sub-population may differ from not only the overall New York City student population (obvious), but also from the sub-population located near the threshold (less obvious). We analyze retention and average complier characteristics [Angrist and Pischke, 2009] using regression discontinuity methods.

We document pronounced heterogeneity in compliance along observable characteristics of the student. Moreover, this heterogeneity departs in important ways from what we had expected *a priori*.[1] In particular, we expected compliance to be highest among the youngest students, who were closest to the age-at-school entry cutoff. These students narrowly missed beginning kindergarten a year later and are on average less developed academically, socially, and physically than peers (particularly in early grades). Using administrative data on birth month, however, we do not find that retained students tend to be young for their grade. Nor do we find older students are more likely to be promoted after failure. Instead, we find

[1] See Tomchin and Impara [1992] for a description of factors affecting the probability of retention.

race and gender to be important. Hispanic students are 60% more likely and Black students 120% more likely to be retained due to exam failure (relative to White students). Female students are 25% more likely to be retained in their grade due to exam failure than boys.[2] Poverty (free or reduced-price lunch eligibility[3]) and poor performance on previous exams also increase the likelihood of retention. Like age for grade, biometric measures of student height and weight do not seem to play a large role beyond the exam score. Again, we had expected smaller-stature students might face a higher retention risk when they fail because they might "fit in" physically in their repeated grade. We also show these biometric and demographic characteristics are smooth at the threshold. Thus it is not the case that, for example, Black students have discontinuously worse characteristics just below the threshold for passing. Nor do we find any evidence of heaping near the threshold.

We discuss two classes of "explanations" for the retention heterogeneity we uncover: student-level differences and school-level differences. Regarding the former, it is not the case that the predictive power of the baseline test score is different for girls or minorities than for the rest of the student population (located near the failure threshold). Thus, we do not see evidence that, for example, girls are more likely to be retained when they fail because failure is a stronger predictor of future (poor) performance. On average, girls perform better in subsequent periods than boys with identical baseline scores. Other factors equal, this would suggest that the compliance rate among girls should be lower than for boys. Higher compliance of girls' retention with exam failure is puzzling. The unexplained gender gap is widest among Whites: failing increases a girl's retention rate to 5.9%, but when a White boy fails, only 0.9% are retained. Indeed, we cannot reject that exam failure has *zero* impact on retention for non-Hispanic White boys.

Turning to school-level characteristics, these are "balanced" by sex so disproportionate retention of girls who fail cannot be attributed to differential exposure to school characteristics. Race and ethnicity, in contrast, do vary with school-level characteristics. Among these school-level factors, "high retention" schools have more minority students on average. Furthermore, predominantly Black schools tend to be high compliance schools, i.e. schools where retention rates jump more below the failure threshold. While school-level factors thus appear important to racial heterogeneity in retention, so too do within-school factors. Blacks are substantially more likely to be retained than Whites (for identical baseline scores) at predominantly non-Black schools.

The existing literature has overlooked compliance heterogeneity: we know of no published

[2]Significant at the 1% level: see Section 5 and footnote 7.

[3]Students are eligible for free lunch if their parents or guardians make less than 130% of the poverty line and reduced lunch if their parents/guardians make less than 185% of the poverty line.

work on the subject.[4] In addition to student composition of schools, we also consider faculty [Dee, 2005]. The final retention decision is made by the school principal. We find a striking pattern whereby girls are substantially more likely to be retained due to exam failure at schools with a female principal. That said, because other (unobserved) characteristics of the school presumably vary by principal's characteristics (*cf.* student gender), we characterize this pattern as descriptive. Furthermore, because girls perform better on average than boys, the *unconditional* retention rates remain lower for girls than boys: girls score better on average and fewer girls fail (overall). This and the fact that relatively few students are retained in a given school each year may have obscured higher retention rates among girls who just fail.

2 Background

2.1 Literature Review

Previous papers have used regression discontinuity approaches to consider impacts of retention on subsequent outcomes, beginning with Jacob and Lefgren [2004]. Among third graders in Chicago public schools, Jacob and Lefgren [2004] found positive effects of retention and more mixed impacts among sixth graders. Jacob and Lefgren [2009] found that retention increased subsequent high school dropout rates. These findings are noteworthy as longer-term endpoints (like high school completion) might be more important endpoints for parents, students, and policy makers than shorter-term achievement. Because compliance rates are an order of magnitude higher in Chicago than in New York,[5] there is a different scope for heterogeneity in compliance in Chicago's context compared to New York.

Mariano and Martorell [2013] follow Jacob and Lefgren [2004, 2009] and exploit test score cutoffs used in assignment to summer school and retention in New York City. Specifically, they consider 2004-2008 data on fifth graders failing proficiency exam in 2004-2006. They find modest positive effects of summer school on English achievement. They estimate cohort-over-cohort test score differences ("external drift") and subtract it from the RD estimates of retention (see 5.5 section). They find large and positive effects of grade retention on both Math and English. As in Jacob and Lefgren [2004, 2009], heterogeneity in compliance is not considered.

Student characteristics, however, might conceivably play a role by shaping interactions between teachers and students. Dee [2005] uses National Education Longitudinal Study

[4]Two recent working papers using Florida records are discussed in Section 2.1.

[5]41% of sixth-graders who failed to meet the promotion cutoff were retained in Chicago from 1993-1994 to 1998-1999 [Jacob and Lefgren, 2004].

of 1988 (NELS:88) to examine the role of demographic similarity between teachers and students on teachers' perceptions of students. Dee [2005] makes within-student comparisons of teachers' perceptions, taking advantage of the structure of NELS:88 data, which surveyed teachers in two different academic subjects, on their perceptions of individual students. Dee [2005] finds that teachers are more likely to have negative perceptions towards students who do not share the same race/ethnicity and gender. His findings suggest that demographic characteristics of students such as gender and race/ethnicity may potentially matter for retention decisions as well, as they are partly based on teachers' evaluations of students.

Labelle and Figlio [2013] and Schwerdt et al. [2015] consider Florida's test-based promotion policy and evaluate various future outcomes. Labelle and Figlio [2013] stands out as most similar to our approach (we discovered their conference draft after conducting our analysis). Labelle and Figlio [2013] examine whether Florida's grade retention policy that mandated promotion to the fourth grade conditional on meeting a minimum standard in third grade reading was being implemented differently depending on maternal education (using matched educational data and birth records). They employ a regression discontinuity design, taking advantage of the score cutoff for determining retention, finding that students whose mothers have less than a high school degree are 20 percent more likely to be retained than students whose mothers have a bachelor's degree or more. Factors besides parental education, including eligibility for free school lunch and other dimensions of student performance, shape heterogeneity in compliance as well. They also estimate the effect of retention on future test scores instrumenting for grade retention with scoring below the promotion cutoff. They find that retention leads to short-term gains in test scores but that the gains fade out over time, consistent with Jacob and Lefgren [2009]. They find no evidence, however, that differential retention by maternal education has differential impacts on students' future test scores.

