NBER WORKING PAPER SERIES

THE EFFECTS OF TEST-BASED RETENTION ON STUDENT OUTCOMES OVER TIME: REGRESSION DISCONTINUITY EVIDENCE FROM FLORIDA

Guido Schwerdt
Martin R. West
Marcus A. Winters

Working Paper 21509
http://www.nber.org/papers/w21509

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
August 2015, Revised April 2017

We are grateful to the Florida Department of Education for providing the primary dataset for this study. We thank Stefan Bauernschuster, Matthew Chingos, Andrew Ho, Paul Peterson, Ludger Woessmann, and seminar participants at the National Bureau of Economic Research, Harvard University, the Ifo Institute, Mathematica Policy Research, Stanford University, the European Economic Association Meeting in Gothenburg, the European Association of Labour Economists Meeting in Turin and the Swedish Institute for Social Research for helpful comments. The Helios Education Foundation provided financial support for this research. The views contained herein are not necessarily those of the Helios Education Foundation. Any errors are our own. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

The Effects of Test-based Retention on Student Outcomes over Time: Regression Discontinuity Evidence from Florida

Guido Schwerdt, Martin R. West, and Marcus A. Winters

NBER Working Paper No. 21509
August 2015, Revised April 2017
JEL No. H52,I21,I28

ABSTRACT

Many American states require that students lacking basic reading proficiency after third grade be retained and remediated. We exploit a discontinuity in retention probabilities under Florida's test-based promotion policy to study its effects on student outcomes through high school. We find large positive effects on achievement that fade out entirely when retained students are compared to their same-age peers, but remain substantial through grade 10 when compared to students in the same grade. Being retained in third grade due to missing the promotion standard increases students' grade point averages and leads them to take fewer remedial courses in high school but has no effect on their probability of graduating.

Guido Schwerdt
Department of Economics
University of Konstanz
Box 133
78457 Konstanz
Germany
guido.schwerdt@uni-konstanz.de

Martin R. West
Harvard Graduate School of Education
Gutman Library 454
6 Appian Way
Cambridge, MA 02138
and NBER
martin_west@gse.harvard.edu

Marcus A. Winters
University of Colorado Colorado Springs
1420 Austin Bluffs Parkway
Colorado Springs, CO 80918
mwinters@uccs.edu

1 Introduction

Sixteen states and the District of Columbia have recently enacted policies requiring that students who do not demonstrate basic reading proficiency at the end of third grade be retained and provided with remedial services (Workman, 2014). Similar policies are under debate in states and school districts across the nation. Although these policies aim to provide incentives for educators and parents to ensure that students meet performance expectations, they can also be expected to increase the incidence of retention in the early grades. Their enactment has therefore renewed a longstanding debate about retention's consequences for low-achieving students.

Roughly 10 percent of American students are retained at least once between kindergarten and eighth grade, with the incidence of retention concentrated among low-income students and traditionally disadvantaged minorities (Planty et al., 2009). Retaining students in the same grade is costly in terms of additional per pupil spending and foregone earnings, if students (as intended) spend an additional year in full-time public education as a result of being held back. Yet consensus is lacking as to whether retention yields benefits for students that could offset these costs and, if so, under what conditions.

Proponents of policies encouraging the retention of low-performing students contend that these students stand to benefit from an improved match of their ability to that of their peers, from the opportunity for additional instruction before confronting more challenging material, and from any additional services provided to students during the retention year. Critics, meanwhile, warn that retained students may be harmed by stigmatization, reduced expectations for their academic performance on the part of teachers and parents, and the challenges of adjusting to a new peer group. In fact, a large literature in educational psychology confirms that retained students achieve at lower levels, complete fewer years of school, and have worse social-emotional outcomes than observably similar students who are promoted.[1] Because the decision to retain a student is typically made based on characteristics unobserved by the researcher, however, even studies that match retained and promoted students based on prior academic achievement are likely to suffer from selection bias. Consistent with this, more recent research in economics exploiting credibly

[1]Influential studies in this discipline include Jimerson (1999), Jimerson et al. (2002), and McCoy and Reynolds (1999). A survey of 47 empirical studies conducted by Holmes (1989) concluded that retained students perform 0.19 to 0.31 standard deviations worse on various measures of academic achievement than similar students who were not retained. In a meta-analysis of post-1990 research, however, Allen et al. (2009) report that a subset of studies that match retained and promoted students based on academic achievement or ability yields more positive estimates of retention effects than do studies that compare all retained and promoted students or match students based on non-academic variables.

exogenous variation in retention probabilities has found less negative and, in some cases, positive effects on student outcomes (Jacob and Lefgren, 2004, 2009; Greene and Winters, 2007).

In this paper, we use statewide administrative data covering all students in Florida public schools to study the causal effect of third grade retention and remediation on student outcomes through high school. The primary outcomes we examine include test scores for eight years following potential third grade retention in reading and six years in math, subsequent retention rates, and high school grade point average (GPA), coursetaking patterns, and graduation outcomes. The Florida database has four key advantages for studying the consequences of grade retention.

First, Florida since 2003 has required that schools retain third grade students who do not demonstrate basic proficiency on the state reading test unless the student is eligible for one of a specified set of exemptions. This test-based promotion policy generates a discontinuity in the probability of retention at the test score cutoff used to determine reading proficiency. We can therefore employ a standard regression discontinuity design to overcome the selection issues plaguing most existing research on this topic (Jacob and Lefgren, 2004, 2009; Greene and Winters, 2007; Winters and Greene, 2012).

Second, the Florida database contains vertically scaled test scores in reading and math that make it possible to compare the achievement of students tested in different grades during the same year. Making this comparison is essential because the counterfactual condition for students who are retained is to have been immediately promoted to the next grade. While often the sole focus of studies of retention, same-grade comparisons conflate any effect of retention with the effect of being a year older and having an additional year of schooling at the time the relevant test is administered.

Third, the availability of annual test scores for up to eight years after the retention decision makes it possible to determine the extent to which any changes over time in the magnitude of the estimated effect of retention are driven by grade-specific effects on achievement. The average amount students learn varies across grades for reasons including differences in teacher quality, the alignment of curricula with test content, and the share of students making school transitions. Because estimates of retention effects based on same-age comparisons capture these grade-specific effects along with the isolated effect of being retained, studies examining the outcomes of retained students after only two years (e.g., Jacob and Lefgren, 2004; Greene and Winters, 2007) are unable to determine whether any short-term effects of retention persist, fade out, or even grow larger over time.

Finally, the availability of high school transcript and graduation data through 2014 makes it possible to study the effects of test-based retention on students' course-taking patterns and performance in high school and on the probability that they graduate. For the first cohort of students affected by the policy, we are also able to provide a preliminary analysis of effects on enrollment in a Florida college.

It is important to note that the Florida policy requires that retained students be given the opportunity to attend a summer reading program prior to the next school year and that they be assigned to a "high-performing" teacher and receive intensive reading interventions during that year. Our estimates of the policy's impact will therefore capture the combined effect of retention and these additional measures and may not be directly comparable to those of some previous studies of retention. Requirements that retained students receive remedial interventions are typical of test-based promotion policies in use and under consideration in other settings, however, giving our results considerable policy relevance.

Due to the availability of exemptions for students scoring below the promotion cutoff, as well as to the voluntary retention of some higher-scoring students, our regression discontinuity design is fuzzy and yields estimates local to students who are retained as a result of the policy but would otherwise have been promoted (i.e., compliers). From a policy perspective, this local average treatment effect is arguably the most relevant parameter. Teachers granting a low-scoring student an exemption or recommending that a student with higher test scores be retained presumably do so because they have strong views as to whether retention would be beneficial for the student in question. In the case of compliers, in contrast, the fact that retention occurs only as a result of the test-based promotion policy implies that local educators are uncertain about whether retention is desirable. Moreover, because the retention policy is based on reading scores alone, we can exploit variation in compliers' math achievement to provide suggestive evidence that our estimates are generalizable to a broader population in terms of third grade achievement.

Our analysis confirms that students retained in third grade under Florida's test-based promotion policy experience substantial short-term gains in both math and reading achievement. On average over the first three years after being held back, retained students outperform their same-age peers who were promoted by 0.31 standard deviations in reading and by 0.23 standard deviations in math. These positive effects fade out over time, becoming statistically insignificant in both subjects within five years, but retained students continue to outperform their promoted peers when tested in the same grade through grade eight in math and grade ten in reading. Consistent with this evidence of improved performance

against grade-level expectations, we find that being retained in third grade as a result of missing the promotion standard improves students' grade point averages (GPAs) and leads them to take fewer remedial courses in high school. Test-based retention delays students' graduation from high school by 0.63 years but has no effect on their overall probability of graduating or their probability of receiving a regular diploma.

These findings contribute to an emerging literature using quasi-experimental research designs to study the effects of retention policies.[2] In prior studies of the Florida policy, Greene and Winters (2007) find that third grade retention improved student achievement after two years, and Winters and Greene (2012) present evidence based on same-grade comparisons that these gains persisted through eighth grade. Looking at behavioral outcomes, Ozek (2015) finds that students retained under the Florida policy were disciplined and suspended more frequently in the first two years after being retained, but that these effects dissipated entirely after two years. Jacob and Lefgren (2004, 2009) study the impact of retention in third, sixth, and eighth grade on achievement and high school completion in Chicago. They find that retention and mandatory summer school had a small positive short-term effect on achievement for third graders but not for sixth graders. They also find that retention reduced high school graduation rates for eighth graders but not for sixth graders. In a comparative setting, Manacorda (2012) finds that retention in junior high school increases dropout rates for Uruguayan students.

Taken as a whole, this evidence suggests that retention in higher grade levels may have detrimental effects on future student outcomes, but that early grade retention may be more beneficial. We confirm that test-based retention in third grade in Florida improves students' achievement in the short run but show that these initial academic benefits fade out over time. At the same time, test-based retention leads students to perform better academically and need less remediation while enrolled in high school and has no effect on their probability of graduating.

Our evidence that test-based retention in third grade reduces the probability of retention in subsequent grades highlights an additional consequence of policies that increase retention rates in early grades. Specifically, we show that many of the students retained as third graders as a result of Florida's test-based promotion policy would otherwise have

[2]In addition to the studies discussed in the text, Eide and Showalter (2001) use variation in kindergarten entry ages across states as an instrument for retention and conclude that retention increases high school completion and earnings for white students, although their results are not statistically significant. Using within-state variation in primary school retention rates from 1960 to 1980, Babcock and Bedard (2011) show that a one standard deviation increase in retention rates is associated with a 0.7 percent increase in mean earnings for adult males.

been retained in a subsequent grade. After five years, students retained in third grade are, on average, only 0.73 grade levels behind their promoted peers. To the extent that later grade retention is in fact less beneficial, students who are retained earlier rather than later may particularly benefit from the policy.

The paper proceeds as follows. In Section 2 we describe the Florida policy and our data and discuss measurement issues. Section 3 presents our identification strategy and provides graphical evidence supporting its validity, while Section 4 presents our findings concerning the effects of third grade retention on student outcomes over time, demonstrates their robustness, and examines potential mechanisms. Section 5 concludes.

2 Institutional Setting, Data, and Measurement

2.1 Test-based Retention Policy in Florida

In 2002, Florida's legislature mandated that third grade students scoring below level two (of five performance levels) on the Florida Comprehensive Assessment Test (FCAT) reading test be retained and provided with remedial services unless they qualify for one of six "good cause exemptions."[3] The Florida policy's exclusive focus on third grade reading distinguishes it from test-based promotion policies in Chicago and New York City, which include retention gates based on reading and math achievement at multiple grade levels. This focus reflects a common belief among educators that acquiring basic reading proficiency by third grade is essential for subsequent performance across disciplines, as well as the fact that third grade is the lowest included in the state testing program.

