# Large Impacts of Grade Retention in New York City*

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

September 25, 2023

**Abstract**

The longer-term consequences of grade retention are largely unknown. We consider students who fail New York's annual statewide proficiency exam, given in grades 3-8. We argue that a naive regression discontinuity approach confounds the effect of retention with that of special education, attenuating retention's estimated impact. A regression kink design allows us to isolate retention's impact, which we find to be large. Retention increases high school graduation odds by 33 percentage points and the likelihood of graduating from a selective "screened" school by 28 percentage points. Minority students drive this LATE effect and have larger IV impacts of retention as well. Identically-scoring White students tend to avoid retention when they fail and are more likely to be classified as special education in response to identically-poor test performance.

JEL Codes: I21, I24, I28

Keywords: grade retention; standardized test; regression discontinuity; regression kink

*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 Michah W. Rothbart and Meryle Weinstein for assistance with the data.

†Columbia University and NBER: da2152@columbia.edu

‡Michigan State University: leeajin@msu.edu

§University of Delaware: aeschwar@udel.edu

# 1 Introduction

Are students retained too often? Existing research consistently finds short run improvements in academic outcomes following retention, including to grades and test scores. *A priori*, this improvement can be mechanical and fleeting, particularly as academic material is being repeated in the first year following retention, but not thereafter. This short run effect may be salient to and readily observed by schools, and taken as evidence for retaining students at the margin (Johnny will do better next year if retained). Not only is it unclear how these short run effects correspond to longer-term endpoints, longer-term endpoints are likely more valued by society, parents, and students themselves. For example, graduating from high school is likely more important than performing well in fourth grade. Teachers/schools likely have a blind spot insofar as longer-term outcomes are concerned, lacking observations across schools let alone longitudinal data.

Empirically, there is evidence for over-retention. In seminal study of Chicago public schools students, Jacob and Lefgren (2004) found positive short run effects of among third graders who were retained. But impacts were mixed for sixth graders. Moreover, Jacob and Lefgren (2009) followed up students to see whether they graduated from high school. They found that retaining sixth grade students had no detectable benefit to high school completion. And that retaining eighth grade students actually increased high school dropout rates. In general, much less is known about such key long-term impacts than the short-term benefits of retention.

Differences in retention by demographic subgroup raise additional concerns. In the nation's largest public school system (New York City), the retention decision is likewise informed by student performance on the annual statewide proficiency exam. But the test score is not deterministic: failing merely puts the student "at risk" for being retained. Conditional on test performance, minority and female students are much more likely to be retained when they fail than when non-Hispanic White or male students fail. Blacks students are 73% more likely to be retained when they fail than whites, and girls are 51% more likely to be retained than boys when they fail. Although the retention decision is keying off the same standardized test, even how identical performance translates into retention differs tremendously. Moreover, these differences correspond to familiar and persistent patterns of discrimination in the US.

We deploy a new identification strategy to address these issues. We focus on the arguably more important long-term impacts of retention in New York City. As in Jacob and Lefgren (2009), we assess these alongside the shorter term benefits about which there is more con-

sensus. But we also uncover a potential pitfall with the standard regression discontinuity analysis of retention: it fails the exclusion restriction, at least in New York. When a student fails the annual proficiency exam, retention is not the only treatment that is triggered (discontinuously). We also see that students are more likely to be classified for special education services when they fail the proficiency exam. This means that the discontinuous increase in retention is confounded by a discontinuous increase in special education at the same cutoff score. Decision makers considering retention need to know whether it *per se* is detrimental or beneficial. I.E., our reduced form effect of failing an exam is not particularly informative to decision makers.

We resolve this identification challenge by pursuing a regression kink design. We leverage the fact that special education classification is not additionally affected by the test score *conditional on failing*. In contrast, the retention decision continues to be affected by "how badly" the student failed the exam, and this differs strikingly from the underlying relationship between test score and retention just above the failure threshold (where the correlation is essentially zero). Indeed, scoring 3 points below the threshold has roughly the same impact on retention likelihood as the discontinuity itself. By shoring up the exclusion restriction, we can isolate the impact of retention in New York.

Finally, we consider longer term impacts of retention in context of marked retention heterogeneity. Yes, minority and female students are retained more even when they have the same test score as Whites or males. Is it also the case that "how badly" a minority or female student fails the exam matters more? Our overall IV impact for retention is identified for the subpopulations whose retention is most affected by failure, the compliant subpopulation. This compliant subpopulation looks quite different from the overall student population, and even those select students scoring near the failure threshold. Given those "first stage" differences and the potential for heterogenous treatment effects, what is the impact of retention on outcomes of interest for subgroups that are disproportionately retained? In retaining some demographic groups much more than others – and meanwhile lacking access to longitudinal data through high school – are primary and middle schools somehow still "getting it right"?

The remainder of the paper is organized as follows. Section 2 describes the retention policy in New York City and provides a brief discussion of previous literature. Section 3 discusses our data and two research designs (regression discontinuity and regression kink designs) along with the main estimation results. In this section, we also examine the validity of our identification strategy and impact heterogeneity. Section 4 concludes.

# 2 Background and Literature Review

## 2.1 Retention Policy in New York City

Particularly for a highly consequential decision, the process determining retention is often opaque. This is despite frequent utilization of standardized tests to inform the decision, which might be expected to increase transparency. In New York, passing essentially guarantees promotion, while failing merely puts one "in jeopardy" of being retained. Only about 16% of students are retained in our data when they fail the annual proficiency exam. School principals make the final promotion decision.

Students in grades 3-8 take New York State Math and English Language Arts (ELA) tests each spring. Students receive the scale score that determines their performance level for each subject. The scores are classified into four levels with Level 1 being not meeting state learning standards, or "failing," Level 2 being partially meeting state learning standards, Level 3 being meeting state learning standards, and Level 4 being exceeding state learning standards. The thresholds for these four levels are specific to each subject, and they vary every year. Scoring Level 2 in both tests essentially guarantees promotion, while students who score Level 1 in either test are at risk of being retained.

## 2.2 Special Education in New York City

Students with disabilities who need special education services have Individualized Education Programs (IEPs) in New York City. Multiple steps are involved in the provision of special education services. First, a parent or a Department of Education (DOE) school official must refer a child for special education services. Second, after receiving informed consent, the DOE evaluates a child in all areas related to a suspected disability. Third, after the evaluation is complete, an IEP team (which includes a parent) reviews information from the evaluation and determines a child's eligibility for special education services. If found eligible, the team develops an IEP, and a child is placed for special education services.

Importantly, while there is no single factor that qualifies a child for special education services, test scores are reviewed and taken into consideration for the IEP eligibility determination. Empirically, we find that whether a child failed a proficiency exam has a significant impact on special education placement.

### 2.3 Brief Literature Review

Seminal work by Jacob and Lefgren (2004, 2009) was described in the Introduction. The short-term effects of summer school and retention in New York City have been studied using a similar regression discontinuity (RD) design. Mariano and Martorell (2013) focused on fifth graders failing the annual exam in 2004/2005-2005/2006 and found positive effects of summer school on English achievement. Additionally, they estimated the effect of retention on sixth grade outcomes, accounting for cohort-over-cohort differences ("external drift"), and found large and positive effects on both Math and English performance. Matsudaira (2008) similarly found positive effects of summer school in New York City and highlighted that the average effects mask considerable heterogeneity in short-term impacts of grade retention. We build on this literature by allowing for heterogeneity in the impacts of retention on both short and longer term endpoints.

This paper exploits a regression kink (RK) design (Nielsen et al., 2010; Card et al., 2015; Turner, 2017) to isolate the effect of retention. Fortunately for us, the provision of special education services independently considers the binary outcome of whether a student failed a proficiency exam. Thus, information on test scores aside from whether they failed does not have an additional impact on the provision of special education services. In other words, the gradient in the provision of special education services with respect to test scores is identical on either side of the threshold. Such heuristics or rules of thumb have previously been used by researchers as the basis of an identification strategy. For example, Almond, Doyle, Kowalski, and Williams (2010) identified marginal returns to medical care using the "very low birth weight" threshold at 1,500 grams. They find that being below 1,500 grams is salient, even when exact birth weight is observable to clinicians. When time and attention are scarce, more complex information is often overlooked (Maćkowiak et al., 2023), which helps us in identifying the effect of retention net of special education.

## 3 Data, Design, and Results

### 3.1 Data

We obtain statewide proficiency test results and retention data for students in grades 3-8 between academic year (AY) 2007/2008 and 2010/2011 from the New York City Department of Education. We merge these data to high school graduation outcomes between AY 2011/2012 and 2018/2019. Graduation outcomes are measured within 4, 5, or 6 years

Table 1: Summary statistics

| | (1) Analysis sample | (2) Full sample |
|---|---|---|
| Retention in year t+1 | 0.037 | 0.015 |
| Special education in year t+1 | 0.246 | 0.123 |
| Graduated in 4 years | 0.583 | 0.758 |
| Dropped out in 4 years | 0.137 | 0.074 |
| Still enrolled in 4 years | 0.301 | 0.190 |
| Enrolled in any screened school in ninth grade | 0.428 | 0.494 |
| Graduated from any screened school in 4 years | 0.383 | 0.465 |
| Enrolled in any testing school in ninth grade | 0.001 | 0.070 |
| Graduated from any testing school in 4 years | 0.001 | 0.068 |
| Female | 0.465 | 0.505 |
| Asian | 0.088 | 0.171 |
| Black | 0.374 | 0.296 |
| Hispanic | 0.472 | 0.395 |
| White | 0.066 | 0.135 |
| Subsidized lunch | 0.940 | 0.854 |
| Observations | 110432 | 1105315 |

*Notes:* The analysis sample refers to students with 10 scale scores below and above the threshold. We restrict our sample to students in grades 3-8 in academic years between 2007/2008 and 2010/2011 matched with high school graduation outcomes measured in academic years between 2011/2012 and 2018/2019. For the analysis sample, we also focus on the first time students show up in the RD window.

of students entering the ninth grade. We focus on 4-year graduation outcomes in our main analysis.

We additionally use the school choice data for 2007-2013 to characterize high school programs based on admission methods. In particular, we focus on two selective admission methods: screened and testing. Testing programs are most selective, and most of these schools admit students based on the High School Application & Specialized High Schools Admissions Test (SHSAT). Screened programs admit top-performing students in group order, admitting the highest-performing students first. As schools can have multiple programs with different admission criteria, we classify schools into screened or testing schools if any of their programs admit students using screened admissions or testing. There are nine testing schools, which include eight SHSAT high schools and one audition high school, and 169 out of 472 schools use screened admissions.

Given that we examine long-term outcomes, we first test whether failing differentially affects the probability of leaving the New York City public school system, but no evidence of that (Appendix Table A.1). The probability of showing up in our data is smooth across

the threshold both in the following year and in ninth grade. Additionally, because the probability of appearing in 4-year graduation data can change discontinuity at the threshold by virtue of differential retention rates, we drop students who would not be in 12th grade by AY 2018/2019 if retained once. With this sample restriction, we find that the probability of having a record of graduation outcomes is smooth across the threshold. In the final sample, we have 1,105,315 student records with 4-year graduation outcomes. Of those, we focus on 110,432 students whose scale scores are within 10 scale scores from the level 1 threshold for the main analysis.