Labelle and Figlio [2013] additionally show that students are more likely to be retained if they are Black (9 percent increase), male (13 percent increase), have a foreign born mother (13 percent increase), and qualify for free or reduced-price lunch (9 percent increase). Within subgroups categorized by student race, free or reduced-price lunch eligibility, and school characteristics, they still find a similar (but imprecise) pattern in retention probabilities by maternal education. Heterogeneity by student gender is not discussed. They attribute differential retention by maternal education to systematic differences in parental behavior in response to retention risk, although they cannot directly test this hypothesis.

Schwerdt et al. [2015] emphasize the impacts of test failure on retention and future outcomes. They attempt to address the endogeneity of the subsequent exam to retention and consider subsequent reading, Math test scores, and high school graduation. Short-term

gains in both Math and reading fade over time. They also find no clear impact on graduation and little evidence of systematic heterogeneity by student and school characteristics. They also look at complier characteristics [Angrist and Pischke, 2009] and find that a complier is more likely to score level 1 in Math. Their conclusion is that early grade retention might be favorable (e.g. short-term gains and no detrimental effects), although long-term benefits are uncertain.

At present, there is no published work using regression discontinuity methods to consider heterogeneity in compliance with exam failure/passing. The magnitude of heterogeneity we find in New York City is substantially larger than that found in recent analyses of Florida students and manifests along additional dimensions, e.g. gender of student and principal.

2.2 Promotion Policy

In New York, students in grades 3-8 take the State Math and English Language Arts (ELA) tests each spring. The "scale score" is the number of correct answers converted into a vertically comparable score (comparable across grades). Scale scores are categorized into four performance levels separately for Math and ELA: level 1 - not meeting State learning standards, level 2 - partially meeting State learning standards, level 3 - meeting State learning standards, and level 4 - exceeding State learning standards.

Scoring level 2 ("partially meeting" standards) in both tests essentially guarantees promotion, whereas students who score level 1 ("not meeting" standards) in either subject are at risk of being retained. The failure threshold for each subject varies by year and grade. Retention procedures are less formalized in New York than Florida, with New York having few explicit exemptions. That said, English Language Learners and students with disabilities who receive special education services are exempt from New York's stated promotion criteria.[6] In our sample, 13% of students who failed to meet the promotion cutoff were retained. Thus, there is substantial scope for heterogeneity in compliance, driven predominantly by the "never takers".

3 Data

We analyze administrative data from the New York City public school system for the 2007-2008 to 2011-2012 academic years. Student-level panel data on New York State English Language Arts (ELA) and Mathematics scale scores are merged to demographic character-

[6]Empirically, however, we find that these groups of students were also affected by the policy and thus do not exclude them in our analysis. That said, our results are not sensitive to excluding them.

istics, including race, gender, free or reduced-price lunch eligibility, and age in months. Additionally, we observe students' weight, height, and BMI, measures further described in Almond, Lee, and Schwartz [2016]. Unique student identifiers allow us to track students over time as long as they stay in the New York City public school system. When the student's grade level in year $t+1$ is the same as that in year t, we code the student as retained. 1,507,700 student records for grades 3-8 are available 2007-2012, and approximately 2% are retained. The retention rates in grades 3-8 have increased over time from 1% in 2007-2008 to 3% in 2010-2011. Over our analysis period, roughly 4% of students are ever retained.

Table 1 reports mean student characteristics for the whole sample (column 1), those who passed both tests but scored within 10 units of the cutoff (column 2), and those who failed to meet the promotion cutoff in either test and within 10 scale score units (column 3). Relative to the overall sample, students in this "retention window" are more likely to be Black or Hispanic, and less likely to be Asian or White. The proportion of female students is *lower*, and the proportion of students who are eligible for free or reduced-price lunch higher near the cutoff. 13% of students below the failure threshold were retained while 0.7% of those "just above" the threshold were retained.

4 Estimation

To assess heterogeneity in how standardized test scores are utilized, we exploit the jump in retention rates at the failure threshold in a regression discontinuity framework. We estimate the following equation both "pooled" and separately by student characteristics:

$$Y_{igs,t+1} = \alpha_0 + \alpha_1 \cdot 1[X_{igst} < 0] + \alpha_2 \cdot X_{igst} + \alpha_3 \cdot 1[X_{igst} < 0] \cdot X_{igst} + \eta_{gst} + \epsilon_{igst} \quad (1)$$

where i is individual, g is grade, s is subject, and t is year. Y is an indicator for whether the student is retained or not. X_{igst} is minimum of the Math and English test scores, re-centered to zero at their respective failure thresholds. We use this measure as the main running variable, since students are at risk of grade retention when they score level 1 in *either* Math or English test.

We fit a linear relationship between the scale score and the probability of retention, allowing for different slopes above and below the cutoff (consistent with our figures). We include year×grade×subject fixed effects, η_{gst}, to control for year-, grade-, and subject-specific cutoffs. We estimate equation (1) by OLS and report robust standard errors.[7] We

[7]We do not cluster our standard errors at the running variable level since we found out that clustered standard errors from separate regressions are inconsistent with clustered standard errors from pooled regressions. In addition, Kolesár and Rothe [2016] argue that the convention of clustering standard errors on the

focus on the roughly 250,000 student observations within 10 scale score (approximately one third of a standard deviation for both Math and English) of the cutoff. In the tables, we report the RD estimate α_1, which measures the size of the discontinuity at the failure threshold.

4.1 Discontinuities in Baseline Covariates?

Figure 1 shows histograms of the running variable both in the full sample (panel (a)) and within 10 scale score from the failure cutoff (panel (b)). We do not observe any heaping around the failure cutoff (normalized to 0).[8] As there is no evidence of manipulation around the cutoff, we expect students to have similar characteristics above and below the cutoff. We summarize covariates by predicting the probability of retention using student gender, race/ethnicity, age in months, BMI, height, weight, free or reduced-price lunch eligibility, special education participation, and previous Math and English scale scores. Figure 2 compares this predicted probability of retention around the cutoff. There is no evidence of a discontinuity at the cutoff in the full sample (panel (a)) nor separately for females (panel (b)) or for Black students (panel (c)). The corresponding regression estimates of the discontinuities are precisely estimated zeros.