Students scoring below the level two cutoff may be granted an exemption from the policy if they fall into any of the following categories: students with disabilities whose Individualized Education Plan indicates that the state test is an inappropriate measure of their achievement; students with disabilities who were previously retained in third grade; Limited English proficiency (LEP) students with less than two years of instruction in English; students who were retained twice previously; students scoring above the 51st percentile nationally on another standardized reading test; and students demonstrating proficiency through a portfolio of work.[4] In light of these exemptions, the term "test-based promotion policy" may be a misnomer. It would be more precise to say that, for students

[3]The description of the Florida program in this section is based on Office of Program Policy Analysis & Government Accountability (2006).

[4]Since the 2004-05 school year, retained students have also been given the opportunity for a midyear promotion to fourth grade if they demonstrate mastery of necessary skills at that time.

not in special education, a low test-score shifts the burden of proof such that educators need to make an affirmative case that the student should be promoted.

Even so, the policy sharply increased the number of students held back in third grade. The number of Florida third graders retained jumped to 21,799 (13.5 percent) as the policy was implemented in 2003, up from 4,819 (2.8 percent) the previous year. The number of Florida students retained in third grade fell steadily over the next five years, reaching 9,562 (5.6 percent) in 2008, due primarily to a reduction in the number of students failing to meet the promotion standard.

As noted above, the policy includes several provisions intended to ensure that retained students acquire the reading skills needed to be promoted the following year. First, retained students must be given the opportunity to participate in their district's summer reading camp. Schools must also develop an academic improvement plan for each retained student and assign them to a "high-performing teacher," as determined by satisfactory performance appraisals. Finally, while repeating third grade, retained students must receive intensive reading interventions including ninety uninterrupted minutes daily of research-based reading instruction.[5]

2.2 Data and Estimation Sample

The data for our analysis are drawn from the Florida Department of Education's PK-20 Education Data Warehouse and contain information on all Florida students attending public schools in grades 3 to 12 from the 2000-01 through 2013-14 school years. We identify retained students based on the grade level of the state tests taken in adjacent years.[6]

The first cohort to be impacted by the test-based promotion policy (which we will refer to as the 2003 cohort) entered third grade in the 2002-03 school year and can be followed until 2013-14, one year after the students in this cohort who were retained under the policy should have graduated from high school. The five subsequent cohorts that we include in our analysis enter third grade in later years and can therefore be tracked for progressively shorter periods of time. Our primary analyses pool the data on all cohorts for which the relevant outcome is available.

Our basic data extract includes the school each student attends and its location; student characteristics such as ethnicity, gender, special education classification, English proficiency, and free lunch eligibility; annual measures of absences; annual FCAT math and

[5]In 2004-05 the uninterrupted ninety minute reading block became mandatory for all K-5 students.

[6]Students receiving mid-year promotions after 2004-05 will therefore be recorded as not being retained.

reading test scores in grades 3 to 10 from the 2000-01 through 2009-10 school years; and annual reading test scores in grades 9 and 10 from a revised state test from the 2010-11 through 2013-14 school years.

In addition to raw test scores, our data extract includes vertically equated Developmental Scale Scores (DSS) intended to support comparisons of student achievement across grade levels. During the 2000-01 school year, when the FCAT assessment system was expanded to include reading and math in all grades three through ten, a special data collection scheme incorporated the use of common items administered to students across multiple grades. Specifically, operational items from each grade's test were also included on the test administered to the higher and lower adjacent grade. These common items permitted the use of Item Response Theory (IRT) methods to place results from each grade's test on a common scale.[7]

As of the 2010-11 school year, Florida replaced the FCAT with the FCAT 2.0, a new assessment system aligned to revised academic content standards. This complicates our analysis in two ways: First, because of this change we do not have directly comparable test score information in grades 9 and 10 for all members of the two earliest cohorts (i.e., those who were retained and promoted). We therefore base our analysis of retention effects on ninth grade test scores on the 2005 to 2007 cohorts and our analysis of tenth grade test scores on the 2004 to 2006 cohorts. Second, FCAT 2.0 does not include a single math test for all students statewide in grades nine and ten, but rather has separate end-of-course tests for students taking different math courses. As a result, we can examine effects on test scores beyond eighth grade in reading only.

In addition, we obtained high school enrollment and transcript data through the 2013-14 school year. This allows us to construct complete enrollment histories for students in the first two cohorts affected by the test-based promotion policy, to develop measures of students' GPA and course-taking patterns, and to identify students who successfully graduated from a public high school in Florida. We also obtained information on enrollment in Florida colleges and universities through the 2013-14 school year, which we use to conduct a preliminary analysis of retention effects on enrollment in post-secondary education in the year following graduation for the 2003 cohort.

Table 1 provides summary statistics of student characteristics in third grade for the pooled sample covering the 2003-2008 cohorts used to study outcomes one year after potential retention. The first column reports mean characteristics (measured in third grade)

[7]See Hoffman et al. (2001) for technical details on the construction of the developmental scale scores.

for all students; columns (2) and (3) include all retained and all promoted students scoring below the cutoff; and columns (4) and (5) include all retained and all promoted students scoring above the cutoff. The table shows that 8.3 percent of all Florida students in these cohorts were retained in grade 3. This includes almost half (47.8 percent) of students scoring below the promotion cutoff, as well as an additional 0.6 percent of students scoring above the cutoff.

Naturally, students scoring below the cutoff and retained students perform at low levels. For example, retained students below the cutoff score 603 points (1.63 standard deviations) below the average student in reading and 415 points (1.36 standard deviations) below the average student in math.[8] Compared to the retained students below the cutoff, the relatively few voluntarily retained students are higher performing on average, more likely to be white, and substantially younger than the average retained student. They are also absent more frequently as third graders, perhaps suggesting the importance of behavioral indicators to voluntary retention decisions.

Table 2 provides summary statistics of student outcomes in high school and beyond. As discussed above, not all outcomes are observed for all cohorts. Thus, Table 2 reports data on each outcome for the respective cohorts for whom it is observed. These data reveal that, descriptively, promoted students clearly outperform their retained peers in the long run. They score higher on the FCAT 2.0 reading test in grades 9 and 10, are more likely to enter and graduate from high school, and take fewer remedial courses while enrolled. They are also more likely to take college preparatory courses in high school and to enroll in college immediately upon graduation. Of course, these differences in outcomes may reflect unobserved differences betweeen students who were retained and promoted in third grade and cannot be interpreted as causal effects of test-based retention.

2.3 Measurement Issues

Analyzing the effects of grade retention necessitates a choice about when to measure and compare students' future outcomes (Allen et al., 2009). The standard approach for any (quasi-)experimental analysis of the effect of a given treatment is to compare outcomes measured at the same point in time (e.g., 2 years after treatment) for treated observations and those standing in for their counter-factual outcomes. In the case of grade retention, this corresponds to a comparison of outcomes when treated and non-treated students are

[8]Students' raw third grade test scores are expressed in the metric of the vertically equated Developmental Scale Scores (DSS) to have a consistent way of reporting student achievement based on FCAT scores. The standard deviations of DSS scores in grade 3 are 306 points in math and 370 points in reading.

of the same age. However, the very nature of the retention treatment implies that the future grade levels of treated and non-treated students will differ, as will their expected graduation date. As achievement is typically measured by grade-specific tests, it is common in the retention literature to deviate from the standard approach and compare outcomes when students have reached the same grade. This enables researchers to address questions such as "What is the effect of third grade retention on student achievement in grade g?".

It can be shown that neither same-age nor same-grade comparisons identify the isolated effect of retention absent further assumptions.[9] For example, in a same-grade comparison treated students will be older and will have been exposed to more schooling than non-treated students when they take any grade-specific test after being retained. Moreover, even if the outcomes of treated and non-treated students were identical in expectation at the time of treatment, effects of any prior interventions that fade out over time will confound a same-grade comparison. A same-grade comparison therefore identifies the isolated effect of retention only in the absence of age effects and time-in-school effects and if any effects of prior interventions (potentially including prior retentions) do not fade out. A same-age comparison will not be affected by these issues.

However, a same-age comparison may nonetheless be confounded by true differences in the average rate of learning across grades. For example, Figure 1, which plots average DSS scores in reading and math by grade for all students in the pooled dataset, shows that Florida students experience very small gains in math achievement in sixth grade relative to the gains made by students in other grades. This pattern likely reflects the fact that most Florida students transition into a middle school in grade six, which Schwerdt and West (2013b) show has a negative impact on their achievement growth. To the extent that retention simply delays students from experiencing a grade in which their own achievement growth is likely to be smaller, policymakers may want to incorporate this information into the metric used to compare their achievement to that of promoted students.

Overall, the jagged trajectory evident in both subjects in Figure 1 indicates that average achievement gains as measured by developmental scale scores vary considerably by grade. This variation provides a first indication of how the point in time at which outcomes are compared can influence estimates of the causal effect of retention.

In practice, both same-age and same-grade comparisons can offer useful evidence on retention's consequences depending on how the desired treatment effect is defined. For example, because attending third grade a second time rather than fourth grade the follow-

[9]See (Schwerdt and West, 2013a) for a more formal discussion of the identification assumptions of same-age and same-grade comparisons.

ing year is a direct consequence of being retained, differences in instructional quality or content across grades may reasonably be considered part of the desired treatment effect. Estimates based on a same-age comparison would therefore represent a meaningful causal effect of retention despite the fact that the total effect is partly driven by instructional differences between grades. Conversely, policymakers or parents may be most interested in what students know when they reach a specific grade or upon graduation from high school, which are same-grade comparisons. Same-grade comparisons may be of particular interest in anticipating retention's potential effects on outcomes that turn on students' performance relative to their same-grade peers, such as admission to selective colleges.

From an economic perspective, the choice between same-age and same-grade comparisons may hinge on assumptions about the functioning of the labor market. If productivity were perfectly observable, then conditional on skills educational credentials should be unrelated to labor market outcomes. A same-age comparison of student skills would therefore provide a direct estimate of retention's effects on those outcomes. However, with imperfect information and sheepskin effects, a same-grade comparison may be preferable. In the extreme case in which the high school diploma provides the only signal allowing employers to distinguish high and low productivity workers, all that matters is whether a student graduates from high school. If retention increases the probability of graduation, retention has a benefit. Whether it is also cost-effective then depends on the size of credential's effect on life-time earnings relative to its opportunity costs.

Because each approach identifies a potentially interesting (combined) treatment effect, in our empirical analysis we report estimates based on both. We focus our interpretation on same-age comparisons, however, due to their advantages in terms of identifying the isolated effect of grade retention – the parameter that conceptually links estimates based on the two approaches. The distinction between same-grade and same-age comparisons is less fundamental for measures of educational attainment, provided enough time has passed for all students who will eventually receive a given credential to have obtained it. Our analysis of the effects of retention on high school graduation focuses on whether students in the first two cohorts affected by the policy had received a high school degree by the end of the 2013-14 year. At this point, students of the 2003 cohort were on average 20 years old, while students of the 2004 cohort were about 19.