Table 1 presents summary statistics. In our analysis sample near the threshold, 3.7% are retained and 24.6% are classified as special education the following year. These are about twice as high as the rates of retention and special education in the full sample. The 4-year graduation rate is also lower in the analysis sample: 58.3% vs. 75.8% in the full sample. About 14% of students in our analysis sample drop out high school in 4 years, compared to 7.4% in the full sample. About 43% of students in our analysis sample attend a screened school in ninth grade, slightly lower than 49% in the full sample. Enrollment in testing schools is rare, 0.1% in the analysis sample and 7% in the full sample. Our analysis sample is less likely to be female, more likely to be a racial minority, and more likely to be eligible for subsidized lunch.

## 3.2 Naive RDD

To consider the effects of grade retention, we first take a conventional approach where we exploit the increased risk of grade retention below the Level 1 threshold using a regression discontinuity design (RDD). Specifically, we estimate the following regressions:

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

$$Y_{igs,t+x} = \beta_0 + \beta_1 1[X_{igst} < 0] + \beta_2 X_{igst} + \beta_3 X_{igst} 1[X_{igst} < 0] + \eta_{gst} + \nu_{igst} \quad (2)$$

where $i$ is individual, $g$ is grade, $s$ is subject, and $t$ is year. $R_{igs,t+1}$ is the indicator for being retained in the following year. $Y_{igs,t+x}$ indicates dependent variables such as short-term test performance and long-term graduation outcomes. For the running variable ($X_{igst}$), we re-center each Math and ELA scale score at the threshold and take the minimum of the two. This way, we capture the effect of failing either test. We include a linear spine of the running variable to control for the trend separately below and above the threshold. Additionally, we

include year×grade×subject dummies (threshold fixed effects), $\theta_{gst}$ and $\eta_{gst}$, given that the threshold changes within these cells. We report robust standard errors.

The key coefficients of interest are $\alpha_1$, the first stage effect of failing on retention, and $\beta_1$, which captures the reduced-form effect of failing on short- and long-term outcomes. Finally, we estimate $\frac{\hat{\beta_1}}{\hat{\alpha_1}}$, the ratio of the reduced-form estimate to the first stage estimate, using a two-stage least squares (2SLS) approach, the effect of retention in our naive RDD. We refer to these estimates as RD-IV estimates.

Figure 1(a) depicts the first stage relationship between test scores and the probability of retention. Consistent with the retention policy, the probability of retention increases just below the threshold. Column (1) of Table 2 summarizes the first stage estimates: failing increases the retention probability by 3.8 percentage points. Using the 2SLS approach, we then examine whether the increased retention due to exam failure is associated with changes in graduation outcomes. Columns (1)-(2) of Table 3 report the RD-IV estimates on 4-year graduation rates and dropout rates in Panel A. Here, we find no effect of retention on these long-term outcomes.

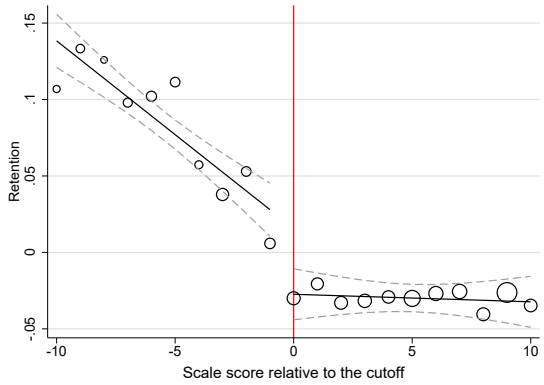
In contrast, we find significant short-term benefits of retention in terms of next year's test scores, consistent with previous literature. Appendix Table A.2 summarizes these effects. However, short-term effects may dissipate and may be confounded by the difference in the specific test taken across the threshold – the exam one takes is endogenous to previous failure. Moreover, while we find no effect of retention on graduation outcomes using the RDD, the exclusion restriction of the standard RDD fails to hold in our setting. When a student fails the exam, we find that the probability of receiving special education services also increases the following year (Figure 1(b) and Table 2 column (2)). This failure of the exclusion restriction makes the binary indicator for exam failure an invalid instrument for retention. Thus, we employ an alternative identification strategy and further examine the impact of retention on long-term graduation outcomes.

## 3.3 RK Design and Impacts

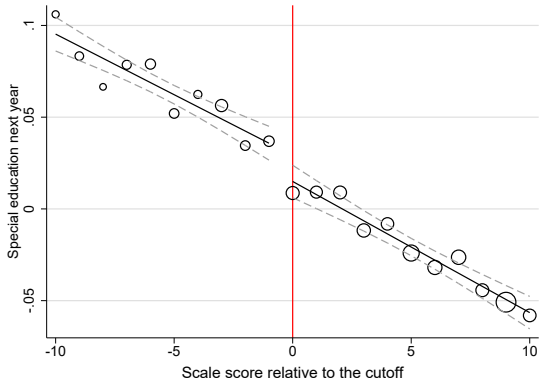
To address the issues above, we pay attention to how test scores are used *conditional on failing*. As Figure 1(b) shows, the relationship between test score and special education is linear and stable across the distribution of test scores, and importantly it does not change below the threshold. In other words, conditional on failing, "how badly" the student failed does not affect special education classification further. This is not the case for retention. Figure 1(a) shows that the retention decision is heavily influenced by test performance below

Figure 1: Effects of test performance near the Level 1 threshold

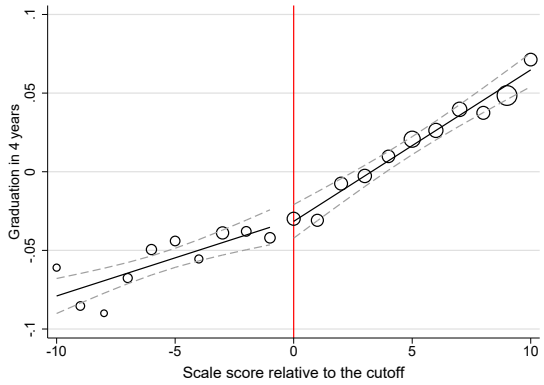
(a) Retention



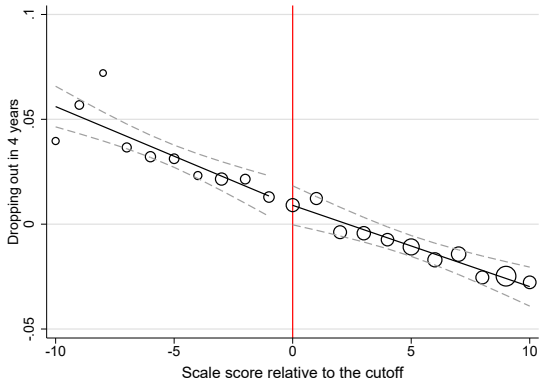

(b) Special education



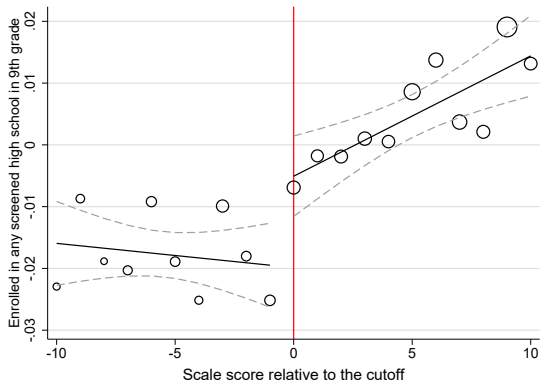

(c) Graduated



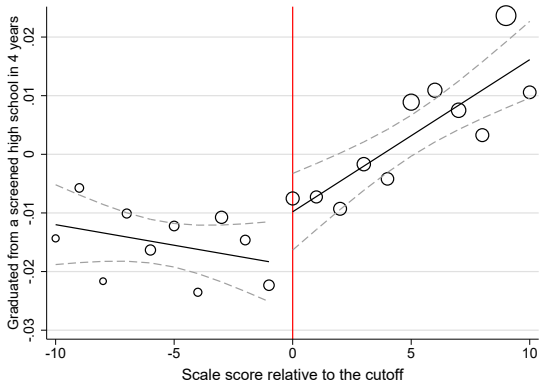

(d) Dropped out




(e) Screened school enrollment in ninth grade




(f) Graduated from a screened school




*Notes:* Hollow circles plot regression-adjusted means of each outcome for one scale score bins. The size of the circle depends on the number of observations in each bin. Solid 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. Dashed lines are 95% confidence intervals.

Table 2: The first-stage and reduced-form effects of exam failure

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| | First stage | | Reduced form | |
| | Retention | Special education | Graduated | Dropped out |
| Below cutoff | 0.038*** | 0.015** | 0.002 | -0.000 |
| | (0.003) | (0.006) | (0.007) | (0.005) |
| Running variable | -0.000** | -0.007*** | 0.010*** | -0.004*** |
| | (0.000) | (0.000) | (0.001) | (0.000) |
| Below cutoff * running variable | -0.014*** | 0.000 | -0.004*** | -0.001 |
| | (0.001) | (0.001) | (0.001) | (0.001) |
| Observations | 110432 | 110413 | 110432 | 110432 |
| Mean below cutoff | 0.116 | 0.314 | 0.516 | 0.174 |
| Mean above cutoff | 0.006 | 0.219 | 0.610 | 0.122 |

*Notes:* Columns (1)-(2) report the estimates from equation (1). Columns (3)-(4) report the estimates from equation (2). Robust standard errors are in the parentheses. * p<0.10, ** p<0.05, *** p<0.01.

the threshold. In contrast to the relationship between test score and retention above the threshold (essentially none), we find a large negative relationship below the threshold. For example, scoring 3 points below the threshold has roughly the same impact on the probability of retention as the discontinuity at the threshold itself.

We leverage this stark change in slopes below the threshold to separately identify the impact of grade retention, using a regression kink design (RKD). Formally, the change in slopes below the threshold is captured by $\alpha_3$ in equation (1) and $\beta_3$ in equation (3). Table 2 shows that scoring one more point below the threshold increases the probability of retention by 1.4 percentage points. Critically, scoring one point lower below the failure threshold has no impact on special education classification. While failure itself has no impact on graduation and dropout rates at the threshold, we find that scoring one point below the threshold *increases* graduation rates by 0.4 percentage points (column (3) of Table 2). Using 2SLS, we then compute the RK-IV estimates, or $\frac{\hat{\beta_3}}{\hat{\alpha_3}}$, the ratio of the slope changes. Columns (3)-(4) of Table 3 report these estimates in Panel A. In contrast to the naive RDD estimates, we find that retention increases graduation rates by 33 percentage points using this approach. See Appendix Section B.1 for a complete set of graduation outcomes.