5 Results

Figure 3 summarizes the mean probability of retention for students near the cutoff. Consistent with stated school policy, the probability of retention drops discontinuously at the cutoff. Moreover, the linear specification seems to fit the data well [Gelman and Imbens, 2014]. Table 2 reports the RD estimates from estimating equation (1) "pooled" and separately by subgroup. Overall, failing to meet the promotion cutoff increases the probability of retention by 5 percentage points (column 1).

We are particularly interested in documenting whether exam failure has different retention consequences by baseline characteristics. Panel A of Table 2 shows that Black students are 3.4 percentage points more likely to be retained than White students, more than double the White retention probability as induced by failure (2.9%). Hispanic students are around 2 percentage points more likely to be retained than non-Hispanic Whites, a 60% increase.

running variable performs poorly in a regression discontinuity framework with a discrete running variable.

[8]Dee et al. [2011] document evidence of manipulation of Regent's exam scores among New York City *high school* students, finding "roughly 3 to 5 percent of the exam scores that qualified for a high school diploma actually had performance below the state requirement". Key for us, they do not find any evidence of manipulation on the statewide Math and English exams given to students in grades 3-8. Likewise, we detect no evidence of manipulation among the proficiency exams taken prior to high school (i.e. grades 3-8).

Asians are, if anything, are less likely to be retained than non-Hispanic White students when they fail to meet the cutoff.

Girls are 1.2 percentage points (or 27%) more likely to be retained than boys failing the exam (panel B of Table 2). The gender difference is statistically significant at the 0.01 level (p-value = 0.005). This is intriguing since the overall retention rate in grades 3-8 is higher for boys (2.1%) than for girls (1.7%). But when we examine the retention rates in a narrow window near the failure threshold, we find the opposite: girls are more likely to be retained than boys. Additionally, we find that low performance on previous year's Math test increases the probability of retention (panel C in Table 2).[9] Finally, those who are eligible for subsidized lunch are 1.3 percentage points more likely to be retained than those ineligible (panel D in Table 2).

Figure 4 presents these findings graphically. Panel (a) shows a large disparity in retention between Black and White students below the cutoff. Likewise, panel (b) shows the mean probability of retention is higher for Hispanics than for Whites, although the gap is smaller. Panel (c) shows that mean retention probabilities are similar between Asians and Whites. Panel (d) shows Blacks and Hispanics are roughly twice as likely to be retained than Asians and Whites when they fail. Panel (e) shows that girls are more likely to be retained than boys conditional on scoring identically below the cutoff. Below the threshold, the girl mean is above the boy mean at each scale score, but means are indistinguishable above the threshold. Additionally, we examine whether the probability of retention differs by age for grade, height for grade, and weight status category. We might expect students who are younger or smaller than their peers in the same grade are more likely to be retained, since they would potentially fit in better in their repeated grade (socially, physically, and academically). Parents might also be less likely to object to the retention decision if their child was a "close call" with respect to age at school entry cutoff. We also test whether students who are "too big to fail" are in fact less likely to be retained. However, we find surprisingly little heterogeneity along these dimensions (Table 3).

Given the stark heterogeneity by ethnicity and gender, we consider interactions between these dimensions. The gender gap is especially large among Whites (5.9 percentage points versus 0.9 percentage points). For all ethnicity groups, we find that girls are more likely to be retained than boys.

[9]Additionally, we examine whether the demographic heterogeneity we find disappears once we condition on previous test scores. We estimate separate regressions by student characteristics for each decile of previous test scores. We find that ethnicity and gender heterogeneity in retention probability generally persist across the distributions. This suggests that there are other factors driving the differential retention probabilities that are independent of previous academic performance.

5.1 Racial Composition of Schools

Here we explore the role of school-level differences in explaining heterogeneity. Retention policies and practices are shaped by principals and teachers, and thus may differ by school. Given pronounced residential sorting within New York City, Black students might disproportionately attend schools that more strictly adhere to a test-based promotion policy than schools White students attend. We examine whether the probability of retention due to exam failure differs between schools with different Black shares, dividing schools into three equal-sized groups by their proportion of Black students: low (mean 5%), middle (mean 24%), and high (mean 60%). The overall mean retention rate in grades 3-8 is higher in high share schools (2.8%) than low share schools (0.8%). Furthermore, predominantly Black schools tend to be high compliance schools, i.e. where discontinuity in retention rates is larger at the failure threshold.[10] Thus school-level differences can "explain" (in a statistical sense) some of the individual differences in retention by race.

However, panel A of Table 4 shows that the Black-White gap in the probability of retention is much larger in schools with *low* share of Black students. At predominantly Black schools, we do not see a racial disparity in the effect of failing the exam. The difference in retention probability between Black and White students is only 0.7 percentage points, and it is not statistically significant. Panel B shows that these findings are not sensitive to including school fixed effects (nor would we expect them to be, as retention's predictors are and should be continuous at the cutoff). Thus, Black students are more likely to be retained *within* predominantly non-Black schools. To summarize, higher black retention rates are attributable to both school-level differences and differential responses to failure within predominantly non-Black schools. More generally, including school fixed effects indeed leaves our impact estimates essentially unchanged, including impact estimates by demographic subgroup.

5.2 Student-level Differences

Because students are not segregated by gender in New York City schools, gender heterogeneity in compliance cannot be driven by differences in school-level characteristics. We examine other observable student-level differences which may explain the gender gap. For instance, students who are more likely to be retained conditional on identical test scores might perform worse in other performance measures. This exercise is necessarily imperfect because we do not observe everything observed by teachers, principals, and parents. On the other hand, as researchers we *do* observe some key information unobserved by schools and

[10]Exam failure increases retention probability by 3.6 percentage points in low share schools, by 4.4 percentage points in middle share schools, and by 6.5 percentage points in high share schools.

parents: information on the *future* academic performance of students.

We compare average performance of girls and boys in baseline test scores, baseline attendance rate, and future test scores. (We depart from usual regression discontinuity analyses by *not* interpreting the jump scores at the failure threshold.[11]) As in previous studies, girls perform better than or as well as boys on average along these dimensions. Conditional on scoring identically on the baseline Math test in our retention window, girls also score better than boys on baseline English test, future Math test, and future English test. Moreover, they have similar slopes in the relationship between other test scores and baseline Math score as boys, implying that the predictive power of baseline test score is not different by gender (panel (a) of Figure 5).[12] Panel (b) shows that the slope of Black students above the cutoff is also similar to that of White students. It remains a puzzle that girls are about 25% more likely to be retained when they fail compared to boys, and that this gender gap is especially large for Whites.