As discussed above, Figure 1 indicates that the achievement gains made by typical students on this scale are not uniform across grades. Thus, estimates based on a same-age approach may vary with the number of years since treatment for at least two reasons: true

fade out of retention effects and grade-specific effects on achievement conditional on the number of prior years of schooling.[10] To back out an approximate estimate of the extent of true fade out of retention effects over time, we construct an alternative vertical scaling of reading and math achievement, which is also plotted in Figure 1. Specifically, we subtract from each student's DSS score the grade-specific mean score and then add the predicted value for each grade based on a linear regression of mean scores on grade level. These rescaled scores increase linearly from grades three to ten by construction. The estimated slope coefficients, which indicate the average annual rate of achievement growth between third and tenth grade, are 80 DSS points in reading and 83 DSS points in math. The assumption of linear achievement growth underlying the rescaling is admittedly arbitrary, and point estimates based on rescaled scores do not necessarily represent an unbiased estimate of the isolated retention effects. Comparing estimates based on rescaled scores across years should nonetheless be informative about the rate at which retention effects fade out over time.

3 Empirical Strategy

Empirical strategies that rely on a selection-on-observables assumption will fail to provide unbiased estimates of the effect of early grade retention on future student outcomes if students are selected for retention based on factors unobserved by the researcher that influence educational outcomes. We address this concern by taking advantage of Florida's test-based promotion policy, which leads to a discontinuous relationship between third grade reading test scores and the probability of grade retention. This discontinuity generates plausibly exogenous variation in retention which we exploit to identify the causal effect of test-based retention on future outcomes.

3.1 Graphical Evidence

Our identification strategy hinges on the assumption that Florida's test-based promotion policy generates exogenous variation in third grade retention that we can exploit for identification using a regression discontinuity. We first present graphical evidence of the existence

[10]Fade out may also be a mechanical artifact of the practice of rescaling grade-specific test scores if a standard deviation in test scores in later grades translates into a larger difference in knowledge (Lang, 2010). This is less of a concern in our case as we report results based on non-standardized vertically scaled scores across all grades. Moreover, (Cascio and Staiger, 2012) demonstrate that this mechanism is unlikely to fully explain fade-out of the effects of educational interventions

of a discontinuity in the relationship between a student's third grade reading test scores and the probability of being retained. We then discuss potential threats to the validity of regression discontinuity studies and provide additional graphical evidence demonstrating that these threats are not applicable in this setting (c.f., Lee and Lemieux, 2010). Unless otherwise noted, all figures are based on the pooled data set of students in the 2003-2008 cohorts.[11]

Panel B of Figure 2, which plots the share of students retained as a function of third grade reading scores (measured relative to the test score cutoff), provides visual evidence of the discontinuity in retention probabilities. The data points represent the share of students retained for each possible score on the third grade reading test, with each marker's size proportional to the number of students receiving that score. The solid line represents predicted values from separate local linear regressions on either side of the cutoff. For students 30 or more points ($>$ .5 standard deviations) below the cutoff, retention probabilities are relatively stable at just under 0.6. The probability of retention then declines as test scores increase, with retention probabilities immediately to the left of the cutoff approaching 0.3. Retention probabilities drop sharply to less than 0.05 at the cutoff, however, and approach zero 50 points above it.

Panel A of Figure 2 displays the same relationship for the two cohorts of students in our data extract entering third grade immediately prior to the introduction of the test-based promotion policy. Note that the probability of retention for students in these cohorts rarely exceeds 20 percent, even for very low-scoring students. More importantly, the probability of retention is essentially continuous around the cutoff, indicating that the discontinuity evident in panel A of Figure 2 was in fact generated by the policy change.

While Figure 2 is based on the full distribution of third grade reading test scores, we limit our regression discontinuity analysis of the causal effects of retention to a narrower sample of students within a 10 test-score-point bandwidth on either side of the cutoff. Figure A-1 in the Supplementary Appendix illustrates the discontinuity within this more restricted sample, again plotting the fraction of students retained by third grade reading test scores measured relative to the cutoff. Local regressions on either side of the cutoff suggest an approximately linear relationship between test scores and retention probabilities in the cutoff region. However, the slope of this relationship clearly differs for students below and above the cutoff. We make use of this observation below when specifying the functional relationship between the forcing variable (reading test scores) and the retention indicator

[11] Cohort-specific graphs are available from the authors upon request.

in our empirical model.

A common concern with regression discontinuity analyses is the possibility of precise manipulation of the forcing variable around the cutoff (c.f., Urquiola and Verhoogen, 2009). In this setting, for example, one might worry that teachers were able to manipulate students' reading scores to push them just above the promotion cutoff. The fact that the FCAT reading test is scored objectively without teacher input makes this possibility unlikely, however, and Figure A-2 in the Supplementary Appendix confirms that the overall distribution of reading test scores shows no evidence of a heaping of observations around the cutoff.

The regression discontinuity identification strategy also assumes that there are no discontinuities in other characteristics associated with student outcomes at the cutoff. Figure A-3 in the Supplementary Appendix addresses this issue by plotting the mean value of the observable student characteristics available in our data against third grade reading test scores. In addition to examining each characteristic individually, we also use a probit model to generate a predicted retention probability for each student based on all available background characteristics (except reading scores). The figure confirms the absence of discontinuities in observed student characteristics at the test-score cutoff used to inform retention decisions.

Finally, we confirm that attrition from the Florida database in subsequent years also does not vary discontinuously at the promotion cutoff. Even in the absence of sorting around the cutoff based on prior characteristics, differential attrition could occur if, for example, being retained in third grade made students more likely to leave the Florida public schools. Figure A-4 in the online appendix therefore plots attrition rates against third grade reading scores around the cutoff.[12] Attrition rates increase as expected with the number of years since potential third grade retention, but they appear to be unrelated to third grade reading scores and there is no evidence of a discontinuity at the promotion cutoff.[13]

[12]To enhance legibility, the figure plots attrition rates after two, four, and six years only; the patterns after three and five years are similar. Because we identify students as having been promoted or retained in third grade based on the grade in which they are observed the following year, attrition rates one year after potential retention are zero by construction. We can, however, examine the rate of attrition among all students tested in third grade regardless of whether we observe them in Florida public schools the following year. Table A-1 in the Supplementary Appendix confirms that attrition rates after one year and subsequently do not vary discontinuously around the promotion cutoff.

[13]In addition to the graphical analyses in figures A-3 and A-4, we used each student characteristic and attrition in each year after potential third grade retention as the outcome variable in regressions with the same specification and bandwidth as our preferred regression discontinuity model. The results (available upon request) confirm the absence of any statistically significant breaks in the relationship between reading

3.2 Estimation

Because only a subset of students scoring below the cutoff in reading test scores were actually retained, our empirical analysis takes the form of a fuzzy regression discontinuity design that can be implemented via instrumental variables (IV) estimation. In our preferred specification we estimate the causal effect of test-based retention on future student outcomes in a two-stage least squares model. The first stage is given by the following equation:

$$retained = \gamma_1 below + \gamma_2 forcevar + \gamma_3 below \times forcevar + \Gamma X + \epsilon, \quad (1)$$

where *retained* indicates retention in grade 3, *below* indicates that the student scored below the promotion cutoff on the grade 3 reading test, $forcevar$ measures student achievement on the grade 3 reading test (centered around the cutoff score), X is a vector of student demographic characteristics including the student's math achievement in grade 3, and ϵ is a standard zero-mean error term. Note that, based on the graphical evidence in Figure 2, we model the relationship between reading scores and the retention indicator as linear with a break in the trend at the cutoff.

The corresponding second stage of our 2SLS model is given by:

$$y = \delta_1 retained + \delta_2 forcevar + \delta_3 below \times forcevar + \Lambda X + \eta, \quad (2)$$

where y denotes the student outcome of interest.[14] We achieve identification of δ_1 by instrumenting for grade retention in grade 3 (*retained*) with the indicator for being below the cutoff for promotion to grade 4 (*below*). As noted above, we estimate the 2SLS model for the sample of students within ten test score points on either side of this cutoff. We select this bandwidth based on the optimal bandwidth algorithm developed by Imbens and Kalyanaraman (2012) and demonstrate the robustness of our results to alternative bandwidths in Section 5.

Throughout the empirical analysis, we estimate and report two-way clustered standard errors clustered at the level of the grade 3 school and the level of the forcing variable for all regressions discontinuity designs. Using robust standard errors or clustering standard

scores and these outcomes at the promotion cutoff.

[14]Equations 1 and 2 represent our preferred specification, but some other choices would be equally justifiable. Fortunately, our results are extremely robust to minor specification changes. In particular, allowing the first stage effect to be different for students with special education or LEP status in grade 3 produces very similar results. Results available upon request.

errors at the level of each unique value of the forcing variable as suggested by Lee and Card (2008) produces quite similar standard errors and does not affect the interpretation of our results.

4 Results

Table 3 reports results from estimating the first-stage model in Equation (1) for each cohort of students separately and for the pooled sample. For purposes of comparison, we also present results for the two cohorts of students in our data that were not impacted by the policy. Note that all estimations are based on our preferred discontinuity sample within a 10 test-score-point bandwidth around the cutoff. Despite this narrow bandwidth, we still have between 9,981 and 15,687 students in each post-2002 cohort and a total of nearly 75,000 students in the pooled sample.

The first row of Table 3 presents estimates of the jump in the probability of retention at the promotion cutoff. Consistent with panel B of Figure 2, the first two columns confirm that there was essentially no such jump in the two years immediately preceding the policy's introduction.[15] In contrast, each of the cohort-specific estimates for students impacted by the policy is positive and highly statistically significant, with F-statistics on the excluded instruments exceeding 100. Point estimates of the jump in retention probabilities at the cutoff range from 0.20 to 0.37, with the largest estimate observed for the initial 2003 cohort and the two smallest estimates observed for the 2007 and 2008 cohorts. This pattern suggests that educators over this period made increasing use of the good cause exemptions within the policy allowing students performing below the promotion cutoff to avoid retention. The overall first stage effect for the pooled sample nonetheless indicates an increase of 0.28 in the probability of retention for typical students scoring immediately below the cutoff, relative to students scoring one point higher.

4.1 The Effect of Test-Based Retention on Student Achievement

We begin our discussion of the effects of grade retention on student outcomes with graphical evidence on the reduced form relationship between students' third grade reading test scores and their future achievement. Figure 3 is based on a same-age comparison and uses local linear regressions estimated separately on each side of the promotion cutoff to depict the

[15] Although the results for the 2002 cohort show a statistically significant increase in the probability of retention for students scoring below the cutoff, the cohort-specific estimates while the policy was in place are all more than ten times as large.

relationship between students' third grade reading test scores and their reading and math achievement up to six years after potential third grade retention.[16] In both subjects, we observe students scoring below the promotion cutoff performing at higher levels in the first three years after potential third grade retention. However, these differences dissipate in later years and, in some cases, turn slightly negative.

Table 4 presents estimates of the effects of test-based retention in third grade on reading and math achievement over time. All estimates are based on our preferred IV model with covariates and are local to students retained as a result of failing to meet the promotion standard.[17] Column (1) reports results from same-grade comparisons, while columns (2) and (3) report the effects of third grade retention on achievement when retained and promoted students are tested at the same age.

Estimates based on same-grade comparisons indicate large positive effects of test-based retention on reading and math achievement that diminimish over time but remain substantial for as long as we are able to observe (i.e., through grade eight in math and grade ten in reading). Specifically, retained students scored 73 (61) percent of a standard deviation higher than their promoted peers in reading (math) when both groups of students were first tested in grade four.[18] By the time students first reached grade eight, retained students scored 19 (13) percent of a standard deviation higher in reading (math). In grade ten, retained students continued to outperform their promoted peers by 22 percent of a standard deviation in reading when both were given the new FCAT 2.0 assessment.[19] As discussed in Section 2.3, these estimates capture the effects of being a year older and having

[16]In addition to previewing our findings, Figure 3, confirms that the reduced form relationship between third grade test scores and future student achievement is approximately linear around the promotion cutoff. Figure A-5 in the Supplementary Appendix similarly suggests a linear relationship between third grade test scores and both high school graduation and college enrollment. Along with the graphical evidence presented on the first stage relationship in Figure 2, these figures support the choice of equations 1 and 2 as our preferred specification for modeling retention effects.