Moreover, we examine the effects of retention on enrollment in and graduation from selective high schools and find big impacts. Focusing on our preferred RK-IV estimates, Panel B of Table 3 shows that retention increases attending a screened school in ninth grade

Table 3: Effects of retention on high school enrollment and 4-year graduation outcomes

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| | RD-IV | | RK-IV | |
| | Graduated | Dropped out | Graduated | Dropped out |
| Panel A. 4-year graduation outcomes from any high school | | | | |
| Retention | 0.044 | -0.012 | 0.331*** | 0.067 |
| | (0.172) | (0.127) | (0.087) | (0.064) |
| Observations | 110432 | 110432 | 110432 | 110432 |
| Mean below cutoff | 0.516 | 0.174 | 0.516 | 0.174 |
| Mean above cutoff | 0.610 | 0.122 | 0.610 | 0.122 |
| F statistic | 142.313 | 142.313 | 411.855 | 411.855 |
| | Enrollment | Graduation | Enrollment | Graduation |
| Panel B. Enrollment in and graduation from screened high schools | | | | |
| Retention | -0.374** | -0.210 | 0.194** | 0.279*** |
| | (0.186) | (0.173) | (0.086) | (0.086) |
| Observations | 105553 | 110202 | 105553 | 110202 |
| Mean below cutoff | 0.409 | 0.366 | 0.409 | 0.366 |
| Mean above cutoff | 0.435 | 0.390 | 0.435 | 0.390 |
| F statistic | 128.504 | 140.651 | 418.174 | 413.777 |

*Notes:* Columns (1)-(2) report the RD-IV estimates. Columns (3)-(4) report the RK-IV estimates. Panel A summarizes the effects of retention on 4-year graduation outcomes from any high school. Panel B reports the estimates on ninth grade enrollment in and 4-year graduation from any screened high school. Refer to the main text for more information. Robust standard errors are in the parentheses. * p<0.10, ** p<0.05, *** p<0.01.

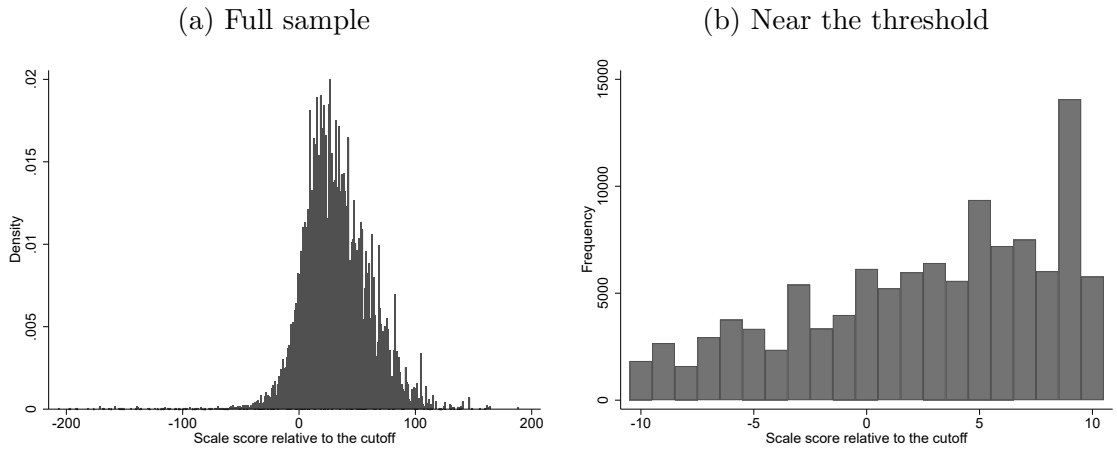
by 19 percentage points and graduating from one by 28 percentage points. This result further highlights potential benefits of retention in New York City. We additionally find that retention increases the likelihood of attending and graduating from a testing school. See Appendix Section B.2 for more details.

Thus, addressing the confounding of retention's impact with that of special education from the conventional RD design reveals a large, positive benefit of retention.

### 3.3.1 Validity of Our Identification Strategy

An identifying assumption for both RDD and RKD designs is that the density of the running variable is smooth across the threshold. Figure 2 shows the distribution in the full sample and in the analysis sample. Visually, we find no evidence of sorting to either side

Figure 2: Histograms of the running variable




*Notes:* Panel (a) shows the distribution of the running variable for the full sample. Panel (b) shows the distribution of the running variable for the analysis sample, 10 scale scores below and above the threshold.

of the threshold. A formal test (McCrary, 2008) also does not reject the smoothness of the density across the failure threshold.

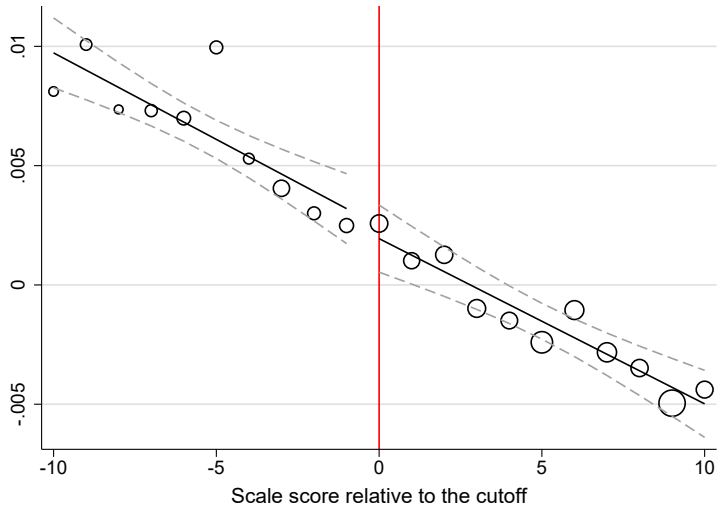
Additionally, we test whether pre-determined characteristics are balanced across the threshold. We calculate a predicted probability of retention based on various observable characteristics: gender, race/ethnicity, age in months, BMI, height, weight, subsidized lunch, special education participation in the baseline year, previous Math and ELA test scores, and school fixed effects. We find that the predicted probability of retention is smooth, and there is no change in the slope of this index of covariates across the threshold (Figure 3), further validating the our identification approach.

## 3.4 IV Impact Heterogeneity

Whether one is retained when one fails differs markedly by subgroup (see section 3.5 below). In this section, we consider how long-term effects of retention vary across subgroups. Note that, even in subgroups, there is no kink in the probability of special education placement, while the relationship between test score and retention probability becomes significantly steeper below the threshold (Appendix Table A.6). This allows us to use the RK-IV approach by subgroups.

Figure 4 summarizes the effects by race/ethnicity, and the corresponding estimates are reported in Appendix Table A.3. We find that the increase in graduation rates is large and precisely estimated for Black and Hispanic students, although the differences are not

Figure 3: Balance of baseline covariates




*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, previous Math and English scale scores, and school fixed effects. Each circle plots regression-adjusted 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. Solid 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. Dashed lines are 95% confidence intervals.

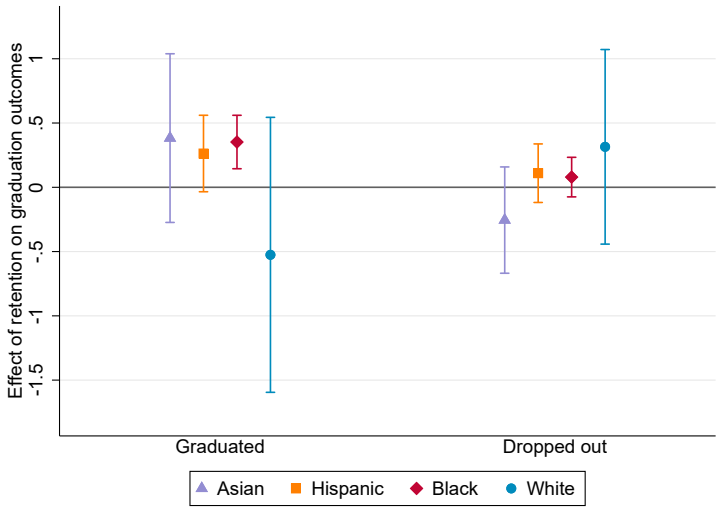
significant across different categories of race/ethnicity.

We also consider heterogeneity by various student characteristics, such as gender, subsidized lunch eligibility, age for grade, height for grade, weight categories. The positive effect of retention on graduation is larger for girls than for boys, although the estimates are not statistically different (Appendix Table A.4). We also find that the increased graduation rates are much more pronounced for students who are eligible for free or reduced-priced lunch (Appendix Table A.5). Note, however, that most students in our sample are eligible for subsidized lunch. We generally find little heterogeneity across student age and physical characteristics. See Appendix Section B.3 for a complete set of RK-IV estimates by subgroups. Overall, our results suggest that students with relatively disadvantaged backgrounds may be more likely to benefit from grade retention in a long run.

## 3.5 First Stage Heterogeneity

We now examine how first stage effects for both retention and special education vary across student characteristics. We find that the discontinuous increase in retention at the

Figure 4: Effects of retention on graduation outcomes (RK-IV estimates) by race/ethnicity




threshold is larger for girls than for boys, for racial minorities, and for those who are eligible for subsidized lunch (Appendix Table A.6). This suggests that the same information (exam failure) may have different implications for students with different characteristics. In addition, the change in slopes below the threshold (i.e., the role of test score below the threshold) is the most pronounced for Black students.

Test performance also has different impacts on special education across student characteristics. For example, we find that boys are more likely to receive special education services when they fail compared to girls. Although insignificant, the point estimate also suggests that white students are more likely to be placed in special education due to exam failure.

The differential impacts of test performance on retention and special education classification across students motivates us to further understand who is driving our results. While we cannot identify who the compliers are, we can estimate their average characteristics without additional assumptions (Angrist and Pischke, 2009). Empirically, this entails calculating the ratio of the first stage effect for the specific subgroup to the first stage effect for the overall sample. For example, to examine whether the RDD compliers are more or less likely to be girls, we calculate the ratio of the estimated $\alpha_1$ (from equation (1)) using the subgroup of girls to the estimated $\alpha_1$ using the whole analysis sample. For RKK compliers, we similarly calculate the ratios of the $\alpha_3$ coefficients.

Table 4 reports these complier ratios along with the analysis sample mean. We find that RDD compliers are 22% more likely to be female and 17% more likely to be Black compared to the average in the analysis sample, which is consistent qualitatively with the heterogeneity in

Table 4: Complier characteristics

| Characteristic | Analysis sample mean | RDD complier ratio | RKD complier ratio |
|---|---|---|---|
| Female | 0.465 | 1.219 | 0.988 |
| Asian | 0.088 | 0.404 | 0.779 |
| Hispanic | 0.472 | 0.973 | 0.818 |
| Black | 0.374 | 1.170 | 1.368 |
| White | 0.066 | 0.670 | 0.640 |
| Subsidized lunch | 0.940 | 1.040 | 1.002 |
| Young for grade | 0.334 | 1.111 | 0.995 |
| Old for grade | 0.320 | 0.958 | 0.933 |
| Short for grade | 0.300 | 1.065 | 0.918 |
| Tall for grade | 0.382 | 0.860 | 1.049 |
| Underweight | 0.033 | 0.962 | 1.008 |
| Healthy | 0.523 | 0.951 | 1.017 |
| Overweight | 0.444 | 1.064 | 0.979 |
| Obese | 0.251 | 1.180 | 0.924 |

*Notes:* The RDD complier ratio is the ratio of the estimated $\alpha_1$ (from equation (1)) using the respective subgroup to the estimated $\alpha_1$ using the whole analysis sample. For RKD compliers, we similarly calculate the ratios of $\alpha_3$s.

the first stage estimates. RKD compliers are even more likely to be Black (36.8%), suggesting that the worse performance below the threshold is particularly meaningful to the retention decision for Black students.