Additionally, we examine whether heterogeneity in short-run benefits of retention can explain higher retention of girls and minorities conditional on test score. As retained and promoted students take different tests in subsequent years, it is fundamentally difficult to compare future test scores below and above the threshold. We attempt to address this issue by comparing same-grade test scores both in the baseline grade (i.e. test scores in the baseline year for the promoted versus test scores in the following year for the retained) and one grade above (i.e. test scores in the following year for the promoted versus test scores two years later for the retained). We find no obvious and robust heterogeneity in these future test scores, suggesting that it is unlikely that larger potential benefits on future performance for girls and minorities drive the differential retention decisions in the baseline.

5.3 Who Done It?

In this section, we focus on the role of principals. Teachers' perceptions of students can be based on their racial/ethnic and gender similarities to students [Dee, 2005]. Unfortunately, we do not observe the classroom to which students are assigned within grade and school (or the demographics of teachers). But according to the New York City Department of Education website:

Principals will review these portfolios in August and make a holistic promotion

[11]The particular exam taken is determined by a student's year in school, so the exam taken changes discontinuously at the threshold due to retention. If one is willing to ignore that potential compositional effect, there is an apparent increase in short-run academic performance due to retention, as has been found in previous literature. See Section 5.5.

[12]The relationship above the failure cutoff is easier to interpret because it is not affected by endogenous retention.

decision for each student. Superintendents will continue to review promotion appeals for cases in which a parent disagrees with the principal's decision.

As the final retention decision is made by principals and superintendents, we utilize data on school principal demographics, which come from a single 2008 cross-section of roughly 1,400 schools. This yields a subsample of 19,421 student records within 10 scale score of the cutoff. We consider whether the gender gap in retention varies by principal's gender.

Table 5 shows that the female-male difference in retention probability is pronounced in schools with female principals, while it essentially disappears in male principal schools. This is consistent with the findings from Hanna and Linden [2012] (admittedly in a radically different context): *"In fact, we observe the opposite, with discrimination against the low-caste children being driven by low-caste graders, and graders from the high-caste groups appearing not to discriminate at all even when controlling for the education and age of grader".* On gender, we do not know of an economics of education paper with a similar finding to ours. Bagues et al. [2015] argue that having women on faculty review committees in Italy and Spain, if anything, leads to fewer female faculty being promoted.

Additionally, we find that the Black-White gap in retention probability is large (11.3 percentage points versus 6.2 percentage points) in schools with White principals. The ethnicity gap disappears and is imprecisely estimated in Black principal schools, although this is partly due to the small number of White students in these schools. Because other (unobserved) characteristics of the school presumably vary by principal's observed characteristics, however, we characterize this pattern as descriptive.

5.4 Statistical Discrimination

Can the canonical theory of statistical discrimination [Phelps 1972; Arrow 1973; Aigner and Cain 1977] explain the heterogeneity we find? Through this lens, principals make the retention decision based on the current test score, x, which is a noisy signal for the true level of academic success in the next grade, q. That is, $x = q + u$ where $u \sim N(0, \sigma_u^2)$. In addition, principals have formed expectations of academic success for different demographic groups from experience: $q_s \sim N(\bar{q_s}, \sigma_{q,s}^2)$. Let $s = \{f, m\}$ denote the gender group. Then, the expected academic success of a student with test score x and gender s can be written as $\alpha_s x + (1 - \alpha_s)\bar{q_s}$, where $\alpha_s = \frac{\sigma_{q,s}^2}{\sigma_{q,s}^2 + \sigma_{u,s}^2}$. (Since the signal s may be more informative for one group than another, we let σ_u^2 to vary across groups and denote it as $\sigma_{u,s}^2$.) Intuitively, if the observed signal is noisy, α_s goes down and thus principals would put more weight on group mean and less weight weight on the observed signal. The female-male difference in the expected academic success conditional on scoring identically $x = k$ on the current test is:

$$E(q|x=k,s=f)-E(q|x=k,s=m)=(\alpha_f k+(1-\alpha_f)\bar{q_f})-(\alpha_m k+(1-\alpha_m)\bar{q_m})$$

If we assume that there is no difference in group mean, $\bar{q_f}=\bar{q}_m=\bar{q}$, the above equation reduces to $(\alpha_f-\alpha_m)(k-\bar{q})$. Therefore, if the current test score is a noisier measure for boys than for girls (i.e. $\alpha_f>\alpha_m$), the female-male difference in the expected academic success is negative for below average students ($k<\bar{q}$).

Turning to our data, we assume that principals have formed expectations of group performance based on previous year's Math test score and observe current year's Math test score. In our full sample, previous year's Math test score is slightly higher on average ($\bar{q_f}=681.7$ and $\bar{q_m}=680.7$) and more precise ($\sigma^2_{q,f}=33.5^2$ and $\sigma^2_{q,m}=34.6^2$) for girls. In addition, the current Math test score is noisier for boys ($\sigma^2_{u,f}=31.9^2$ and $\sigma^2_{u,m}=32.8^2$). Using these estimates, $E(q|x=k,s=f)-E(q|x=k,s=m)=-0.002k+1.998$. Evaluating this at the mean current Math test score below the cutoff $x=637.4$, we find that the female-male difference is small and rather positive ($-0.002(637.4)+1.998=0.7$). In this simple framework, the signal for boys is *not* noisy enough for our findings to be consistent with statistical discrimination.

5.5 BMI Impacts?

As in previous econometric studies of retention, considering the causal effects on subsequent academic performance is not straight-forward even with a valid instrument for retention. This is because the grade level of the exam students take in subsequent year is endogenous to retention decision. Therefore, it is difficult to distinguish the endogenous "exam taken" effect from the effect of retention on academic performance. We do not have a "silver bullet" solution to this problem.[13]

However, BMI testing does not vary by grade, and thus its evaluation is not compromised by endogenous retention. Moreover, as BMI percentiles vary by age in months and age itself is unaffected by retention, BMI percentiles are comparable for retained versus non-retained students. Furthermore, we observe BMI for all students, and have sufficient power to consider biometric impacts. Following a health economics literature on peer effects in BMI [Halliday and Kwak, 2009], timing of puberty and its responsiveness to social/environmental factors [Bharadwaj and Cullen, 2013], we test whether the higher probability of retention due to exam failure affects BMI in the following year. We instrument for retention with

[13]Mariano and Martorell [2013] address the endogeneity by estimating "external drift", which we do not pursue here.

scoring below the failure threshold and estimate the effect of retention on next year BMI using 2SLS. Table 6 shows that retention due to exam failure does not have a statistically significant impact on next year BMI, although point estimates indicate that grade retention might lower BMI relative to promoted peers. We conclude the peer effect on BMI is not large in our compliant sub-population, although our 2SLS estimates are somewhat imprecise.