[17]Tables A-2 to A-4 in the Supplementary Appendix additionally report OLS estimates from Equation (2) with and without covariates, as well as IV estimates without covariates. As expected, the inclusion of covariates does not notably influence the IV point estimates (although it modestly improves their precision) but substantially alters the OLS results. Relative to our preferred IV estimates, OLS estimates of the effects of third grade retention are always lower. In reading after one year, for example, the difference between the OLS and IV point estimates is more than one third of a standard deviation. This confirms the extent to which OLS estimates fail to control adequately for unobserved confounding factors and, thus, understate any benefits (and exaggerate any harms) of grade retention.

[18]We express the size of effects on achievement through grade 8 relative to the statewide standard deviation in third grade DSS scores, which are 370 in reading and 306 in math.

[19]We obtained FCAT 2.0 results only for students in our sample, not for all students taking the FCAT 2.0 in a given year. We therefore express the size of effects on reading achievement as measured in the FCAT 2.0 based on the statewide standard deviation of 21 for grade 10 students reported in Foorman et al. (2013) based on data for 2011 and 2012.

received an additional year of schooling along with the isolated effect of retention; they also incorporate any differential fade out of interventions students experienced prior to grade 3. Even so, they may be of interest to policymakers seeking evidence on how test-based promotion policies affect the performance of retained students measured relative to other students in the same grade.

Consistent with Figure 3, the same-age IV estimates in column 2 of Table 4 indicate that test-based retention improves students' reading and math achievement dramatically in the short run. Reading achievement improves by 23 percent of a standard deviation after one year and by as much as 49 percent of a standard deviation after two years. The estimated impact of retention on math achievement is 30 percent of a standard deviation after one year and grows to 36 percent of a standard deviation after three years. On average over the first three years after being held back, retained students outperform their promoted peers by 31 percent of a standard deviation in reading and by 23 percent of a standard deviation in math.

As with the same-grade comparisons, however, these initial benefits fade out in subsequent years. The effect of test-based retention on reading achievement remains statistically significant after six years, but is reduced to 11 percent of a standard deviation, and dissipates entirely after seven years. In the case of math achievement, the estimated effects become slightly negative in years four and five but are statistically insignificant after six years.[20]

One unusual aspect of the results in column (2) of Table 4 is the non-monotonic relationship between the size of the estimated impacts of retention and the time elapsed since the student was retained. The estimated impact is largest after two years in the case of reading achievement and after three years in math. Given the overall pattern of fade out and the fact that remedial services were required only in the year the student was retained, one would expect the impact of retention to be largest at the end of that year. This pattern likely stems in part from the grade-to-grade variation in the average achievement gains of Florida public school students as measured by DSS scores. For example, Figure 1 shows that Florida students experience particularly large gains in DSS reading achievement in fourth grade, which promoted students enter immediately and (most) retained students enter one year later. This difference in timing could explain the unexpected growth from

[20]Tables A-5 and A-6 in the Supplementary Appendix present the same year-by-year results separately for each cohort and confirm that this apparent fade out in the effects of third grade retention over time does not simply reflect smaller impacts of retention on the earliest cohorts, whose outcomes we are able to observe for more years.

year one to year two in the estimated impact of retention on DSS reading achievement. The alternative scaling of the DSS scores discussed in section 2.3 eliminates variation in average achievement gains across grades and thereby allows us to approximate the true rate of fade out over time.

Column (3) of Table 4 presents IV estimates of Equation (2) based on these rescaled DSS scores. In both reading and math, the magnitude of the estimated impacts now decreases monotonically with distance from treatment. In reading, the impacts based on the rescaled DSS scores are as large as 58 percent of a standard deviation after one year but fade to 11 percent of a standard deviation by year four and are statistically insignificant thereafter. In math, the impacts start at 42 percent of a standard deviation but are statistically insignificant by year four and become modestly negative after five years. Qualitatively, however, the results concerning achievement impacts of third grade retention do not depend on the test scaling. Both sets of same-age comparisons show large positive initial impacts of retention that fade out gradually over time.

Overall, our analysis of student achievement suggests that test-based retention in third grade has substantial positive effects on achievement in the short-run but that these effects fade out completely over time. Retained students continue to perform better than their promoted peers in reading when they are tested in the same grade through at least grade 10, but this is likely due to the effects of age and schooling and cannot necessarily be interpreted as a long-run effect of grade retention.

4.2 The Effect of Test-Based Retention on Grade Progression

We next present estimates of the effect of test-based retention in third grade on students' subsequent grade progression through grade 8.[21] Grade progression is an important outcome to consider for at least two reasons. First, the effects of retention on outcomes such as student achievement and attainment could vary according to the grade level at which the student is retained. If retention in early grades is more beneficial to students than later retention, test-based promotion policies targeting early grades could benefit students who would eventually be retained by ensuring that they are retained at a younger age. Second, if low-achieving students who narrowly avoid retention in third grade are more likely to be retained in subsequent grades, this could explain some or all of the fade out of the test score effects we have documented for students retained in third grade.

[21] We also examined the impact of test-based retention on student absences and special education placement and confirmed that it had no impact on these outcomes (results available upon request.)

Table 5 reports estimates of the effect of test-based retention on future retention probabilities and subsequent grade progression based on the regression discontinuity sample within 10 points of the promotion cutoff.[22] We limit this analysis to upper elementary and middle school grades because the nature of grade retention changes in high school, when students are typically asked to repeat specific courses they have failed rather than an entire grade. The estimates in column (1) show that third grade retention reduces the probability that the student will be in the process of repeating a grade two years later by 11 percentage points. The effect is smaller in subsequent years, but remains statistically significant and ranges from 2 to 4 percentage points in magnitude in years three to five. The estimates in column (2) of Table 5 use grade level as the outcome variable in Equation (2), thereby providing direct evidence on the differences in the grade progression of retained and promoted students. These estimates show that five years after being retained in third grade, students retained under Florida's test-based promotion policy are only 0.73 grade levels behind comparable peers who were promoted.

Table 5 confirms that test-based retention substantially reduced the probability that Florida students at the promotion cutoff would be retained in future grades. Could these differences in subsequent grade progression explain the fade out of test score impacts for students retained in third grade? To evaluate this possibility, we assume that (1) the effects of retention on student achievement after one year are in fact fully persistent and (2) that students retained in subsequent grades experience the same short-term benefits, regardless of the grade in which they were retained. We then ask how much of the observed fade out in test score impacts from year one to year two would be explained by the additional gains made by students retained in year two. The results suggest that differences in subsequent retention could account for no more than 38 percent of the observed fade out in reading effects after two years and 25 percent of the fade out in math effects.[23] Additional analyses also confirm that the test score impacts in both subjects fade out even when students who were subsequently retained are excluded from the sample.

[22]Table A-7 in the Supplementary Appendix provides estimates of the impact of third grade retention on subsequent grade progression by cohort.

[23]For example, the simple calculation in terms of reading is as follows: Observed fade out in reading effects between year one and two is given by 214.9 - 152 = 62.9 DSS points (see column 3 of Table 4). Fade out resulting from a 11 percentage point reduction in the probability of being retained after two years (see column 1 of Table 5) is given by 0.11* 214.9 = 23.64 DSS points. Thus, roughly 38 percent of the fade out in reading effects after two years could be explained by effects on future grade retention.

4.3 The Effect of Test-Based Retention on High School Graduation

In addition to studying students' subsequent grade progression, we also estimate the effect of test-based retention on the probability of graduating from a Florida public high school for students in the 2003 and 2004 cohorts, the first two cohorts subjected to the state's test-based promotion policy.[24] Assuming typical grade progression, students in the 2003 (2004) cohort who were retained in third grade would be expected to graduate from high school at the end of the 2012-13 (2013-14) school year, one year after their promoted peers. As in the case of other outcomes, we focus our analysis on the regression discontinuity sample within 10 points of the promotion cutoff.

Figure 4 tracks school enrollment, average grade levels, and graduation outcomes for students just above and below the promotion cutoff from the 2002-03 school year through 2013-14 separately for the 2003 and 2004 cohorts. Students above and below the cutoff from the 2003 cohort remained enrolled in Florida public schools at very similar rates through the 2011-12 school year, when roughly 52 percent of students above the cutoff graduated. Among students below the cutoff, roughly half of whom were retained in third grade, only 32 percent graduated in 2011-12. In 2012-13, however, 25 percent of students below the cutoff graduated, as compared with just 10 percent of students above the cutoff. In 2013-14, an additional 3 percent of students below the cutoff and 2 percent of students above the cutoff graduated. The total share of the 2003 cohort graduating by 2013-14 was 65 percent and 61 percent, respectively, for students above and below the cutoff. The patterns of enrollment, grade progression, and graduation among students above and below the promotion cutoff are similar for the 2004 cohort, who we can follow for one less year.

Table 6 presents estimates of the effect of test-based retention in third grade on the probability that students entered and graduated from a public high school in Florida by the 2013-14 school year, as well as on the school year in which graduating students received their diploma. Column (1) indicates that the marginal students retained as a result of missing the promotion standard in third grade were no less likely to enter a public high school in Florida, while column (2) shows that third grade retention had no causal effect on high school graduation for these students as of the end of the 2013-14 school year.[25]

[24]Tables A-5, A-6, and A-7 in the Supplementary Appendix indicate that the effects of retention on student achievement and future grade retention for the 2003 and 2004 cohorts were broadly similar to those for the pooled sample.

[25]Note that 5 percent of the students in the 2003 cohort and 24 percent of the students in the 2004 cohort remained enrolled in 2013-14 but did not successfully graduate; some of these students can be expected

The specification reported in column (2) classifies students as high school graduates if they received any type of diploma. Columns (3)-(5) show that, conditional on graduating, third grade retention also did not affect the probability that students received a regular diploma, a certificate of completion, or a GED. Finally, column (6) shows that, conditional on graduating from high school, being retained in third grade delayed the timing of high school graduation by only 0.63 school years. The fact that this number is less than one is consistent with our previous findings that third grade retention reduced the probability of retention in future grades. Overall, we interpret the results in Table 7 as evidence that third grade retention did not significantly affect students' high school graduation rates or the type of credential they received and delayed the progress of graduating students by substantially less than a full year.

Even without influencing high school graduation outcomes, test-based retention could affect the number of grades (and therefore core academic courses) students complete prior to leaving school. This could occur in part due to them having completed fewer grades at the time they first exceed Florida's compulsory schooling age of 16. We therefore also estimated the effect of test-based retention on the highest grade students completed, their age when they left Florida public schools, and the probability that they completed the highest grade at age 16, 17, and 18 (or older). We conducted these analyses separately for the 2003 cohort and for the 2003 and 2004 cohorts combined. The results, which we present in Table 7, show no statistically significant effect of retention on the highest grade students in these cohorts completed. While the point estimate is negative, the standard error is small enough for us to rule out negative effects as small as one third of a grade. Consistent with this, we find that retention increased the age at which students completed their highest grade by 0.54 years. Retention led students to be 1.5 percentage points less likely to leave at age 16, 2.7 percentage points less likely to leave at age 17, and 3.6 percentage points more likely to leave at age 18 or older. In sum, we find no statistically significant evidence that test-based retention reduced the number of grades students successfully completed prior to leaving school. Rather, the marginal students retained appear to have responded to being older when they reached a given grade level by staying enrolled longer.

to graduate in subsequent years. However, estimates based on the 2003 cohort alone, 95 percent of whom had left school before the 2013-14 school year, also do not suggest a significant effect of retention on high school graduation (results available upon request).