Anecdotally, we have heard that New York City schools target younger students for retention, i.e. those who were close to the age cutoff for school entry. This would make some intuitive sense. Interestingly, we do not find strong evidence that young students who fail the exam are more likely to be retained. Indeed, for the RKD, young students are if anything less likely to be retained due to poor exam scores (Appendix Table A.7).

# 4 Conclusion

Roughly 50,000 students are "held back" in New York City each year. Despite being momentous and costly, there is surprisingly little evidence on the long-term consequences of grade retention. We argue that progression in high school is more important to students and parents than shorter-term outcomes, like performance in fourth grade, which unambiguously improve with retention. This means that at present, retention practice and policies exist largely in ignorance of key evidence. In the United States as a whole, around 2% or 1 million

students are retained each year.

In New York, special education seems to be used — at least in part — as a substitute for retention when a student fails. This substitutability seems to be more of an implicit practice than a stated policy, and differs by student demographics. White male students are particularly unlikely to be retained and more likely to be promoted to next grade with special education services. Econometrically, this means a conventional RD design will tend to understate the benefits of retention in New York.

Nevertheless, New York's statewide proficiency exam constitutes a shock to perception of a given student's appropriateness for their grade. Leveraging this shock in a RK design, we find retention increases the likelihood of completing high school. The magnitude of this impact is large, increasing high school completion by 33 percentage points. Minority students drive this large benefit, and indeed have a stronger first stage response to variation in exam score below the failure cutoff. In contrast to the cautionary evidence from Chicago (Jacob and Lefgren, 2009), we find no evidence of over-retention in New York.

# References


Almond, D., J. J. Doyle, Jr., A. E. Kowalski, and H. Williams (2010). Estimating marginal returns to medical care: Evidence from at-risk newborns. *The Quarterly Journal of Economics 125*(2), 591–634.

Angrist, J. D. and J.-S. Pischke (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, New Jersey: Princeton University Press.

Card, D., D. S. Lee, Z. Pei, and A. Weber (2015). Inference on causal effects in a generalized regression kink design. *Econometrica 83*(6), 2453–2483.

Jacob, B. A. and L. Lefgren (2004, February). Remedial Education and Student Achievement: A Regression-Discontinuity Analysis. *The Review of Economics and Statistics 86*(1), 226–244.

Jacob, B. A. and L. Lefgren (2009, July). The Effect of Grade Retention on High School Completion. *American Economic Journal: Applied Economics 1*(3), 33–58.

Maćkowiak, B., F. Matějka, and M. Wiederholt (2023). Rational inattention: A review. *Journal of Economic Literature 61*(1), 226–273.

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

Matsudaira, J. D. (2008). Mandatory summer school and student achievement. *Journal of Econometrics 142*(2), 829–850.

McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics 142*(2), 698 – 714. The regression discontinuity design: Theory and applications.

Nielsen, H. S., T. Sørensen, and C. Taber (2010). Estimating the effect of student aid on college enrollment: Evidence from a government grant policy reform. *American Economic Journal: Economic Policy 2*(2), 185–215.

Turner, L. J. (2017). The economic incidence of federal student grant aid. *University of Maryland, College Park, MD 1000*.

# Appendix A. Appendix Tables

Table A.1: Testing differential exit across the threshold

| Probability of having records | In the following year | In ninth grade | For 4-year graduation outcomes | For 5-year graduation outcomes | For 6-year graduation outcomes |
|---|---|---|---|---|---|
| Below cutoff | 0.000 | -0.002 | 0.005 | 0.002 | -0.007 |
| | (0.001) | (0.004) | (0.005) | (0.005) | (0.005) |
| Observations | 137391 | 137391 | 137391 | 137391 | 137391 |
| Mean below cutoff | 0.989 | 0.878 | 0.800 | 0.763 | 0.649 |
| Mean above cutoff | 0.992 | 0.901 | 0.805 | 0.786 | 0.661 |

*Notes:* Each column tests differential exits from the New York City public school system across the threshold at a different point in time, by examining the probability of having non-missing records in each time period as an outcome. Robust standard errors are in the parentheses. * p<0.10, ** p<0.05, *** p<0.01.

Table A.2: Effect of retention on next year's test scores

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| | A. RD-IV | | B. RK-IV | |
| | Math | English | Math | English |
| Retention | 29.373*** | 18.119*** | 39.829*** | 21.238*** |
| | (8.222) | (4.393) | (5.133) | (2.749) |
| Observations | 84168 | 84155 | 84168 | 84155 |
| Mean below cutoff | 656.164 | 642.697 | 656.164 | 642.697 |
| Mean above cutoff | 662.764 | 648.507 | 662.764 | 648.507 |
| F statistic | 133.825 | 135.627 | 280.011 | 280.181 |

*Notes:* Panel A reports the RD-IV estimates. Panel B reports the RK-IV estimates. Refer to the main text for more information. Robust standard errors are in the parentheses. * p<0.10, ** p<0.05, *** p<0.01.

Table A.3: Effect on 4-year graduation outcomes (RK-IV estimates) by race

| | A. Graduated | | | | B. Dropped out | | | |
|---|---|---|---|---|---|---|---|---|
| | Asian | Hispanic | Black | White | Asian | Hispanic | Black | White |
| Retention | 0.383 | 0.263* | 0.352*** | -0.525 | -0.255 | 0.110 | 0.079 | 0.315 |
| | (0.335) | (0.152) | (0.106) | (0.546) | (0.211) | (0.116) | (0.078) | (0.386) |
| Observations | 9696 | 52177 | 41293 | 7245 | 9696 | 52177 | 41293 | 7245 |
| Mean below cutoff | 0.721 | 0.497 | 0.479 | 0.626 | 0.089 | 0.187 | 0.182 | 0.138 |
| Mean above cutoff | 0.783 | 0.590 | 0.574 | 0.707 | 0.063 | 0.135 | 0.122 | 0.106 |
| F statistic | 32.456 | 151.184 | 226.524 | 13.937 | 32.456 | 151.184 | 226.524 | 13.937 |

*Notes:* Panel A reports the RK-IV estimates on graduation rates by race. Panel B reports the RK-IV estimates on dropout rates by race. Robust standard errors are in the parentheses. * p<0.10, ** p<0.05, *** p<0.01.

Table A.4: Effect on 4-year graduation outcomes (RK-IV estimates) by gender

| | A. Graduated | | B. Dropped out | |
|---|---|---|---|---|
| | Female | Male | Female | Male |
| Retention | 0.416*** | 0.265** | 0.054 | 0.082 |
| | (0.131) | (0.116) | (0.096) | (0.086) |
| Observations | 51371 | 59059 | 51371 | 59059 |
| Mean below cutoff | 0.549 | 0.489 | 0.173 | 0.175 |
| Mean above cutoff | 0.639 | 0.584 | 0.116 | 0.127 |
| F statistic | 179.763 | 233.900 | 179.763 | 233.900 |

*Notes:* Panel A reports the RK-IV estimates on graduation rates by race. Panel B reports the RK-IV estimates on dropout rates by race. Robust standard errors are in the parentheses. * p<0.10, ** p<0.05, *** p<0.01.

Table A.5: Effect on 4-year graduation outcomes (RK-IV estimates) by subsidized lunch eligibility

| | A. Graduated | | B. Dropped out | |
|---|---|---|---|---|
| | Not eligible | Eligible | Not eligible | Eligible |
| Retention | -0.728* | 0.368*** | 0.235 | 0.046 |
| | (0.405) | (0.090) | (0.242) | (0.067) |
| Observations | 5835 | 91827 | 5835 | 91827 |
| Mean below cutoff | 0.648 | 0.507 | 0.093 | 0.180 |
| Mean above cutoff | 0.741 | 0.599 | 0.067 | 0.127 |
| F statistic | 18.892 | 390.821 | 18.892 | 390.821 |

*Notes:* Panel A reports the RK-IV estimates on graduation rates by race. Panel B reports the RK-IV estimates on dropout rates by race. Robust standard errors are in the parentheses. * p<0.10, ** p<0.05, *** p<0.01.

Table A.6: Effects of test performance on retention and special education

| | | A. Gender | | B. Race/ethnicity | | | | C. Subsidized lunch | |
|---|---|---|---|---|---|---|---|---|---|
| | All | Female | Male | Asian | Hispanic | Black | White | Not eligible | Eligible |
| Panel A. Dependent variable: probability of retention | | | | | | | | | |
| Below cutoff | 0.038*** | 0.047*** | 0.031*** | 0.016* | 0.037*** | 0.045*** | 0.026** | 0.011 | 0.036*** |
| | (0.003) | (0.005) | (0.004) | (0.009) | (0.004) | (0.006) | (0.011) | (0.013) | (0.003) |
| Running variable | -0.000** | -0.000* | -0.000 | -0.000 | -0.000 | -0.000** | 0.000 | -0.001*** | -0.000* |
| | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) |
| Below cutoff * running variable | -0.014*** | -0.013*** | -0.014*** | -0.011*** | -0.011*** | -0.019*** | -0.009*** | -0.013*** | -0.014*** |
| | (0.001) | (0.001) | (0.001) | (0.002) | (0.001) | (0.001) | (0.002) | (0.003) | (0.001) |
| Observations | 110432 | 51371 | 59059 | 9696 | 52177 | 41293 | 7245 | 5835 | 91827 |
| Mean below cutoff | 0.116 | 0.122 | 0.110 | 0.075 | 0.103 | 0.148 | 0.073 | 0.097 | 0.117 |
| Mean above cutoff | 0.006 | 0.005 | 0.007 | 0.003 | 0.006 | 0.007 | 0.006 | 0.007 | 0.006 |
| Panel B. Dependent variable: probability of receiving special education services next year | | | | | | | | | |
| Below cutoff | 0.015** | 0.001 | 0.029*** | -0.019 | 0.018** | 0.016* | 0.024 | -0.000 | 0.013** |
| | (0.006) | (0.008) | (0.008) | (0.018) | (0.009) | (0.010) | (0.027) | (0.030) | (0.007) |
| Running variable | -0.007*** | -0.007*** | -0.007*** | -0.007*** | -0.008*** | -0.007*** | -0.010*** | -0.010*** | -0.007*** |
| | (0.000) | (0.001) | (0.001) | (0.001) | (0.001) | (0.001) | (0.002) | (0.002) | (0.001) |
| Below cutoff * running variable | 0.000 | -0.001 | 0.002 | 0.005 | 0.002 | -0.003 | 0.005 | -0.005 | -0.000 |
| | (0.001) | (0.001) | (0.001) | (0.003) | (0.002) | (0.002) | (0.005) | (0.005) | (0.001) |
| Observations | 110413 | 51356 | 59055 | 9694 | 52170 | 41286 | 7242 | 5834 | 91812 |
| Mean below cutoff | 0.314 | 0.259 | 0.361 | 0.194 | 0.325 | 0.307 | 0.467 | 0.430 | 0.310 |
| Mean above cutoff | 0.219 | 0.179 | 0.255 | 0.165 | 0.228 | 0.198 | 0.338 | 0.294 | 0.213 |

*Notes:* Panel A reports the first stage effect of failing on retention by student characteristics. Panel B reports the first stage effect of failing on special education placement by student characteristics. Robust standard errors are in the parentheses. * p<0.10, ** p<0.05, *** p<0.01.