5.6 Complier Characteristics

In this section, we take a more systematic approach to describing heterogeneity in compliance to the retention policy. The LATE theorem states that if treatment effects are heterogeneous, an instrument captures the causal effect for the sub-population of compliers (in our application, those who are retained as a result of exam failure). While it is not possible to identify individual compliers, it is possible to describe the distribution of complier characteristics. We estimate compliers' mean observable characteristics following Angrist and Pischke [2009], Almond and Doyle [2011].[14]

Table 7 shows that mean characteristics in fact vary substantially across different samples. Compliers are less likely to be Asian or White, while they are much more likely to be Black (49%) compared to those both in our retention window (38%) and in the full sample (31%). Insofar as race is concerned, compliance appears more selective than does scoring near the threshold. Turning to income, scoring near the threshold increases the share receiving a reduced-price lunch from 86 to 93%, while compliers are "only" 95% poor. Thus, performance on the test is more strongly related to income rather than how the test is used. Turning to gender, compliers are on average 48% female, versus 46% in our retention window (and 50% overall). The fraction obese is remarkably similar across these subgroups.

6 Discussion

The process by which retention decisions are made is often opaque despite utilization of standardized test scores and common thresholds. There is little systematic evidence on this "black box". We find both the magnitude and nature of this heterogeneity surprising. Why are younger students not more likely to be retained conditional on their exam score? In contrast, both race and gender help predict retention *conditional* on the baseline test score. Compliance with proficiency exams in New York City is thus selective. We find these descriptive patterns interesting *per se* and invite additional research on whether retention

[14] Curiously, seven years after Angrist and Pischke [2009] recounted a straight-forward approach to describe compliers, empirical economists seldom do. Recent methodological contributions in Angrist and Fernández-Val [2013], Dehejia et al. [2015], and Kowalski [2016] are notable exceptions.

decisions are "fair". Are girls and minorities over-retained? The need for such work is underscored by previous research (from other contexts where students can be tracked for longer time periods) that there may be long-term impacts on marginally-retained students [Jacob and Lefgren, 2009]. Such outcomes may be more important than the shorter-term benefits students show somewhat mechanically from repeating material they have seen in the previous year. Thus, it is not merely the case that the retention decision is perceived at the time as momentous by parents and students.

References

D. J. Aigner and G. G. Cain. Statistical theories of discrimination in labor markets. *Industrial and Labor Relations Review*, pages 175–187, 1977.

Douglas Almond and Joseph J. Doyle. After midnight: A regression discontinuity design in length of postpartum hospital stays. *American Economic Journal: Economic Policy*, 3(3): 1–34, 2011. doi: 10.1257/pol.3.3.1. URL `http://www.aeaweb.org/articles.php?doi=10.1257/pol.3.3.1`.

Douglas Almond, Ajin Lee, and Amy Ellen Schwartz. Impacts of classifying new york city students as overweight. *Proceedings of the National Academy of Sciences*, 113(13), 2016.

Joshua D. Angrist and Iván Fernández-Val. Extrapolate-ing: External validity and overidentification in the late framework. In Daron Acemoglu, Manuel Arellano, and Eddie Dekel, editors, *Advances in Economics and Econometrics, Tenth World Congress*, volume 3, chapter 11. Cambridge University Press, May 2013.

Joshua D. Angrist and Jörn-Steffen Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, Princeton, New Jersey, 2009.

K.J. Arrow. The theory of discrimination. In O. Ashenfelter and A. Rees, editors, *Discrimination in Labor Markets*, pages 3–33. Princeton University Press, 1973.

Manuel Bagues, Mauro Sylos-Labini, and Natalia Zinovyeva. Does the gender composition of scientific committees matter? *IZA Discussion Paper Series*, (9199), 2015.

Prashant Bharadwaj and Julie Berry Cullen. Coming of age: Timing of adolescence and gender identity formation. (Preliminary and incomplete. Do not cite without permission), September 2013.

Thomas S Dee. A teacher like me: Does race, ethnicity, or gender matter? *American Economic Review*, pages 158–165, 2005.

Thomas S. Dee, Brian A. Jacob, Jonah E. Rockoff, and Justin McCrary. Rules and discretion in the evaluation of students and schools: The case of the new york regents examinations. manuscript, Columbia Business School, 2011.

Rajeev Dehejia, Christian Pop-Eleches, and Cyrus Samii. From local to global: External validity in a fertility natural experiment. NBER Working Paper 21459, August 2015.

Andrew Gelman and Guido Imbens. Why high-order polynomials should not be used in regression discontinuity designs. Working Paper 20405, National Bureau of Economic Research, August 2014. URL http://www.nber.org/papers/w20405.pdf.

Timothy J. Halliday and Sally Kwak. Weight gain in adolescents and their peers. *Economics & Human Biology*, 7(2):181 – 190, 2009. ISSN 1570-677X. doi: DOI:10.1016/j.ehb.2009.05.002. URL http://www.sciencedirect.com/science/article/B73DX-4W91PW9-2/2/3efb336fa8c00a962aa98d27ad9696d5.

Rema N. Hanna and Leigh L. Linden. Discrimination in grading. *American Economic Journal: Economic Policy*, 4(4):146–68, 2012. doi: 10.1257/pol.4.4.146. URL http://www.aeaweb.org/articles.php?doi=10.1257/pol.4.4.146.

Elizabeth A. Harris and Ford Fessenden. 'opt out' becomes anti-test rallying cry in new york state. *The New York Times*, May 2015. URL http://www.nytimes.com/2015/05/21/nyregion/opt-out-movement-against-common-core-testing-grows-in-new-york-state.html?_r=0.

Brian A. Jacob and Lars Lefgren. Remedial Education and Student Achievement: A Regression-Discontinuity Analysis. *The Review of Economics and Statistics*, 86(1):226–244, February 2004. URL http://ideas.repec.org/a/tpr/restat/v86y2004i1p226-244.html.

Brian A. Jacob and Lars Lefgren. The Effect of Grade Retention on High School Completion. *American Economic Journal: Applied Economics*, 1(3):33–58, July 2009. URL http://ideas.repec.org/a/aea/aejapp/v1y2009i3p33-58.html.

Michal Kolesár and Christoph Rothe. Inference in regression discontinuity designs with a discrete running variable. manuscript, arXiv:1606.04086 [stat.AP], 2016.

Amanda E. Kowalski. Doing more when you're running late: Applying marginal treatment effect methods to examine treatment effect heterogeneity in experiments. NBER Working Paper 22363, June 2016.