4.4 The Effect of Test-Based Retention on High School GPA and Course-taking

While test-based retention in third grade did not impact high school graduation rates among the first two cohorts of students affected by the policy, the results in section 4.1 demonstrate that retention caused students to enter high school performing at higher levels in reading and math than their same-grade peers. In reading, this advantage with respect to test score performance persisted through at least grade ten. These differences in academic preparation may have translated into differences in course-taking patterns and performance in high school.

Table 8 therefore presents estimates of the effect of test-based retention on a series of outcomes generated using students' high school transcripts, which are available for the same two cohorts for which we examined graduation outcomes. In particular, we study student GPAs; the number of courses students took offering remedial instruction in English language arts, reading, and math; and the number of courses students took that meet an admissions requirement for Florida universities (which we refer to as college prep).[26] The descriptive statistics for these outcomes reported in Table 2 indicate that students in these cohorts on average earned a 2.77 GPA and took 2.55 remedial and 5.74 college prep courses while enrolled in high school.

Column (1) of Table 8 shows that test-based retention in third grade increased students' GPAs by 0.067 grade points, or 12 percent of a standard deviation. Column (2) shows that retention also led students to take 1.61 fewer remedial courses overall, a sizable reduction relative to the average of 5.35 among students who scored below the cutoff but were not retained in third grade (see Table 2). Columns (3)-(5) reveal that this overall effect was driven primarily by a reduction of 1.26 remedial courses in reading, but that retention also led students to take 0.35 fewer remedial courses in math. Finally, column (6) shows that retention had no clear effect on the number of college prep courses students took. Overall, the results in Table 8 suggest that students who were retained in grade three performed modestly better in high school and required less reading and math remediation than would have been the case had they been promoted, but that they did not take more courses aligned to college admissions requirements.

[26] We calculate student GPAs by multiplying the numerical equivalent of the letter grade earned (i.e., A = 4, B = 3, etc.) in each course by the number of credits the course was worth, taking the sum across all of a given students courses, and dividing the sum by the total number of credits the student attempted.

4.5 The Effect of Test-Based Retention on College Enrollment

Table 9 presents estimates based on the 2003 cohort of the effects of test-based retention in third grade on a series of outcomes related to students' post-secondary enrollment patterns as of the 2013-14 school year. The outcome in column (1) is a binary indicator of enrollment in any post-secondary education institution in the state of Florida, while columns (2) and (3) respectively consider enrollment in four-year colleges and two-year community colleges in the state. The outcome in column (4) is again enrollment in any post-secondary education institution, but in this case the model is restricted to students who had graduated from a public high school in Florida.

The results show no significant relationship between third grade retention and the likelihood that a student enrolls in post-secondary education in Florida. That holds whether we look at attendance at any post-secondary institution or if we look separately at attendance at four-year or two-year schools. We also find no significant effect for the sub-sample of successful high school graduates in the 2003 cohort. The estimated coefficients in the three models based on the full sample are all very close to zero. The point estimate for the model based on the restricted sample is larger and positive, but also more imprecisely estimated.

These results suggest that retention under Florida's test-based promotion policy may not have influenced college enrollment patterns but need to be interpreted cautiously. The analysis is based on only the first cohort of students impacted by the policy, it would not capture the post-secondary enrollment of any students in that cohort who were retained under the policy and subsequently took longer than the expected number of years to graduate from high school, and it would not capture any enrollment in colleges outside of Florida. It is possible that some students retained under the policy will enroll in college later or in another state. That said, the results are consistent with the evidence in Table 8 suggesting that third grade retention helped students avoid the need for remediation in high school but did not lead them to take more courses aligned with college admissions requirements.

4.6 Sensitivity Analyses

The empirical results presented above are robust to a wide variety of alternative specification choices and validity checks. For example, Figure A-6 in the Supplementary Appendix confirms that our estimation results are stable across alternatives to the ten test-score-

point bandwidth ranging from five to 25 points on either side of the cutoff.[27] Table A-8 further shows that results are not influenced by the exclusion of students at or within one test score point of the promotion cutoff, are essentially unchanged when we use school fixed effects to restrict comparisons to students attending the same school in third grade, and are robust to the use of quadratic terms in modeling the relationship between third grade reading scores and the probability of retention on either side of the cutoff.

One potential concern with interpreting our results as the causal effect of test-based retention is the possibility of labeling effects (Papay et al., 2016). Students scoring below the cutoff are labeled as level 1 readers, while students above the cutoff are labeled as level 2 readers. Although there are no explicit consequences apart from the promotion decision of being a level 1 rather than a level 2 reader, these labels could alter the behavior of students, teachers, and parents in ways that affect students' subsequent achievement. To test whether labeling effects bias our estimates of test-based retention, we conduct a placebo test using the two cohorts of students in our data that entered third grade before 2003 and therefore were unaffected by the promotion policy. The results in Table A-9 confirm that being labeled a level 1 reader had no effect on future achievement for these students. Labeling effects are thus unlikely to confound our estimates of retention effects.

The analyses described so far focus on the local average treatment effect of test-based retention for all students performing at the promotion cutoff. This approach could conceal qualitative differences in effects across subgroups. For example, our results might be driven by large positive effects for specific subgroups, while grade retention is in fact detrimental for other students. In Tables A-10 and A-11 we address this concern by replicating our main analyses for subgroups defined based on their own characteristics or those of the schools they attended in third grade.

The results of these analyses provide little evidence of qualitative differences across student subgroups defined based on gender, ethnicity, or free/reduced-price lunch eligibility. Exploiting wide variation in the math achievement of Florida students who are retained on the basis of their reading test scores, we also document that the short-term benefits of test-based retention in both subjects are not limited to students achieving at a specific level.[28]

Similarly, we find little evidence of qualitative differences in the effects of test-based

[27] These alternatives more than encompass the informal sensitivity test suggested by Nichols (2007) of using twice and half the preferred bandwidth.

[28] Among students in our preferred bandwidth, 20,537 (27 percent) were classified as performing at level one (of five) based on the third grade math test, 26,357 (35 percent) performed at level two, and 29,253 (29 percent) performed at level three or higher.

retention across elementary schools categorized based on pupil/teacher ratio, expenditure per student, average teacher experience, and average teacher salary. There is some evidence, however, that the positive effects of test-based retention are more pronounced in schools with below-median retention rates. This could indicate that retained students receive more attention when there are fewer of them, potentially reinforcing any beneficial impact of test-based retention.

4.7 Potential Mechanisms

As discussed above, Florida requires that students retained under its test-based promotion policy receive remedial services intended to help them acquire the reading skills needed to be promoted the following year. These include the opportunity to attend a summer reading program prior to the next school year, assignment to a "high-performing" teacher, and intensive reading interventions during the retention year. Any of these program components could in theory account for part or all of the short-term academic gains we have documented for retained students.

Unfortunately, a lack of detailed information on the implementation and take-up of the policy's summer programming component makes it impossible to disentangle its separate effect. We note, however, that Matsudaira's (2008) regression discontinuity study of mandatory summer school for low-achieving grade 3-5 students in a large urban district finds average effects of 0.12 standard deviations in both reading and math. Jacob and Lefgren (2004) find that attending summer school after third grade improved the achievement of retained students in Chicago by 0.05 standard deviations in reading and 0.07 standard deviations in math after two years. Even if summer school attendance among students retained under Florida's policy were quite high, it is therefore unlikely that it accounts for more than a fraction of the short-term academic gains we observe for retained students.

We do have information on the teachers to which roughly 60 percent of the retained students were assigned in both their initial and repeated third grade year. Because the evaluation systems Florida school districts used during this period rated very few teachers as ineffective, the requirement that retained students be assigned to a high-performing teacher did not meaningfully constrain classroom placements. Even so, our data indicate that 94 percent of retained students were assigned to a different teacher during their retention year. Average class sizes for retained students also fell by almost two students, from 19.6 to 17.7, between their first and second years in the third grade.

In Table 10, we therefore use our regression discontinuity approach (same-grade com-

parison) to estimate the effect of being retained on two characteristics of the teachers to which students are assigned, as well as on their class size, in grades 3-5. The first row of Table 10 confirms that students retained in third grade are assigned to smaller classes compared to those non-retained students had experienced in third grade. Moreover, they are roughly 8 percentage points less likely to be assigned to a teacher with less than 2 years of experience. However, in grades 4 and 5 (rows 2 and 3) we no longer observe any significant differences with respect to class size or teacher experience. Nor do we find any evidence for systematic differences in grades 4 and 5 with respect to teacher quality as measured by value-added to student achievement.[29]

In sum, this evidence suggests that Florida schools did take steps to ensure that students were placed with different and possibly more effective teachers when repeating the third grade, but that any effects on teacher assignments were limited to that year. The lack of prior test scores for third grade students prevents us from constructing value-added estimates that would allow us to examine the effectiveness of third grade teachers directly. However, a recent review by Hanushek and Rivkin (2010) indicates that the within-school standard deviation of teacher value added to reading (math) test scores is, on average, 0.13 (0.17) standard deviations. Feasible improvements in teacher effectiveness during the retention year could therefore explain some of the short-term gains made by students retained under the Florida policy, but are unlikely in our view to be the only mechanism. Rather, it appears that the majority of the gains are attributable to the combination of a pure retention effect and whatever supplemental interventions students received during the retention year.

5 Conclusion

Our analysis exploits a discontinuity in the probability of grade retention under Florida's test-based promotion policy to study the policy's long-run effects on students retained in the third grade. Based on same-age comparisons, we find evidence of substantial short-

[29]We construct a single value-added measure for each math and reading teacher who could be linked to students in grades 4-5 that combines value-added estimates from all available years, grades, tests, and subjects. During our analysis period, Florida administered both the Florida Comprehensive Achievement Test and the Stanford Achievement Test in math and reading in these grades. In a given year, a teacher in a self-contained elementary classroom therefore has up to four separate value-added estimates. The methods used to construct these value-added estimates and average them across subjects, tests, and years are described in detail in Chingos and West (2012). We follow their procedures exactly, except that we exclude estimates based on the year for which the teacher assignment is the outcome when calculating teachers' average effectiveness.

term gains in both math and reading achievement. However, these positive effects fade out over time and become statistically insignificant within five years. We also find that test-based retention (and remediation) in third grade substantially reduces the probability of being retained in later grades and has no impact on the probability of graduating from high school.

In sum, our analysis provides more favorable evidence on the effects of early grade retention than found in many previous studies—in particular those that do not rely on credible quasi-experimental methods to address unobserved selection into the retention treatment. We show that test-based retention has substantial positive effects on reading and math achievement in the short run, has no detrimental effects on the limited set of outcomes we can measure, and leads students to perform at higher levels against grade-level expectations and need less remediation while in high school. To the extent that early grade retention is more beneficial than later grade retention (as suggested by the results of Jacob and Lefgren, 2004, 2009), students who were retained in third grade and would have been retained later clearly benefited from the introduction of the Florida policy. However, we also do not provide definitive evidence that test-based retention in early grades is beneficial for students in the long run, even when it is accompanied by the requirement that students receive additional services.