Table A.7: Effects of test performance on retention and special education

| | A. Age for grade | | | B. Height for grade | | | C. Weight status | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Young | Middle | Old | Short | Middle | Tall | Underweight | Healthy | Overweight | Obese |
| Panel A. Dependent variable: probability of retention | | | | | | | | | | |
| Below cutoff | 0.043*** | 0.035*** | 0.037*** | 0.041*** | 0.043*** | 0.033*** | 0.037* | 0.037*** | 0.035*** | 0.046*** |
| | (0.006) | (0.006) | (0.005) | (0.006) | (0.006) | (0.005) | (0.019) | (0.005) | (0.008) | (0.007) |
| Running variable | -0.000 | -0.000* | -0.000 | -0.000 | -0.000 | 0.000 | 0.000 | -0.000* | 0.000 | 0.000 |
| | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) | (0.001) | (0.000) | (0.000) | (0.000) |
| Below cutoff * running variable | -0.013*** | -0.015*** | -0.013*** | -0.013*** | -0.014*** | -0.015*** | -0.014*** | -0.014*** | -0.015*** | -0.013*** |
| | (0.001) | (0.001) | (0.001) | (0.001) | (0.001) | (0.001) | (0.004) | (0.001) | (0.002) | (0.001) |
| Observations | 36722 | 38042 | 35116 | 28550 | 30276 | 36370 | 3136 | 49651 | 18339 | 23840 |
| Mean below cutoff | 0.120 | 0.118 | 0.108 | 0.115 | 0.120 | 0.111 | 0.111 | 0.117 | 0.113 | 0.114 |
| Mean above cutoff | 0.006 | 0.006 | 0.006 | 0.005 | 0.004 | 0.005 | 0.002 | 0.005 | 0.004 | 0.004 |
| Panel B. Dependent variable: probability of receiving special education services next year | | | | | | | | | | |
| Below cutoff | 0.026*** | 0.003 | 0.018* | 0.009 | 0.007 | 0.031*** | 0.008 | 0.011 | 0.023 | 0.024* |
| | (0.010) | (0.010) | (0.011) | (0.011) | (0.011) | (0.011) | (0.036) | (0.009) | (0.015) | (0.013) |
| Running variable | -0.004*** | -0.007*** | -0.007*** | -0.007*** | -0.008*** | -0.007*** | -0.012*** | -0.006*** | -0.006*** | -0.009*** |
| | (0.001) | (0.001) | (0.001) | (0.001) | (0.001) | (0.001) | (0.003) | (0.001) | (0.001) | (0.001) |
| Below cutoff * running variable | -0.000 | -0.001 | 0.001 | 0.000 | 0.001 | 0.001 | 0.007 | 0.000 | -0.001 | 0.003 |
| | (0.002) | (0.002) | (0.002) | (0.002) | (0.002) | (0.002) | (0.006) | (0.002) | (0.003) | (0.002) |
| Observations | 36715 | 38037 | 35109 | 28547 | 30272 | 36362 | 3136 | 49643 | 18335 | 23837 |
| Mean below cutoff | 0.259 | 0.277 | 0.399 | 0.267 | 0.299 | 0.339 | 0.304 | 0.288 | 0.302 | 0.342 |
| Mean above cutoff | 0.177 | 0.181 | 0.314 | 0.187 | 0.206 | 0.239 | 0.200 | 0.207 | 0.209 | 0.228 |

*Notes:* Panel A reports the first stage effect of failing on retention by student characteristics. Panel B reports the first stage effect of failing on special education placement by student characteristics. Robust standard errors are in the parentheses. * p<0.10, ** p<0.05, *** p<0.01.

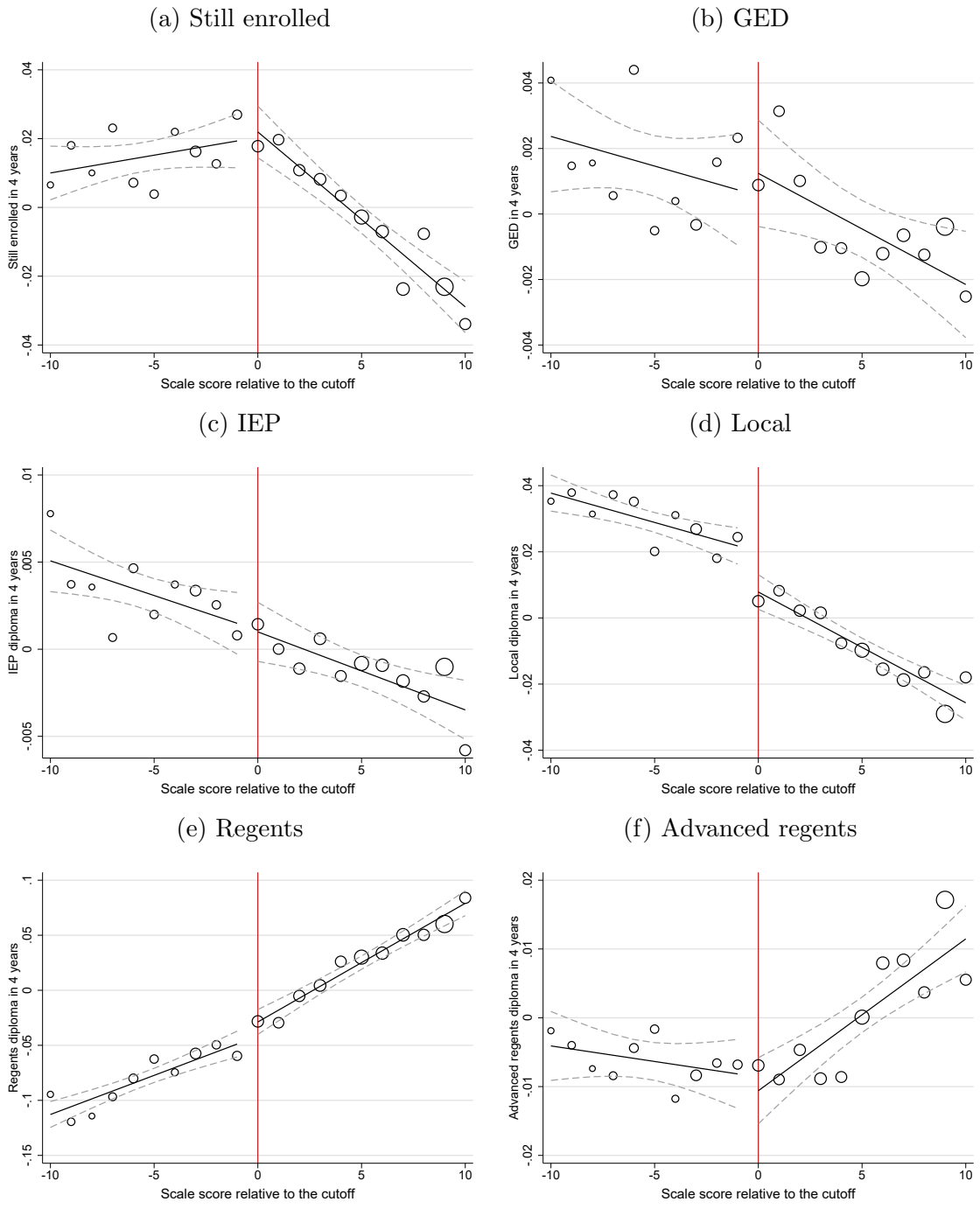
# Appendix B. Additional results

## B.1 Graduation outcomes

In this section, we examine the full set of graduation outcomes. Figure B.1 depicts the relationship between test scores and six other 4-year graduation outcomes besides graduation and dropouts. Table B.1 reports the RK-IV estimates for all 4-year graduation outcomes in Panel A. We find that the increased graduation effect is associated with more Regents and advanced Regents diplomas and fewer local diplomas. Regents and advanced Regents diplomas indicate students' additional skills in math, science, and languages. To earn the Regents diploma, students need to pass 5 Regents exams. To earn the advanced Regents diploma, students need to pass 9 Regents exam. Certain students have an option to earn the local diploma by passing 5 Regents exams with lower exam scores.

Panels B and C show the RK-IV estimates for 5-year and 6-year graduation outcomes, respectively. While the estimates become less precise, we generally find the same patterns. Retention increases graduation rates, and the effect is driven by more Regents or advanced Regents diplomas.

Figure B.1: Effects of test performance near the Level 1 threshold on other 4-year graduation outcomes




*Notes:* Hollow circles plot regression-adjusted means of each outcome for one scale score bins. The size of the circle depends on the number of observations in each bin. Solid lines are the fitted values from equation (2). Dashed lines are 95% confidence intervals.

Table B.1: Graduation outcomes, RK-IV estimates

| | | | | Type of diploma | | | | |
|---|---|---|---|---|---|---|---|---|
| | Graduated | Dropped out | Still enrolled | GED | IEP | Local | Regents | Advanced Regents |
| Panel A. 4-year outcomes | | | | | | | | |
| Retention | 0.331*** | 0.067 | -0.457*** | -0.008 | -0.002 | -0.128** | 0.237*** | 0.220*** |
| | (0.087) | (0.064) | (0.083) | (0.022) | (0.017) | (0.056) | (0.084) | (0.033) |
| Observations | 110432 | 110432 | 110432 | 110432 | 110432 | 110432 | 110432 | 110432 |
| Mean below cutoff | 0.516 | 0.174 | 0.323 | 0.016 | 0.010 | 0.126 | 0.362 | 0.029 |
| Mean above cutoff | 0.610 | 0.122 | 0.292 | 0.013 | 0.006 | 0.082 | 0.485 | 0.042 |
| F statistic | 411.855 | 411.855 | 411.855 | 411.855 | 411.855 | 411.855 | 411.855 | 411.855 |
| Panel B. 5-year outcomes | | | | | | | | |
| Retention | 0.263*** | -0.086 | -0.191*** | -0.012 | 0.007 | -0.195*** | 0.196* | 0.261*** |
| | (0.100) | (0.087) | (0.070) | (0.028) | (0.022) | (0.074) | (0.101) | (0.040) |
| Observations | 105040 | 105040 | 105040 | 105040 | 105040 | 105040 | 105040 | 105040 |
| Mean below cutoff | 0.602 | 0.242 | 0.132 | 0.018 | 0.012 | 0.157 | 0.415 | 0.030 |
| Mean above cutoff | 0.691 | 0.180 | 0.112 | 0.016 | 0.007 | 0.103 | 0.544 | 0.043 |
| F statistic | 335.147 | 335.147 | 335.147 | 335.147 | 335.147 | 335.147 | 335.147 | 335.147 |
| Panel C. 6-year outcomes | | | | | | | | |
| Retention | 0.152 | -0.140 | -0.011 | -0.019 | 0.025 | -0.197** | 0.102 | 0.247*** |
| | (0.109) | (0.101) | (0.059) | (0.030) | (0.027) | (0.084) | (0.113) | (0.043) |
| Observations | 87496 | 87496 | 87496 | 87496 | 87496 | 87496 | 87496 | 87496 |
| Mean below cutoff | 0.626 | 0.270 | 0.071 | 0.017 | 0.014 | 0.165 | 0.432 | 0.029 |
| Mean above cutoff | 0.715 | 0.203 | 0.055 | 0.016 | 0.008 | 0.108 | 0.565 | 0.042 |
| F statistic | 299.357 | 299.357 | 299.357 | 299.357 | 299.357 | 299.357 | 299.357 | 299.357 |

*Notes:* Panel A reports the reduced-form effects of failing on 4-year graduation outcomes. Panel B reports RD-IV estimates. Panel C reports RK-IV estimates. Robust standard errors are in the parentheses. * p<0.10, ** p<0.05, *** p<0.01.