Christina LiCalsi Labelle and David N. Figlio. The uneven implementation of universal school policies: Maternal education and florida's mandatory grade retention policy. Conference draft, Association for Education Finance and Policy (accessed 9/2015), September 2013. URL http://www.aefpweb.org/sites/default/files/webform/39th/Uneven%20Implementation%20of%20Universal%20School%20Policies.pdf.

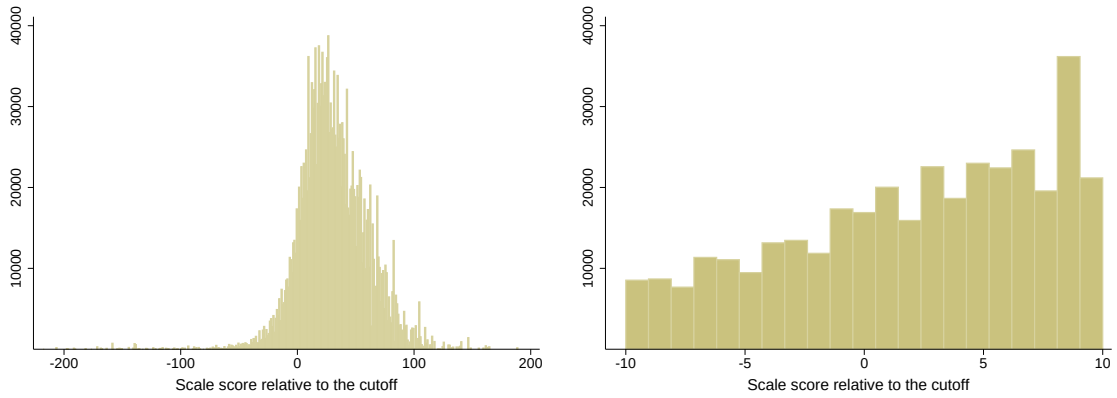
Louis T. Mariano and Paco Martorell. The academic effects of summer instruction and retention in new york city. *Educational Evaluation and Policy Analysis*, 35(1):96–117, 2013. doi: 10.3102/0162373712454327. URL http://epa.sagepub.com/content/35/1/96.abstract.

Edmund S. Phelps. The statistical theory of racism and sexism. *American Economic Review*, 62(4):659–661, September 1972.

Guido Schwerdt, Martin R. West, and Marcus A. Winters. The effects of test-based retention on student outcomes over time: Regression discontinuity evidence from florida. Working Paper 21509, National Bureau of Economic Research, August 2015. URL http://www.nber.org/papers/w21509.

Ellen M. Tomchin and James C. Impara. Unraveling teachers' beliefs about grade retention. *American Educational Research Journal*, 29(1):199–223, 1992.

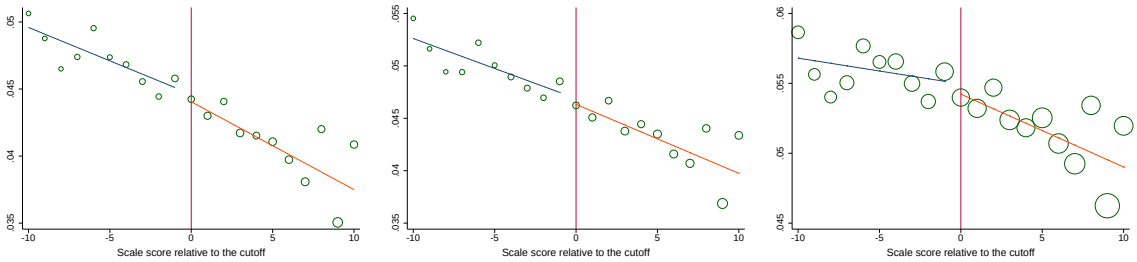
7 Figures



(a) All (b) Retention Window

Figure 1: Distribution of the running variable

Notes: The running variable is minimum of the Math and English test scores re-centered to zero at their own failure thresholds.



(a) All (b) Female only (c) Black only

Figure 2: Predicted probability of retention

Notes: We estimate the predicted probability of retention using gender, race/ethnicity, age in months, BMI, height, weight, free or reduced-price lunch eligibility, special education participation, and previous Math and English scale scores. Each circle plots mean predicted probability of retention within each one scale score bin. The size of the circle depends on the number of observations in each bin. Lines are the fitted values from a regressions of the predicted probability on the exam failure dummy, allowing for different slopes above and below the cutoff.

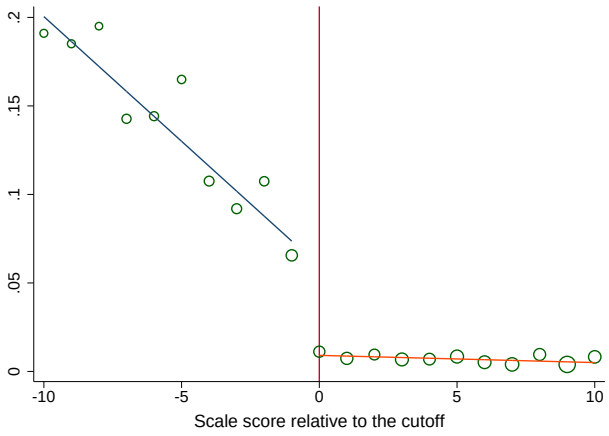
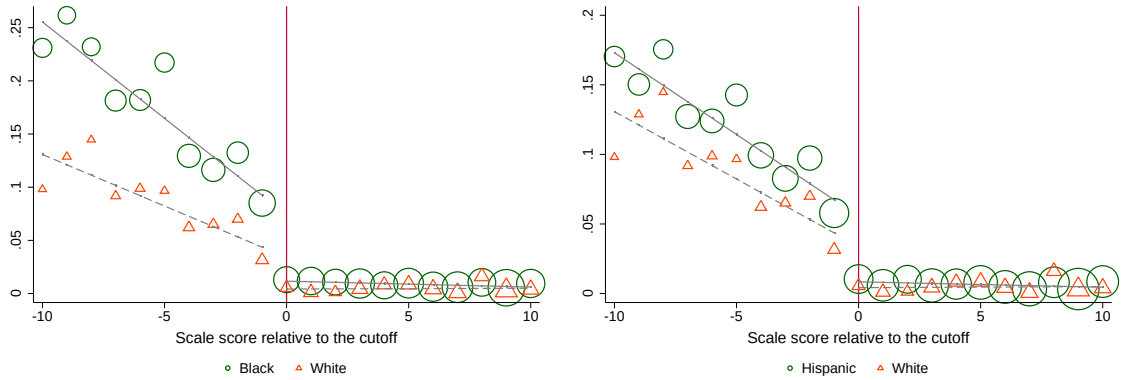


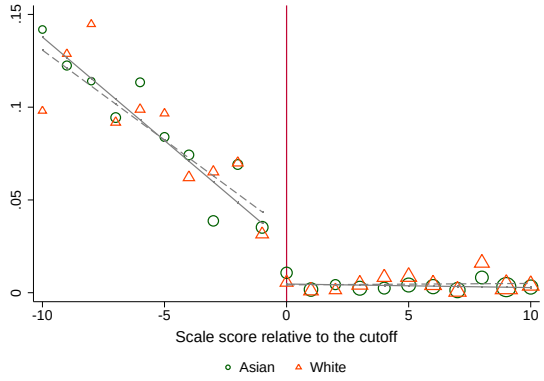
Figure 3: Probability of retention

Notes: Each circle plots mean probability of retention within each one scale score bin. The size of the circle depends on the number of observations in each bin. Lines are the fitted values from a regressions of a retention dummy on the exam failure dummy, allowing for different slopes above and below the cutoff.