The fade out of test score impacts is a common pattern in the literature on educational interventions, including those which have been shown to generate lasting impacts on adult outcomes. For example, Chetty et al. (2011) show that kindergarten classroom quality improves college enrollment and adult earnings despite the complete fade out of short-term test score gains. The same appears to be true of early childhood interventions such as the Perry and Abecederian preschool demonstration projects and the Head Start program (see Almond and Currie [2011] for a review). Whether students retained under Florida's test-based promotion policy will also experience benefits as adults remains uncertain. Test-based retention led students to perform better academically when in high school, but this advantage did not translate into improved graduation rates for the first two cohorts of students affected by the policy. An analysis of the effects of test-based retention on post-secondary attainment and labor market outcomes should be feasible in Florida within a few years.

The Florida policy we have analyzed in this paper has emerged as a model for policymakers in other states. Arizona, Indiana, Oklahoma, and Ohio enacted test-based promotion policies modeled on Florida's between 2010 and 2012, and similar bills have

been introduced in the legislatures of several other states. In light of this interest, we should emphasize that their consequences for retained students are only one component of a comprehensive analysis of these policies' merits. Test-based promotion policies also aim to provide incentives for educators and parents to improve the skills of low-performing students prior to third grade. There are also a variety of other potential mechanisms, such as the creation of grade cohorts that are more homogenous with respect to student achievement, that could influence outcomes for higher-performing students. The broader consequences of policies influencing retention rates have received little attention from researchers and deserve further scrutiny.

References

Allen, C. S., Chen, Q., Willson, V. L., and Hughes, J. N. (2009). Quality of research design moderates effects of grade retention on achievement: A meta-analytic, multilevel analysis. *Educational Evaluation and Policy Analysis*, 31(4):480–499.

Almond, D. and Currie, J. (2011). Human capital development before age five. In Ashenfelter, O. and Card, D., editors, *Handbook of Labor Economics*, volume 4b, pages 1315–1486. Elsevier.

Babcock, P. and Bedard, K. (2011). The wages of failure: New evidence on school retention and long-run outcomes. *Education Finance and Policy*, 6(3):293–322.

Cascio, E. U. and Staiger, D. O. (2012). Knowledge, tests, and fadeout in educational interventions. NBER Working Paper 18038.

Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W., and Yagan, D. (2011). How does your kindergarten classroom affect your earnings? Evidence from project star. *The Quarterly Journal of Economics*, 126(4):1593–1660.

Chingos, M. M. and West, M. R. (2012). Do more effective teachers earn more outside of education? *Education Finance and Policy*, 7(1):8–43.

Eide, E. R. and Showalter, M. H. (2001). The effect of grade retention on educational and labor market outcomes. *Economics of Education Review*, 20(6):563–576.

Foorman, B. R., Kershaw, S., and Petscher, Y. (2013). *Evaluating the screening accuracy of the Florida assessments for instruction in reading (FAIR) (REL 2013008).* Number Retrieved from http://ies.ed.gov/ncee/edlabs. Washington, DC: U.S. Department of Education, Institute of EducationSciences.

Greene, J. P. and Winters, M. A. (2007). Revisiting grade retention: An evaluation of Florida's test-based promotion policy. *Education Finance and Policy*, 2(4):319–340.

Hanushek, E. A. and Rivkin, S. G. (2010). Generalizations about using value-added measures of teacher quality. *American Economic Review*, 100(2):267–271.

Hoffman, R., Wise, L. L., and Thacker, A. A. (2001). Florida comprehensive assessment test: Technical report on vertical scaling for reading and mathematics. Technical report, San Antonio, TX: Harcourt Educational Measurement.

Holmes, C. T. (1989). Grade level retention effects: A meta-analysis of research studies. In Shepard, L. A. and Smith, M. L., editors, *Flunking Grades: Research and Policies on Retention*, pages 16–33. New York: The Falmer Press.

Imbens, G. and Kalyanaraman, K. (2012). Optimal bandwidth choice for the regression discontinuity estimator. *The Review of Economic Studies*, 79(3):933–959.

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

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

Jimerson, S. R. (1999). On the failure of failure: Examining the association between early grade retention and education and employment outcomes during late adolescence. *Journal of School Psychology*, 37(3):243–272.

Jimerson, S. R., Anderson, G. E., and Whipple, A. D. (2002). Winning the battle and losing the war: Examining the relation between grade retention and dropping out of high school. *Psychology in the Schools*, 39(4):441–457.

Lang, K. (2010). Measurement matters: Perspectives on education policy from an economist and school board member. *Journal of Economic Perspectives*, 24(3):167182.

Lee, D. S. and Card, D. (2008). Regression discontinuity inference with specification error. *Journal of Econometrics*, 142(2):655–674.

Lee, D. S. and Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of Economic Literature*, 48(2):281–355.

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

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

McCoy, A. R. and Reynolds, A. J. (1999). Grade retention and school performance: An extended investigation. *Journal of School Psychology*, 37(3):273–298.

Nichols, A. (2007). Causal inference with observational data. *Stata Journal*, 7(4):507–541.

Office of Program Policy Analysis & Government Accountability (2006). Third grade retention leading to better student performance statewide. OPPAGA Report 06-66, http://www.oppaga.state.fl.us/reports/pdf/0666rpt.pdf.

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

Papay, J. P., Murnane, R. J., and Willett, J. B. (2016). The impact of test score labels on human-capital investment decisions. *Journal of Human Resources*, 51(2):357–388.

Planty, M., Hussar, W., Snyder, T., Kena, G., Ramani, A. K., Kemp, J., Bianco, K., and Dinkes, R. (2009). *The Condition of Education 2009*. (NCES 2009-081). Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

Schwerdt, G. and West, M. R. (2013a). The effects of test-based retention on student outcomes over time: Regression discontinuity evidence from Florida. CESifo Working Paper Series 4203.

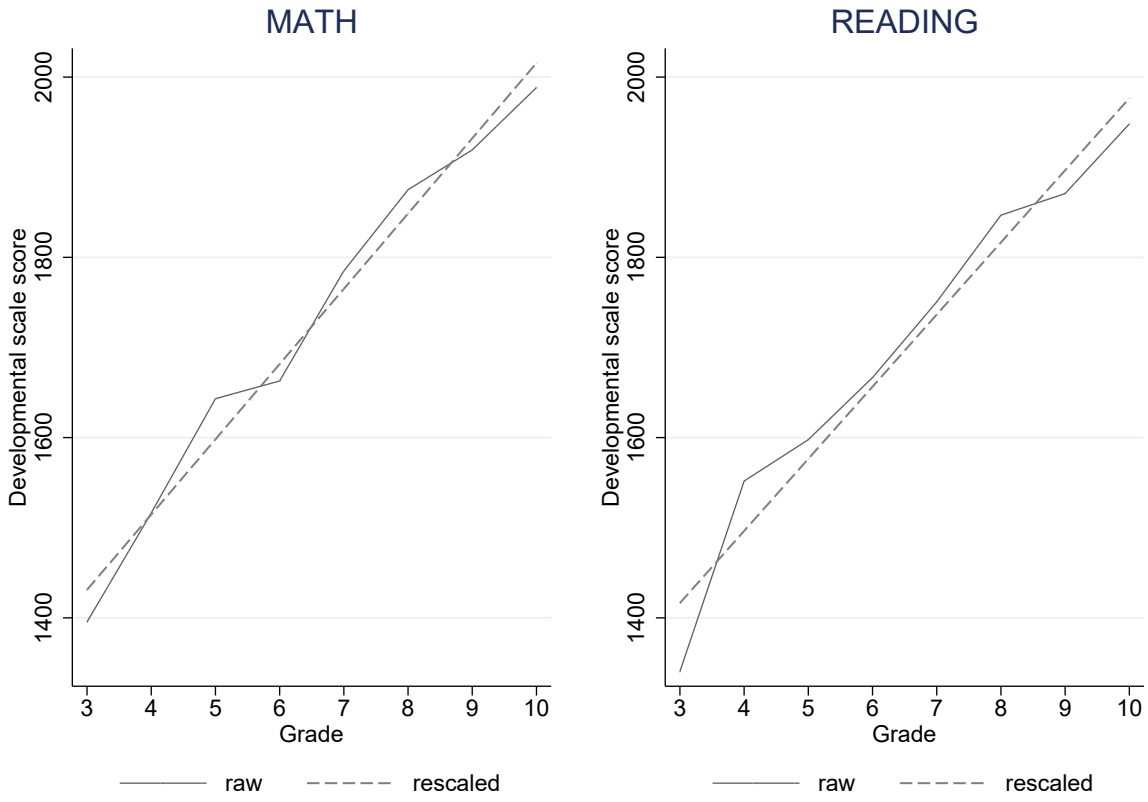
Schwerdt, G. and West, M. R. (2013b). The impact of alternative grade configurations on student outcomes through middle and high school. *Journal of Public Economics*, 97(C):308–326.

Urquiola, M. and Verhoogen, E. (2009). Class-size caps, sorting, and the regression-discontinuity design. *American Economic Review*, 99(1):179–215.

Winters, M. A. and Greene, J. P. (2012). The medium-run effects of Floridas test-based promotion policy. *Education Finance and Policy*, 7(3):305–330.

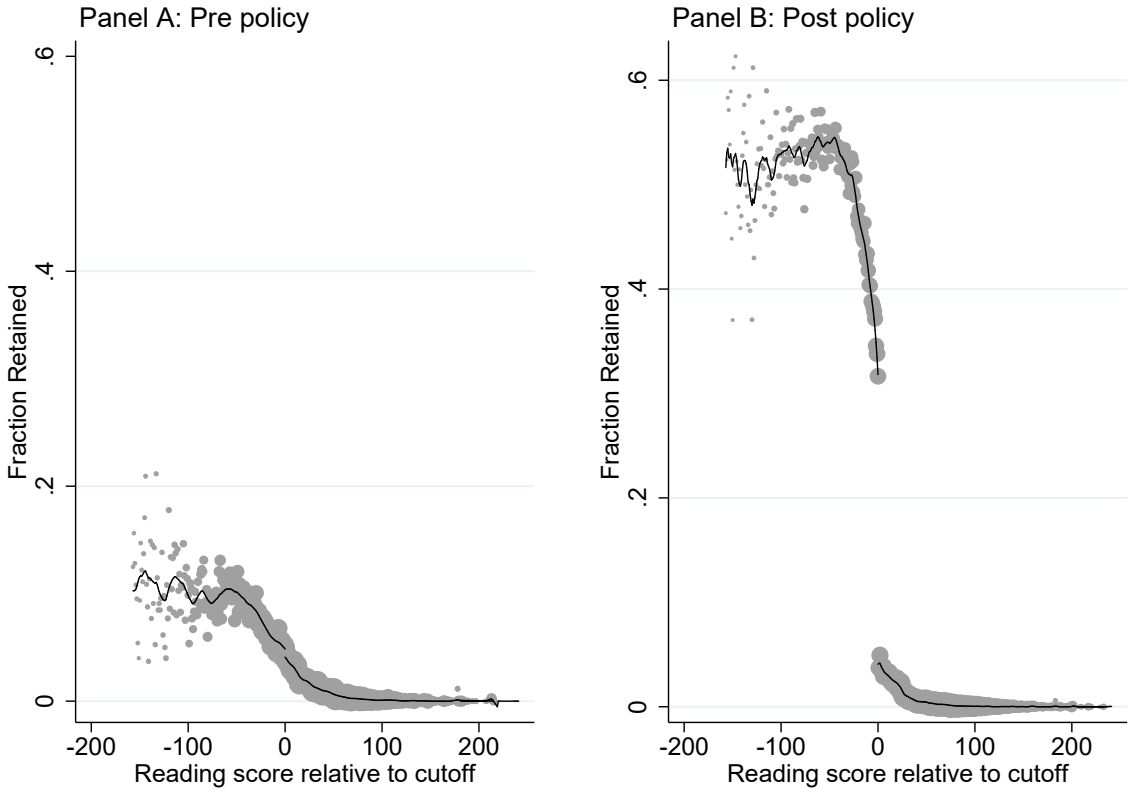
Workman, E. (2014). Third grade reading policies. Technical Report, Denver, CO: Education Commission of the States.