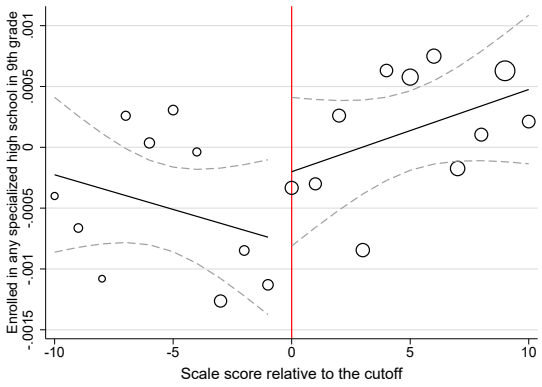
## B.2 Effects of retention on selective high school enrollment and graduation

In addition to screened high schools, we consider whether retention affects attending and graduating from specialized high schools that use testing for admission. These schools include eight SHSAT high schools[1] and one audition high school.[2]

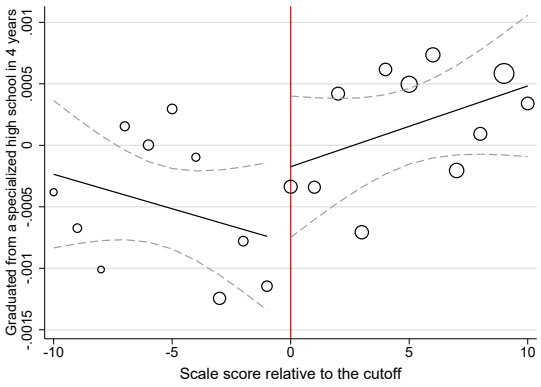
Given that very few students in our analysis sample attend these high schools, Figure B.2 presents much noisier relationships between test scores and these outcomes. That said, Table B.2 shows that retention significantly increases ninth grade enrollment in and graduation odds from these schools, further highlighting the large impact of retention on academic achievement.

Figure B.2: Effect of failing on enrollment in and graduation from a testing high school

(a) Enrollment in ninth grade




(b) Graduation in 4 years




[1]The SHSAT high schools are:

- The Bronx High School of Science
- The Brooklyn Latin School
- Brooklyn Technical High School
- High School for Mathematics, Science and Engineering at City College of New York
- High School of American Studies at Lehman College
- Queens High School for the Sciences at York College
- Staten Island Technical High School
- Stuyvesant High School

[2]Fiorello H. LaGuardia High School of Music & Art and Performing Arts.

Table B.2: Effect of failing on enrollment and graduation from a testing high school

| | A. RD-IV | | B. RK-IV | |
|---|---|---|---|---|
| | Enrollment | Graduation | Enrollment | Graduation |
| Retention | -0.022** | -0.022** | 0.013*** | 0.012*** |
| | (0.009) | (0.004) | (0.004) | (0.004) |
| Observations | 105553 | 110202 | 105553 | 110202 |
| Mean below cutoff | 0.001 | 0.001 | 0.001 | 0.001 |
| Mean above cutoff | 0.002 | 0.002 | 0.002 | 0.002 |
| F statistic | 128.504 | 140.651 | 418.174 | 413.777 |

*Notes:* Enrollment is measured in ninth grade. Graduation rate is measured in 4 years from ninth grade. Robust standard errors are in the parentheses. * p<0.10, ** p<0.05, *** p<0.01.

## B.3 IV Impact Heterogeneity for all subgroups and all 4-year graduation outcomes

In this section, we consider various subgroups and report the RK-IV estimates for all 4-year graduation outcomes. As discussed in the main text, we find that the positive effect of retention is driven by racial minorities and those who are eligible for subsidized lunch. We find little heterogeneity across student age and physical characteristics. Even within subgroups, we find that the increased graduation rate reflects more Regents and advanced Regents diplomas and fewer local diplomas.

Figure B.3: Effects of retention on graduation outcomes (RK-IV estimates) by race/ethnicity

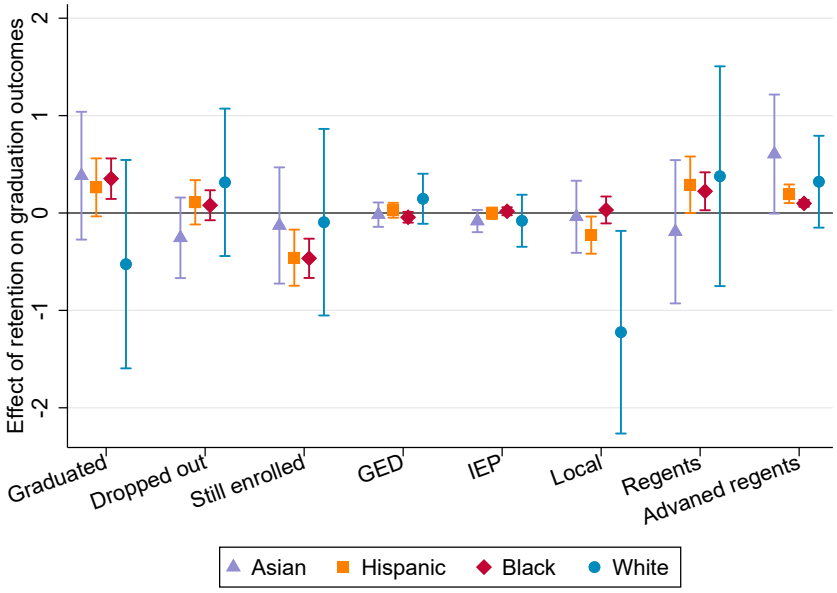


Table B.3: Effect on 4-year graduation outcomes (RK-IV estimates) by race

| | | | | Type of diploma | | | | |
|---|---|---|---|---|---|---|---|---|
| | Graduated | Dropped out | Still enrolled | GEP | IEP | Local | Regents | Advanced regents |
| Asian | | | | | | | | |
| Retention | 0.383 | -0.255 | -0.128 | -0.017 | -0.083 | -0.039 | -0.193 | 0.605* |
| | (0.335) | (0.211) | (0.304) | (0.064) | (0.058) | (0.189) | (0.375) | (0.312) |
| Observations | 9696 | 9696 | 9696 | 9696 | 9696 | 9696 | 9696 | 9696 |
| Mean below cutoff | 0.721 | 0.089 | 0.211 | 0.007 | 0.007 | 0.069 | 0.467 | 0.185 |
| Mean above cutoff | 0.783 | 0.063 | 0.182 | 0.007 | 0.003 | 0.041 | 0.542 | 0.201 |
| F statistic | 32.456 | 32.456 | 32.456 | 32.456 | 32.456 | 32.456 | 32.456 | 32.456 |
| Hispanic | | | | | | | | |
| Retention | 0.263* | 0.110 | -0.459*** | 0.028 | -0.003 | -0.227** | 0.291** | 0.197*** |
| | (0.152) | (0.116) | (0.147) | (0.039) | (0.030) | (0.097) | (0.148) | (0.049) |
| Observations | 52177 | 52177 | 52177 | 52177 | 52177 | 52177 | 52177 | 52177 |
| Mean below cutoff | 0.497 | 0.187 | 0.332 | 0.016 | 0.011 | 0.122 | 0.357 | 0.018 |
| Mean above cutoff | 0.590 | 0.135 | 0.298 | 0.013 | 0.005 | 0.081 | 0.477 | 0.032 |
| F statistic | 151.184 | 151.184 | 151.184 | 151.184 | 151.184 | 151.184 | 151.184 | 151.184 |
| Black | | | | | | | | |
| Retention | 0.352*** | 0.079 | -0.465*** | -0.044 | 0.018 | 0.031 | 0.223** | 0.099*** |
| | (0.106) | (0.078) | (0.103) | (0.028) | (0.021) | (0.070) | (0.099) | (0.018) |
| Observations | 41293 | 41293 | 41293 | 41293 | 41293 | 41293 | 41293 | 41293 |
| Mean below cutoff | 0.479 | 0.182 | 0.349 | 0.018 | 0.011 | 0.136 | 0.338 | 0.004 |
| Mean above cutoff | 0.574 | 0.122 | 0.327 | 0.015 | 0.007 | 0.087 | 0.474 | 0.013 |
| F statistic | 226.524 | 226.524 | 226.524 | 226.524 | 226.524 | 226.524 | 226.524 | 226.524 |
| White | | | | | | | | |
| Retention | -0.525 | 0.315 | -0.095 | 0.146 | -0.080 | -1.224** | 0.377 | 0.321 |
| | (0.546) | (0.386) | (0.489) | (0.131) | (0.136) | (0.531) | (0.576) | (0.241) |
| Observations | 7245 | 7245 | 7245 | 7245 | 7245 | 7245 | 7245 | 7245 |
| Mean below cutoff | 0.626 | 0.138 | 0.243 | 0.010 | 0.013 | 0.179 | 0.404 | 0.043 |
| Mean above cutoff | 0.707 | 0.106 | 0.203 | 0.010 | 0.005 | 0.116 | 0.528 | 0.064 |
| F statistic | 13.937 | 13.937 | 13.937 | 13.937 | 13.937 | 13.937 | 13.937 | 13.937 |

*Notes:* Panel A reports the RK-IV estimates on graduation rates by race. Panel B reports the RK-IV estimates on dropout rates by race. Robust standard errors are in the parentheses. * p<0.10, ** p<0.05, *** p<0.01.