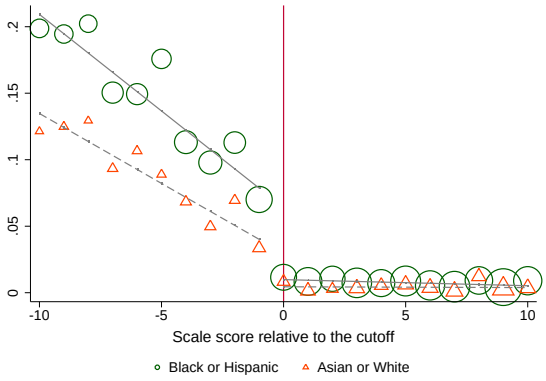


(a) Black vs. White

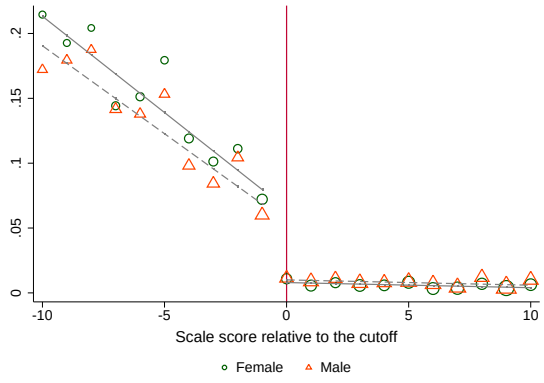
(b) Hispanic vs. White



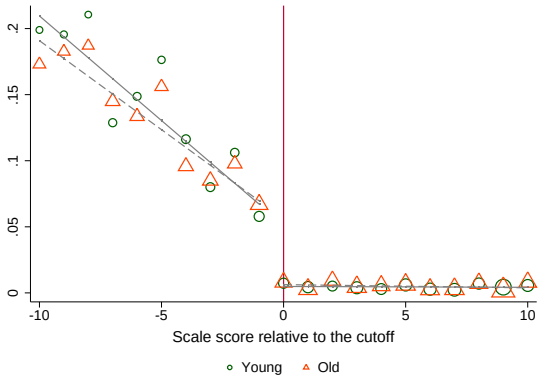
(c) Asian vs. White



(d) Black or Hispanic vs. Asian or White



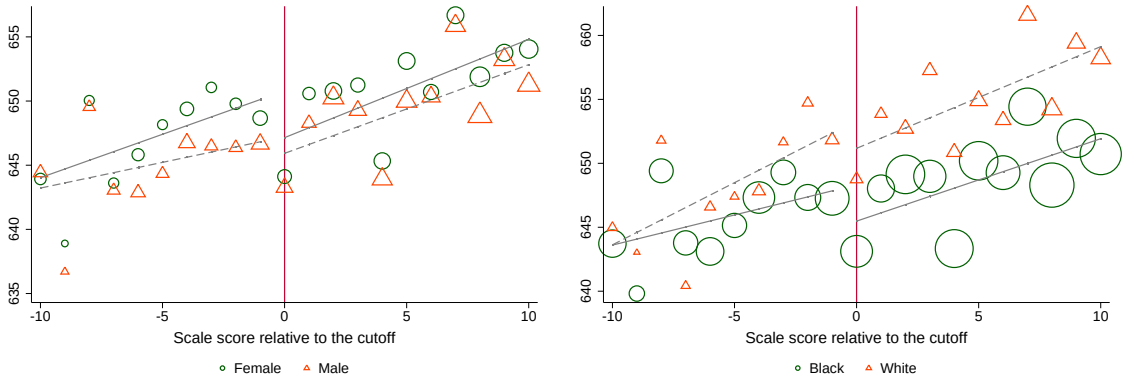
(e) Female vs. male



(f) Young vs. old

Figure 4: Heterogeneity in compliance

Notes: Each circle (or triangle) plots mean probability of retention within each one scale score bin. The size of the circle (or triangle) depends on the number of observations in each bin. Lines are the fitted values from a regression of a retention dummy on the exam failure dummy, allowing for different slopes above and below the cutoff. We divide each grade into three equal-sized groups based on age in months for grade. Panel (f) compares the youngest group with the oldest group.



(a) Female vs. male

(b) Black vs. White

Figure 5: Next year Math scale score

Notes: We use the re-centered baseline Math test score conditional on passing English as the running variable. Each circle (or triangle) plots mean Math test scores in the subsequent year within each one scale score bin. The size of the circle (or triangle) depends on the number of observations in each bin. Lines are the fitted values from a regressions of next year Math test score on the exam failure dummy, allowing for different slopes above and below the cutoff.

8 Tables

Table 1: Summary statistics

		Retention window	
	All	Above	Below
Asian	0.154	0.079	0.069
Black	0.307	0.376	0.388
Hispanic	0.394	0.469	0.485
White	0.142	0.073	0.054
Female	0.500	0.467	0.448
Free or reduced-price lunch	0.860	0.930	0.941
Age in months	133.1	134.5	135.7
Weight (lbs)	101.2	104.7	105.4
Height (inches)	58.2	58.4	58.5
Math level 1	0.049	0.000	0.376
Math level 2	0.245	0.712	0.482
Math level 3	0.461	0.261	0.132
Math level 4	0.245	0.027	0.010
English level 1	0.074	0.000	0.723
English level 2	0.370	0.957	0.256
English level 3	0.504	0.042	0.021
English level 4	0.053	0.001	0.000
Retention	0.018	0.007	0.128
N	1,507,700	77,543	168,047

Notes: Retention window indicates 10 scale score above and below the failure threshold.