Figure 1: Average Developmental Scale Scores by Subject and Grade



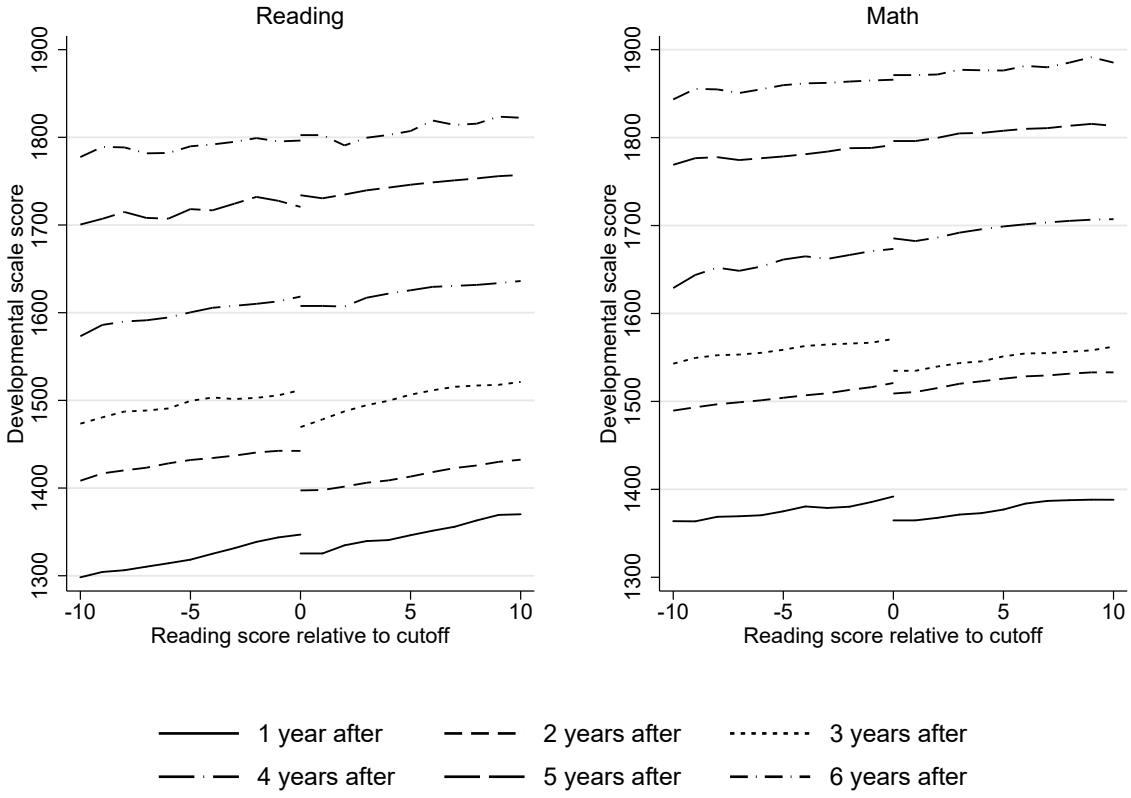
Note: Based on all students in grades 3 to 10 between 2002 and 2009. Rescaled scores stem from predicted values of a linear regression of developmental scale scores on grade levels.

Figure 2: The Relationship between Grade 3 Reading Scores and the Probability of Grade 3 Retention



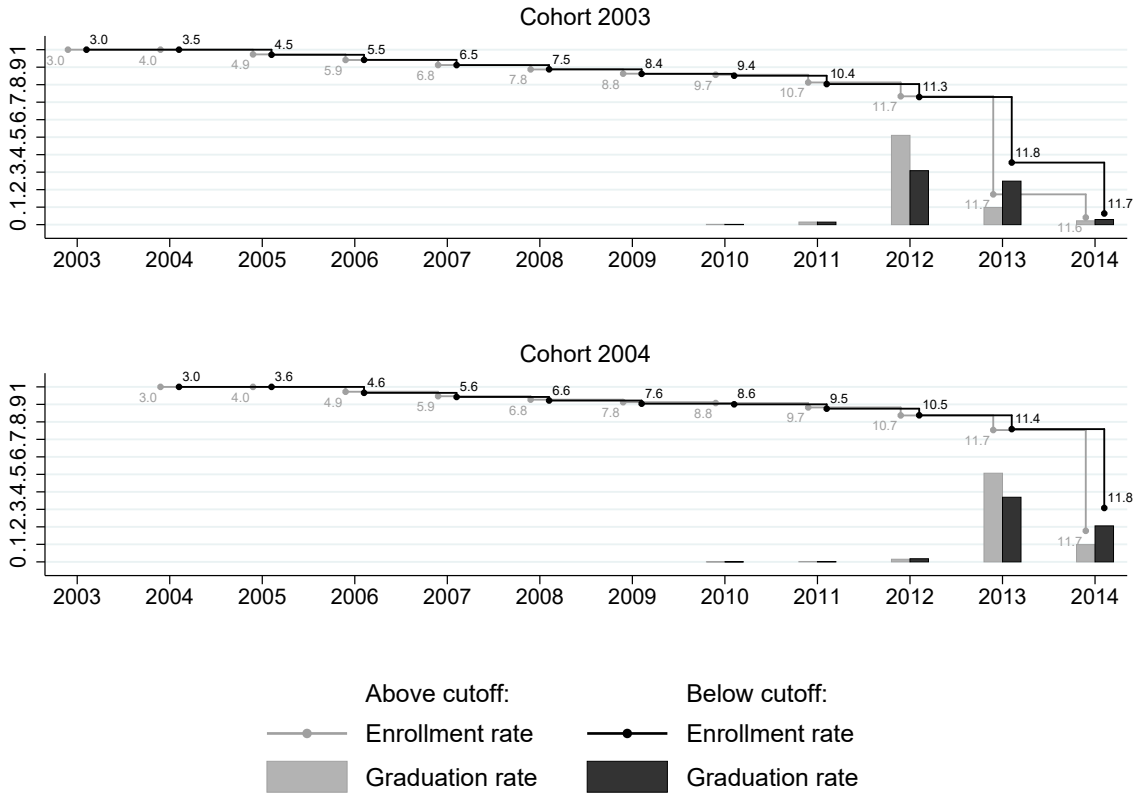
Note: Panel A based on 2001-2002 cohorts; panel B based on 2003-2008 cohorts. Full sample. Solid line represents predicted values from local linear regressions on both sides of the cutoff. Marker size represents relative group size.

Figure 3: The Relationship between Reading Scores in Grade 3 and Future Achievement around the Cutoff



Note: Based on 2003-2008 cohorts. Discontinuity sample with 10-point bandwidth. Lines represent predicted values from local linear regressions on both sides of the cutoff.

Figure 4: School History of the 2003 and 2004 Cohorts



Note: Based on discontinuity sample with 10-point bandwidth for the 2003 and 2004 cohorts. Figure displays enrollment rates and annual high school graduation rates by school-year for students above and below the grade 3 retention cutoff with respect to the group size in third grade. High school graduation is defined as receiving a regular diploma, special diploma, certificate of completion, or GED. Marker labels indicate average grade levels of students enrolled in each year.

Table 1: Summary Statistics: Grade 3 Characteristics

Grade 3 characteristic	Total	Below cutoff /retained	Below cutoff /promoted	Above cutoff /retained	Above cutoff /promoted
DSS Math	1,413	998	1,084	1,128	1,488
	(306)	(266)	(279)	(225)	(254)
DSS Reading	1,373	770	807	1,196	1,488
	(370)	(242)	(242)	(135)	(269)
Female	0.49	0.42	0.42	0.46	0.51
Age	8.84	8.90	9.21	8.77	8.80
	(0.60)	(0.64)	(0.74)	(0.62)	(0.57)
White	0.48	0.26	0.29	0.50	0.51
Black	0.22	0.40	0.35	0.29	0.19
Hispanic	0.24	0.30	0.31	0.15	0.23
Asian	0.02	0.01	0.01	0.01	0.02
Other	0.04	0.03	0.03	0.04	0.04
Free or reduced lunch	0.52	0.79	0.76	0.65	0.47
Limited English proficiency	0.19	0.30	0.31	0.11	0.17
Special Education	0.16	0.30	0.43	0.15	0.11
Days absent	7.46	9.23	8.97	10.13	7.13
	(7.48)	(9.22)	(8.75)	(9.74)	(7.09)
Number of students	983,308	76,398	83,468	4,959	818,483

Note: Based on full sample for the 2003-2008 cohorts. Means (and standard deviations) for the grade 3 characteristics indicated in each row.

Table 2: Summary Statistics: High School and College Outcomes

Outcome - [*sample*]	Total	Below cutoff /retained	Below cutoff /promoted	Above cutoff /retained	Above cutoff /promoted
Grade 9 FCAT2.0 Reading	241	220	218	229	245
- *Cohorts: 2005-2007*	(21)	(17)	(18)	(16)	(19)
Grade 10 FCAT2.0 Reading	246	226	226	234	250
- *Cohorts: 2004-2006*	(20)	(16)	(17)	(15)	(18)
Ever enter High School	0.86	0.83	0.85	0.79	0.87
- *Cohorts: 2003-2004*					
Graduation	0.70	0.49	0.57	0.50	0.75
- *Cohorts: 2003-2004*					
GPA	2.77	2.44	2.50	2.47	2.84
- *Cohorts: 2003-2004*	(0.58)	(0.53)	(0.53)	(0.54)	(0.57)
Remedial Courses	2.55	4.71	5.35	3.84	1.97
- *Cohorts: 2003-2004*	(3.64)	(3.92)	(4.56)	(4.10)	(3.22)
College Prep Courses	5.74	4.86	4.94	2.43	5.76
- *Cohorts: 2003-2004*	(9.43)	(7.81)	(8.46)	(6.34)	(9.73)
College Enrollment	0.39	0.21	0.24	0.20	0.44
- *Cohort: 2003*					

Note: Based on full sample. Means (and standard deviations) for outcomes and cohorts indicated in each row.

Table 3: Effect of Reading Performance on the Probability of Retention in Grade 3

	Pre Policy		Post Policy						
Cohorts	2001	2002	2003	2004	2005	2006	2007	2008	2003-2008
Below cutoff	0.006	0.019***	0.373***	0.268***	0.295***	0.338***	0.198***	0.217***	0.283***
	[0.006]	[0.007]	[0.012]	[0.013]	[0.013]	[0.016]	[0.012]	[0.013]	[0.005]
Reading	–0.000	–0.001	–0.001	–0.000	–0.001	–0.001	–0.001	0.001	–0.001*
	[0.001]	[0.001]	[0.001]	[0.001]	[0.001]	[0.001]	[0.001]	[0.001]	[0.000]
Reading	0.000	0.000	–0.010***	–0.009***	–0.006***	–0.005**	–0.007***	–0.008***	–0.008***
× Below cutoff	[0.001]	[0.001]	[0.002]	[0.002]	[0.002]	[0.002]	[0.002]	[0.002]	[0.001]
Performance and demographic covariates	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	No	No	No	No	No	No	No	No	Yes
Students	17,676	16,516	15,687	12,040	12,435	9,981	12,995	11,536	74,674
R^2	0.018	0.020	0.297	0.207	0.229	0.253	0.158	0.169	0.227
F-statistic on instrument	0.92	8.01	895.72	402.03	497.35	473.78	289.23	292.30	2,778.43
Pr > F	0.34	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

* p<0.10, ** p<0.05, *** p<0.01

Note: OLS estimates. Based on discontinuity sample with 10-point bandwidth. Dependent variable is a dummy indicating retention in grade 3 in all columns. The table displays estimates with performance and demographic covariates of students. Performance and demographic covariates include math scores in grade 3, gender, age, race, special education status in grade 3, LEP status in grade 3, and free or reduced-price lunch status in grade 3. Standard errors clustered by third grade school and by third grade reading score in parentheses.