Figure B.4: Effects of retention on graduation outcomes (RK-IV estimates) by gender

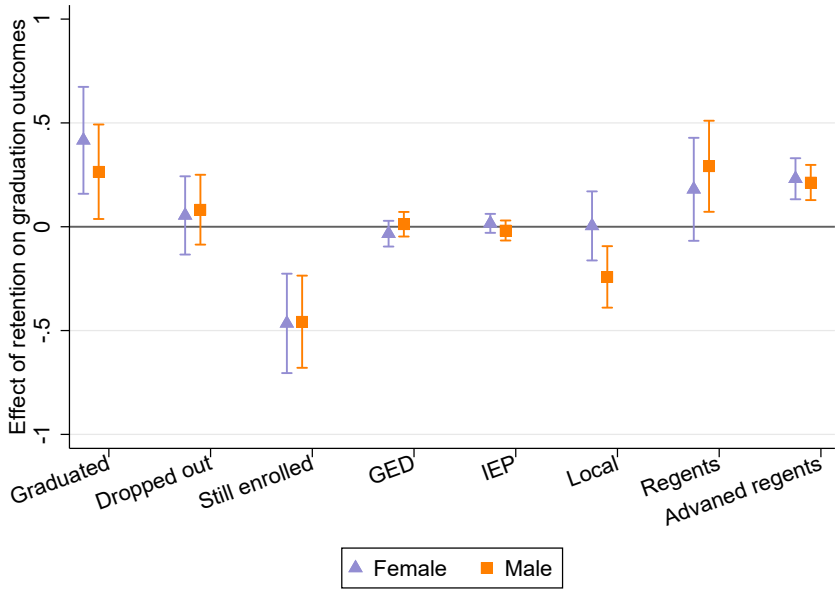



Table B.4: Effect on 4-year graduation outcomes (RK-IV estimates) by gender

| | | | | Type of diploma | | | | |
|---|---|---|---|---|---|---|---|---|
| | Graduated | Dropped out | Still enrolled | GED | IEP | Local | Regents | Advanced regents |
| Girls | | | | | | | | |
| Retention | 0.416*** | 0.054 | -0.465*** | -0.033 | 0.016 | 0.004 | 0.180 | 0.231*** |
| | (0.131) | (0.096) | (0.122) | (0.032) | (0.023) | (0.085) | (0.127) | (0.050) |
| Observations | 51371 | 51371 | 51371 | 51371 | 51371 | 51371 | 51371 | 51371 |
| Mean below cutoff | 0.549 | 0.173 | 0.298 | 0.014 | 0.008 | 0.131 | 0.388 | 0.030 |
| Mean above cutoff | 0.639 | 0.116 | 0.271 | 0.011 | 0.004 | 0.085 | 0.510 | 0.043 |
| F statistic | 179.763 | 179.763 | 179.763 | 179.763 | 179.763 | 179.763 | 179.763 | 179.763 |
| Boys | | | | | | | | |
| Retention | 0.265** | 0.082 | -0.457*** | 0.012 | -0.018 | -0.242*** | 0.291*** | 0.213*** |
| | (0.116) | (0.086) | (0.113) | (0.030) | (0.025) | (0.076) | (0.112) | (0.043) |
| Observations | 59059 | 59059 | 59059 | 59059 | 59059 | 59059 | 59059 | 59059 |
| Mean below cutoff | 0.489 | 0.175 | 0.344 | 0.017 | 0.012 | 0.121 | 0.341 | 0.027 |
| Mean above cutoff | 0.584 | 0.127 | 0.311 | 0.015 | 0.007 | 0.079 | 0.463 | 0.042 |
| F statistic | 233.900 | 233.900 | 233.900 | 233.900 | 233.900 | 233.900 | 233.900 | 233.900 |

*Notes:* Panel A reports the RK-IV estimates on graduation rates by race. Panel B reports the RK-IV estimates on dropout rates by race. Robust standard errors are in the parentheses. * p<0.10, ** p<0.05, *** p<0.01.

Figure B.5: Effects of retention on graduation outcomes (RK-IV estimates) by subsidized lunch eligibility

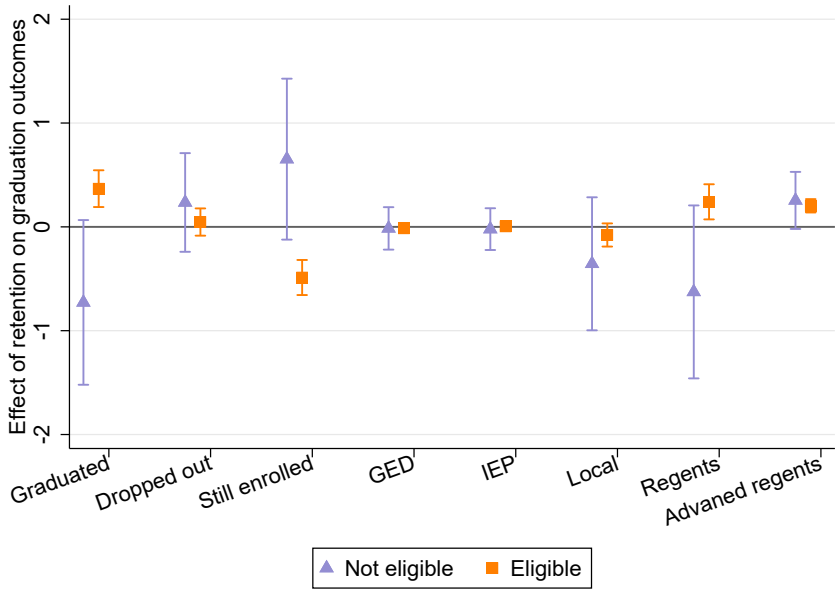



Table B.5: Effect on 4-year graduation outcomes (RK-IV estimates) by subsidized lunch eligibility

| | | | | Type of diploma | | | | |
|---|---|---|---|---|---|---|---|---|
| | Graduated | Dropped out | Still enrolled | GED | IEP | Local | Regents | Advanced regents |
| Not eligible | | | | | | | | |
| Retention | -0.728* | 0.235 | 0.652* | -0.015 | -0.022 | -0.356 | -0.626 | 0.255* |
| | (0.405) | (0.242) | (0.395) | (0.104) | (0.103) | (0.327) | (0.425) | (0.140) |
| Observations | 5835 | 5835 | 5835 | 5835 | 5835 | 5835 | 5835 | 5835 |
| Mean below cutoff | 0.648 | 0.093 | 0.274 | 0.014 | 0.013 | 0.207 | 0.421 | 0.020 |
| Mean above cutoff | 0.741 | 0.067 | 0.212 | 0.011 | 0.005 | 0.110 | 0.572 | 0.058 |
| F statistic | 18.892 | 18.892 | 18.892 | 18.892 | 18.892 | 18.892 | 18.892 | 18.892 |
| Eligible | | | | | | | | |
| Retention | 0.368*** | 0.046 | -0.488*** | -0.014 | 0.005 | -0.078 | 0.241*** | 0.203*** |
| | (0.090) | (0.067) | (0.086) | (0.023) | (0.017) | (0.057) | (0.086) | (0.033) |
| Observations | 91827 | 91827 | 91827 | 91827 | 91827 | 91827 | 91827 | 91827 |
| Mean below cutoff | 0.507 | 0.180 | 0.326 | 0.016 | 0.010 | 0.121 | 0.358 | 0.027 |
| Mean above cutoff | 0.599 | 0.127 | 0.298 | 0.013 | 0.005 | 0.080 | 0.480 | 0.040 |
| F statistic | 390.821 | 390.821 | 390.821 | 390.821 | 390.821 | 390.821 | 390.821 | 390.821 |

*Notes:* Panel A reports the RK-IV estimates on graduation rates by race. Panel B reports the RK-IV estimates on dropout rates by race. Robust standard errors are in the parentheses. * p<0.10, ** p<0.05, *** p<0.01.

Figure B.6: Effects of retention on graduation outcomes (RK-IV estimates) by age for grade

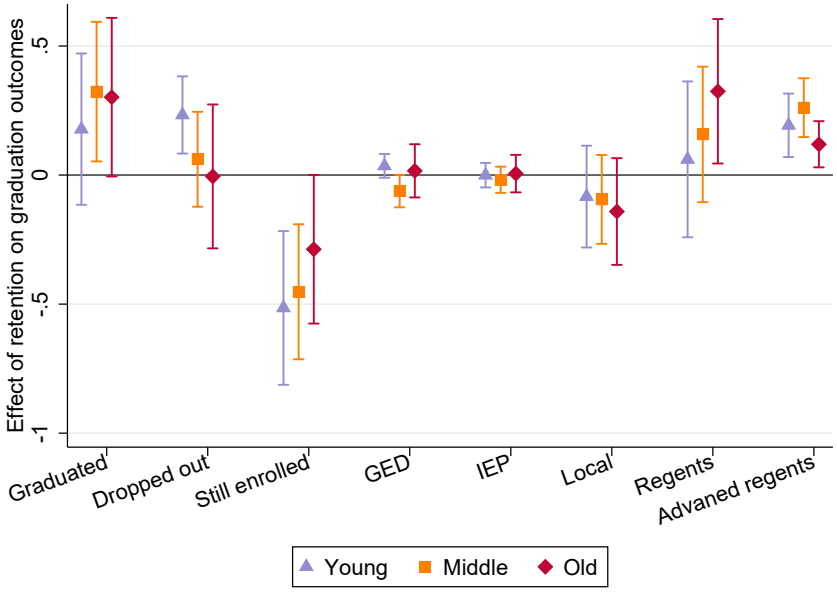


Table B.6: Effect on 4-year graduation outcomes (RK-IV estimates) by age for grade

| | | | | Type of diploma | | | | |
|---|---|---|---|---|---|---|---|---|
| | Graduated | Dropped out | Still enrolled | GED | IEP | Local | Regents | Advanced regents |
| Young | | | | | | | | |
| Retention | 0.178 | 0.233*** | -0.515*** | 0.036 | -0.000 | -0.083 | 0.061 | 0.193*** |
| | (0.150) | (0.076) | (0.152) | (0.023) | (0.024) | (0.101) | (0.154) | (0.063) |
| Observations | 36722 | 36722 | 36722 | 36722 | 36722 | 36722 | 36722 | 36722 |
| Mean below cutoff | 0.615 | 0.069 | 0.346 | 0.005 | 0.006 | 0.121 | 0.462 | 0.033 |
| Mean above cutoff | 0.687 | 0.047 | 0.303 | 0.004 | 0.003 | 0.071 | 0.562 | 0.054 |
| F statistic | 123.858 | 123.858 | 123.858 | 123.858 | 123.858 | 123.858 | 123.858 | 123.858 |
| Middle | | | | | | | | |
| Retention | 0.323** | 0.061 | -0.452*** | -0.062* | -0.018 | -0.094 | 0.158 | 0.261*** |
| | (0.138) | (0.094) | (0.133) | (0.032) | (0.026) | (0.088) | (0.134) | (0.058) |
| Observations | 38042 | 38042 | 38042 | 38042 | 38042 | 38042 | 38042 | 38042 |
| Mean below cutoff | 0.550 | 0.141 | 0.330 | 0.013 | 0.010 | 0.123 | 0.391 | 0.036 |
| Mean above cutoff | 0.636 | 0.100 | 0.290 | 0.012 | 0.005 | 0.072 | 0.517 | 0.047 |
| F statistic | 154.840 | 154.840 | 154.840 | 154.840 | 154.840 | 154.840 | 154.840 | 154.840 |
| Old | | | | | | | | |
| Retention | 0.302* | -0.005 | -0.287* | 0.016 | 0.006 | -0.141 | 0.325** | 0.119*** |
| | (0.157) | (0.142) | (0.147) | (0.053) | (0.037) | (0.105) | (0.143) | (0.046) |
| Observations | 35116 | 35116 | 35116 | 35116 | 35116 | 35116 | 35116 | 35116 |
| Mean below cutoff | 0.398 | 0.294 | 0.299 | 0.027 | 0.015 | 0.132 | 0.248 | 0.018 |
| Mean above cutoff | 0.490 | 0.233 | 0.284 | 0.023 | 0.009 | 0.108 | 0.360 | 0.022 |
| F statistic | 135.534 | 135.534 | 135.534 | 135.534 | 135.534 | 135.534 | 135.534 | 135.534 |

*Notes:* Panel A reports the RK-IV estimates on graduation rates by race. Panel B reports the RK-IV estimates on dropout rates by race. Robust standard errors are in the parentheses. * p<0.10, ** p<0.05, *** p<0.01.