Table 2: Effect of exam failure on the probability of retention

		A. Ethnicity				B. Gender	
	All	Asian	Black	Hispanic	White	Female	Male
Below cutoff	0.050***	0.021***	0.063***	0.047***	0.029***	0.057***	0.045***
	(0.002)	(0.006)	(0.004)	(0.003)	(0.007)	(0.003)	(0.003)
Observations	245,590	18,636	93,331	116,477	16,383	113,207	132,383
Mean below cutoff	0.128	0.080	0.162	0.113	0.080	0.137	0.122
Mean above cutoff	0.007	0.004	0.009	0.006	0.005	0.006	0.008

	C. Previous Math test score			D. Subsidized lunch	
	Low	Middle	High	Eligible	Not eligible
Below cutoff	0.092***	0.049***	0.021***	0.051***	0.038***
	(0.004)	(0.005)	(0.004)	(0.002)	(0.008)
Observations	55,800	53,868	55,312	223,661	15,956
Mean below cutoff	0.157	0.133	0.086	0.129	0.112
Mean above cutoff	0.010	0.007	0.005	0.007	0.009

Notes: Each column reports the estimated discontinuity in the probability of retention for different subsamples. We assume linear relationship between the retention probability and test scores, and allow for different slopes above and below the threshold. We control for year×grade×subject fixed effects. Robust standard errors are in the parentheses. In panel C, we divide the sample into three equal-sized groups based on last year's Math scale score. Subsidized lunch in panel D indicates free or reduced-price lunch eligibility.

Table 3: Effect of exam failure on the probability of retention

	A. Age for grade			B. Height for grade			C. Weight status			
	Young	Middle	Old	Short	Middle	Tall	Underweight	Healthy	Overweight	Obese
Below cutoff	0.045***	0.050***	0.050***	0.049***	0.045***	0.052***	0.063***	0.049***	0.042***	0.056***
	(0.004)	(0.004)	(0.003)	(0.004)	(0.004)	(0.003)	(0.013)	(0.003)	(0.005)	(0.004)
Observations	60,054	61,948	105,338	54,018	73,424	89,368	7,512	112,641	40,757	84,680
Mean below cutoff	0.128	0.131	0.123	0.123	0.130	0.124	0.116	0.129	0.126	0.130
Mean above cutoff	0.005	0.004	0.005	0.005	0.004	0.004	0.004	0.005	0.004	0.012

Notes: Each column reports the estimated discontinuity in the probability of retention for different subsamples. We assume linear relationship between the retention probability and test scores, and allow for different slopes above and below the threshold. We control for year×grade×subject fixed effects. We divide each grade into three equal-sized groups based on age in months for grade (panel A) and height for grade (panel B). Each student's body mass index (BMI) is classified to be underweight, healthy, overweight, and obese based on age- and sex-specific BMI cutoffs from Centers for Disease Control.

Table 4: Heterogeneity by school's proportion of Black students

	Low (mean=5%)				High (mean=60%)			
	Asian	Black	Hispanic	White	Asian	Black	Hispanic	White
A. Without school fixed effects								
Below cutoff	0.027***	0.055***	0.039***	0.017**	0.023	0.066***	0.064***	0.059**
	(0.008)	(0.017)	(0.005)	(0.008)	(0.017)	(0.005)	(0.007)	(0.024)
B. With school fixed effects								
Below cutoff	0.026***	0.058***	0.040***	0.017**	0.022	0.067***	0.063***	0.065**
	(0.008)	(0.017)	(0.005)	(0.008)	(0.019)	(0.004)	(0.007)	(0.028)
Observations	11,475	4,133	40,165	10,612	2,988	68,828	26,948	1,797
Mean below cutoff	0.069	0.123	0.094	0.070	0.102	0.167	0.132	0.105
Mean above cutoff	0.003	0.008	0.004	0.004	0.005	0.010	0.010	0.009

Notes: We divide schools into three equal-sized groups by schools' proportion of Black students. The mean proportion of Black students is 5% in low share schools. It is 60% in high share schools. Each column reports the estimated discontinuity in the probability of retention for different race/ethnicity groups. We assume linear relationship between the retention probability and test scores, and allow for different slopes above and below the threshold. We control for year×grade×subject fixed effects. Robust standard errors are in the parentheses.

Table 5: Gender heterogeneity in retention probability by principal's gender

	Female principal		Male principal	
Student:	Female	Male	Female	Male
Below cutoff	0.135***	0.081***	0.070***	0.083***
	(0.017)	(0.015)	(0.024)	(0.022)
Observations	6,481	7,188	2,689	3,063
Mean below cutoff	0.133	0.123	0.101	0.120
Mean above cutoff	0.005	0.006	0.005	0.011

Notes: We utilize data on principal gender from 2007-2008. First two columns compare the estimated discontinuity in the probability of retention by student gender in schools with a female principal. Last two columns examine heterogeneity by student gender in schools with a male principal. We assume linear relationship between the retention probability and test scores, and allow for different slopes above and below the threshold. We control for year×grade×subject fixed effects. Robust standard errors are in the parentheses.

Table 6: Effect of retention on next year BMI

		By ethnicity				By gender	
	All	Asian	Black	Hispanic	White	Female	Male
Retention	-0.413	4.773	-0.143	-0.981	3.275	-1.019	0.321
	(0.958)	(7.809)	(1.283)	(1.435)	(6.366)	(1.273)	(1.450)
Observations	208,916	17,171	76,138	100,075	14,841	96,119	112,797
Mean below cutoff	21.9	20.1	21.9	22.2	21.3	22.0	21.7
Mean above cutoff	21.8	20.0	21.9	22.1	21.1	21.9	21.6

Notes: Each column reports the estimated effect of retention on next year BMI for different subsamples. We instrument for retention with scoring below the failure threshold and estimate the effect of retention on next year BMI using 2SLS. We assume linear relationship between next year BMI and test scores, and allow for different slopes above and below the threshold. We control for year×grade×subject fixed effects. Robust standard errors are in the parentheses.

Table 7: Mean characteristics

Characteristic	Complier $E(X\|D_1 = 1, D_0 = 0)$	Retention window (N=245,590)	All (N=1,507,700)
Asian	0.043	0.076	0.154
Black	0.492	0.380	0.307
Hispanic	0.428	0.474	0.394
White	0.033	0.067	0.142
Female	0.482	0.461	0.501
Free or reduced-price lunch	0.950	0.933	0.860
Age in months	135.2	134.9	133.1
Weight (lbs)	105.1	104.9	101.2
Height (inches)	58.5	58.5	58.2
Obese	0.338	0.345	0.332
N		245,590	1,507,700

Notes: First column summarizes mean characteristics of compliers following Angrist and Pischke [2009], Almond and Doyle [2011].