Figure B.7: Effects of retention on graduation outcomes (RK-IV estimates) by height for grade

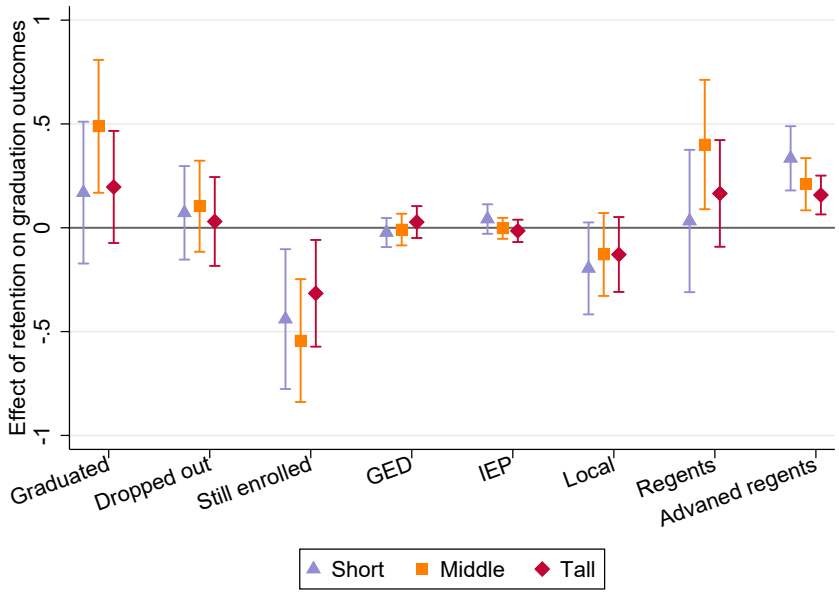


Table B.7: Effect on 4-year graduation outcomes (RK-IV estimates) by height for grade

| | | | | Type of diploma | | | | |
|---|---|---|---|---|---|---|---|---|
| | Graduated | Dropped out | Still enrolled | GED | IEP | Local | Regents | Advanced regents |
| Short | | | | | | | | |
| Retention | 0.169 | 0.072 | -0.440** | -0.023 | 0.042 | -0.196* | 0.033 | 0.335*** |
| | (0.174) | (0.115) | (0.172) | (0.036) | (0.036) | (0.113) | (0.175) | (0.079) |
| Observations | 28550 | 28550 | 28550 | 28550 | 28550 | 28550 | 28550 | 28550 |
| Mean below cutoff | 0.573 | 0.122 | 0.326 | 0.010 | 0.009 | 0.118 | 0.419 | 0.036 |
| Mean above cutoff | 0.659 | 0.083 | 0.289 | 0.009 | 0.004 | 0.074 | 0.531 | 0.055 |
| F statistic | 100.985 | 100.985 | 100.985 | 100.985 | 100.985 | 100.985 | 100.985 | 100.985 |
| Middle | | | | | | | | |
| Retention | 0.489*** | 0.104 | -0.543*** | -0.009 | -0.003 | -0.129 | 0.401** | 0.210*** |
| | (0.163) | (0.112) | (0.151) | (0.039) | (0.026) | (0.102) | (0.159) | (0.064) |
| Observations | 30276 | 30276 | 30276 | 30276 | 30276 | 30276 | 30276 | 30276 |
| Mean below cutoff | 0.548 | 0.161 | 0.309 | 0.015 | 0.008 | 0.127 | 0.389 | 0.033 |
| Mean above cutoff | 0.631 | 0.109 | 0.285 | 0.012 | 0.004 | 0.078 | 0.504 | 0.049 |
| F statistic | 121.641 | 121.641 | 121.641 | 121.641 | 121.641 | 121.641 | 121.641 | 121.641 |
| Tall | | | | | | | | |
| Retention | 0.196 | 0.030 | -0.316** | 0.028 | -0.015 | -0.129 | 0.165 | 0.158*** |
| | (0.138) | (0.109) | (0.131) | (0.039) | (0.027) | (0.092) | (0.131) | (0.048) |
| Observations | 36370 | 36370 | 36370 | 36370 | 36370 | 36370 | 36370 | 36370 |
| Mean below cutoff | 0.485 | 0.204 | 0.319 | 0.019 | 0.012 | 0.131 | 0.328 | 0.026 |
| Mean above cutoff | 0.587 | 0.147 | 0.287 | 0.016 | 0.006 | 0.090 | 0.463 | 0.034 |
| F statistic | 169.435 | 169.435 | 169.435 | 169.435 | 169.435 | 169.435 | 169.435 | 169.435 |

*Notes:* Panel A reports the RK-IV estimates on graduation rates by race. Panel B reports the RK-IV estimates on dropout rates by race. Robust standard errors are in the parentheses. * p<0.10, ** p<0.05, *** p<0.01.

Figure B.8: Effects of retention on graduation outcomes (RK-IV estimates) by weight status

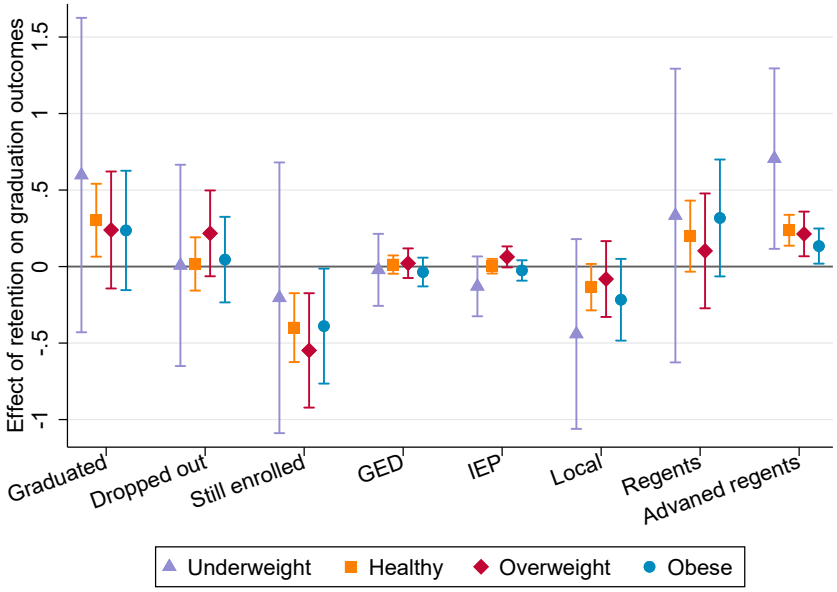


Table B.8: Effect on 4-year graduation outcomes (RK-IV estimates) by weight status

| | | | | Type of diploma | | | | |
|---|---|---|---|---|---|---|---|---|
| | Graduated | Dropped out | Still enrolled | GED | IEP | Local | Regents | Advanced regents |
| Underweight | | | | | | | | |
| Retention | 0.598 | 0.007 | -0.204 | -0.022 | -0.129 | -0.441 | 0.333 | 0.705** |
| | (0.524) | (0.336) | (0.451) | (0.120) | (0.100) | (0.316) | (0.490) | (0.301) |
| Observations | 3136 | 3136 | 3136 | 3136 | 3136 | 3136 | 3136 | 3136 |
| Mean below cutoff | 0.573 | 0.143 | 0.304 | 0.012 | 0.009 | 0.124 | 0.390 | 0.058 |
| Mean above cutoff | 0.672 | 0.095 | 0.255 | 0.009 | 0.004 | 0.075 | 0.532 | 0.065 |
| F statistic | 12.787 | 12.787 | 12.787 | 12.787 | 12.787 | 12.787 | 12.787 | 12.787 |
| Healthy | | | | | | | | |
| Retention | 0.303** | 0.017 | -0.399*** | 0.013 | 0.002 | -0.135* | 0.199* | 0.237*** |
| | (0.121) | (0.089) | (0.115) | (0.031) | (0.024) | (0.077) | (0.119) | (0.051) |
| Observations | 49651 | 49651 | 49651 | 49651 | 49651 | 49651 | 49651 | 49651 |
| Mean below cutoff | 0.531 | 0.170 | 0.310 | 0.016 | 0.011 | 0.121 | 0.373 | 0.037 |
| Mean above cutoff | 0.622 | 0.117 | 0.286 | 0.012 | 0.005 | 0.080 | 0.492 | 0.050 |
| F statistic | 214.764 | 214.764 | 214.764 | 214.764 | 214.764 | 214.764 | 214.764 | 214.764 |
| Overweight | | | | | | | | |
| Retention | 0.239 | 0.217 | -0.548*** | 0.022 | 0.063* | -0.082 | 0.102 | 0.213*** |
| | (0.195) | (0.143) | (0.191) | (0.049) | (0.035) | (0.126) | (0.191) | (0.074) |
| Observations | 18339 | 18339 | 18339 | 18339 | 18339 | 18339 | 18339 | 18339 |
| Mean below cutoff | 0.535 | 0.164 | 0.312 | 0.015 | 0.009 | 0.121 | 0.389 | 0.026 |
| Mean above cutoff | 0.631 | 0.113 | 0.281 | 0.012 | 0.005 | 0.078 | 0.506 | 0.047 |
| F statistic | 79.772 | 79.772 | 79.772 | 79.772 | 79.772 | 79.772 | 79.772 | 79.772 |
| Obese | | | | | | | | |
| Retention | 0.236 | 0.045 | -0.389** | -0.036 | -0.025 | -0.217 | 0.318 | 0.134** |
| | (0.199) | (0.143) | (0.192) | (0.048) | (0.034) | (0.136) | (0.195) | (0.059) |
| Observations | 23840 | 23840 | 23840 | 23840 | 23840 | 23840 | 23840 | 23840 |
| Mean below cutoff | 0.522 | 0.162 | 0.342 | 0.013 | 0.008 | 0.138 | 0.365 | 0.019 |
| Mean above cutoff | 0.613 | 0.116 | 0.298 | 0.013 | 0.005 | 0.087 | 0.494 | 0.032 |
| F statistic | 81.315 | 81.315 | 81.315 | 81.315 | 81.315 | 81.315 | 81.315 | 81.315 |

*Notes:* Panel A reports the RK-IV estimates on graduation rates by race. Panel B reports the RK-IV estimates on dropout rates by race. Robust standard errors are in the parentheses. * p<0.10, ** p<0.05, *** p<0.01.