Table 4: Effect of Retention in Grade 3 on Student Achievement

	Same Grade Comparison		Same Age Comparison	
Dependent Variable measured in	DSS/FCAT2.0 (1)	Dependent Variable measured after	DSS/FCAT2.0 (2)	rescaled DSS (3)
Reading				
Based on vertically scaled FCAT scores (SD= 370)				
		1 year (n = 74,443)	83.64***	214.90***
grade 4 (n = 76,208)	269.07***		(8.67)	(8.67)
	(12.98)	2 years (n = 70,596)	182.23***	152.03***
grade 5 (n = 59,562)	204.48***		(11.24)	(11.28)
	(9.58)	3 years (n = 57,122)	97.64***	88.33***
grade 6 (n = 45,804)	159.08***		(11.20)	(11.15)
	(13.93)	4 years (n = 43,909)	37.55***	40.56***
grade 7 (n = 35,051)	102.43***		(10.96)	(10.97)
	(16.43)	5 years (n = 34,311)	1.71	13.16
grade 8 (n = 23,253)	69.90***		(13.83)	(13.95)
	(9.67)	6 years (n = 22,999)	39.82***	4.29
			(14.38)	(14.37)
Based on FCAT 2.0 scores (SD= 21)				
grade 9 (n = 28,939)	7.48***			
	(.85)	7 years (n = 27,063)	.29	
grade 10 (n = 24,944)	4.79***		(.74)	
	(1.00)			
Math				
Based on vertically scaled FCAT scores (SD= 306)				
		1 year (n = 74,327)	92.51***	129.97***
grade 4 (n = 76,091)	186.25***		(9.75)	(9.75)
	(8.56)	2 years (n = 70,596)	34.06***	72.48***
grade 5 (n = 59,334)	133.21***		(4.34)	(4.29)
	(7.76)	3 years (n = 57,042)	110.10***	58.62***
grade 6 (n = 45,760)	159.70***		(7.47)	(7.42)
	(13.89)	4 years (n = 43,884)	–23.58**	5.78
grade 7 (n = 35,057)	105.17***		(9.83)	(9.94)
	(16.07)	5 years (n = 34,290)	–22.69***	–16.99***
grade 8 (n = 23,230)	40.97***		(5.69)	(5.79)
	(8.22)	6 years (n = 22,977)	–7.77	–32.60***
			(7.21)	(7.23)

* p<0.10, ** p<0.05, *** p<0.01

Note: IV estimates. Based on discontinuity sample with 10-point bandwidth. Dependent variables are unadjusted developmental scale scores in reading and math in columns (1) and (2) and rescaled developmental scale scores in reading and math in column (3); reported standard deviations for developmental scale scores (DSS) are for grade 3, while standard deviations for FCAT 2.0 scores are for grade 10 (see Foorman et al. (2013)). All estimations control for a linear function in grade 3 reading scores that allows for different trends on both sides of the cutoff and cohort dummies. The table displays IV estimates with performance and demographic covariates of students. Performance and demographic covariates include math scores in grade 3, gender, age, race, special education status in grade 3, LEP status in grade 3, and free or reduced-price lunch status in grade 3. Standard errors clustered by third grade school and by third grade reading score in parentheses.

Table 5: Effect of Retention in Grade 3 on Grade Progression

Dependent Variable	Retention Probability	Grade Level
	(1)	(2)
2 years (n = 72,644)	–.11***	–.88***
	(.01)	(.01)
3 years (n = 70,811)	–.03***	–.83***
	(.01)	(.01)
4 years (n = 69,237)	–.04***	–.78***
	(.00)	(.01)
5 years (n = 67,933)	–.02***	–.73***
	(.01)	(.01)

* p<0.10, ** p<0.05, *** p<0.01

Note: IV estimates. Based on discontinuity sample with 10-point bandwidth. Dependent variable is a dummy indicating grade retention in the top panel and the student's grade level in the bottom panel. All estimations control for a linear function in grade 3 reading scores that allows for different trends on both sides of the cutoff and cohort dummies. The table displays IV estimates with performance and demographic covariates of students. Performance and demographic covariates include math scores in grade 3, gender, age, race, special education status in grade 3, LEP status in grade 3, and free or reduced-price lunch status in grade 3. Standard errors clustered by third grade school and by third grade reading score in parentheses.

Table 6: The Effect of Retention in Grade 3 on High School Completion

Outcome	High school entry	High school graduation				High school graduation year
		Type of diploma				
		Any	Regular	GED	CoC	
			conditional on grad.			
	(1)	(2)	(3)	(4)	(5)	(6)
Retained	–.006	–.003	.005	.006	.018	.628***
in grade 3	(.020)	(.036)	(.034)	(.015)	(.026)	(.052)
Reading	.001	.000	–.003***	–.001***	–.001**	–.004*
	(.001)	(.002)	(.001)	(.000)	(.001)	(.002)
Reading	–.000	.003	.004***	.002***	.002*	.004**
× Below cutoff	(.001)	(.002)	(.001)	(.001)	(.001)	(.002)
Students	27,724	27,724	17,147	17,147	17,147	17,147
R^2	.015	.056	.036	.031	.038	.561

* p<0.10, ** p<0.05, *** p<0.01

Note: IV estimates. Based on discontinuity sample with 10-point bandwidth for the 2003 and 2004 cohorts. Dependent variables: dummy indicating whether students enter grade 9 by 2013-14 in column (1); dummy indicating whether students complete high school by 2013-14 in column (2); dummy indicating whether students obtained a regular high school degree (column (3)), a GED (column (4)), or a certificate of completion (column (5)) conditional on graduation by 2013-14; year of high school graduation conditional on graduation in column (6). The table displays IV estimates with performance and demographic covariates of students. Performance and demographic covariates include math scores in grade 3, gender, age, race, special education status in grade 3, LEP status in grade 3, and free or reduced-price lunch status in grade 3. Standard errors clustered by third grade school and by third grade reading score in parentheses.

Table 7: Effect of Grade Retention on Highest Grade Completed and School Leaving Age

Outcome:	Highest grade completed		School leaving age		Left at age 16		Left at age 17		Left at age 18+	
Cohort(s)	2003	2003/04	2003	2003/04	2003	2003/04	2003	2003/04	2003	2003/04
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Retained in grade 3	−.079	−.116	.585***	.542***	−.010	−.015*	−.028	−.027**	.031	.036*
	(.126)	(.121)	(.137)	(.150)	(.012)	(.008)	(.020)	(.013)	(.022)	(.019)
Reading	.011	.008*	.008	.005	−.000	−.000	−.001	−.001*	.001	.001**
	(.008)	(.005)	(.007)	(.006)	(.000)	(.000)	(.001)	(.001)	(.001)	(.001)
Reading	−.005	−.003	−.001	.002	−.000	−.001	−.000	.000	.001	.000
× Below cutoff	(.009)	(.005)	(.008)	(.006)	(.001)	(.001)	(.001)	(.001)	(.001)	(.001)
Students	15,687	27,724	15,687	27,724	15,687	27,724	15,687	27,724	15,687	27,724
R^2	.028	.029	.032	.032	.004	.002	.002	.003	.015	.014

* p<0.10, ** p<0.05, *** p<0.01

Note: Based on discontinuity sample with 10-point bandwidth for the 2003 and 2004 cohorts. Dependent variables: highest grade completed in columns (1) and (2); age in highest grade completed in columns (3) and (4); dummy indicating that age in highest grade completed was equal to 16 in columns (5) and (6); dummy indicating that age in highest grade completed was equal to 17 in columns (7) and (8); dummy indicating that age in highest grade completed was 18 or above in columns (9) and (10). The table displays IV estimates with performance and demographic covariates of students. Performance and demographic covariates include math scores in grade 3, gender, age, race, special education status in grade 3, LEP status in grade 3, and free or reduced-price lunch status in grade 3. Standard errors clustered by third grade school and by third grade reading score in parentheses.

Table 8: The Effect of Retention in Grade 3 on High School GPA and Course-taking

Outcome	GPA	Remedial				College
		All	ELA	Read	Math	Prep
	(1)	(2)	(3)	(4)	(5)	(6)
Retained	0.067*	-1.610***	-0.001	-1.263***	-0.345***	0.092
in grade 3	(0.036)	(0.366)	(0.078)	(0.264)	(0.099)	(0.646)
Reading	0.001	-0.055***	-0.004	-0.045***	-0.005	0.051
	(0.001)	(0.016)	(0.003)	(0.013)	(0.004)	(.009)
Reading	-0.001	0.022	0.004	0.021	-0.003	-0.065***
× Below cutoff	(0.002)	(0.020)	(0.003)	(0.013)	(0.007)	(0.023)
Students	23,642	23,816	23,816	23,816	23,816	23,816
R^2	0.203	0.093	0.010	0.092	0.030	0.099

* p<0.10, ** p<0.05, *** p<0.01

Note: IV estimates. Based on discontinuity sample with 10-point bandwidth for the 2003 and 2004 cohorts. Dependent variables: students grade point average as of 2013-14 in column (1), the number of remedial courses of any type, in ELA, Reading, and Math as of 2013-14 in columns (2)-(5), and the number of courses taken classified as college preparatory as of 2013-14 in column (6). Standard errors clustered by third grade school and by third grade reading score in parentheses.

Table 9: The Effect of Retention in Grade 3 on College Enrollment

Outcome	College Enrollment			
	Any type	Four-year university	Community college	Any type *(conditional on high school graduation)*
	(1)	(2)	(3)	(4)
Retained in grade 3	.004	.015	–.007	.031
	(.036)	(.030)	(.013)	(.055)
Reading	.002	.002	.000	.003*
	(.002)	(.001)	(.001)	(.002)
Reading	–.002	–.002	–.001	–.005
× Below cutoff	(.002)	(.002)	(.001)	(.003)
Students	15,687	15,687	15,687	9,816
R^2	.068	.048	.030	.060

* p<0.10, ** p<0.05, *** p<0.01

Note: IV estimates. Based on discontinuity sample with 10-point bandwidth for the 2003 cohort. Dependent variables: dummy indicating whether students are enrolled in college in 2013-14. The table displays IV estimates with performance and demographic covariates of students. Performance and demographic covariates include math scores in grade 3, gender, age, race, special education status in grade 3, LEP status in grade 3, and free or reduced-price lunch status in grade 3. Standard errors clustered by third grade school and by third grade reading score in parentheses.

Table 10: Mechanisms: IV estimates of the Effect of Retention in Grade 3 on Teacher Assignment and Class Size in Elementary School Grades

Outcome	Teacher quality estimates based on math scores	Teacher quality estimates based on reading scores	Teacher experience (in years)	Teacher with $\leq$ 2 years of experience	Class size
	(1)	(2)	(3)	(4)	(5)
Grade 3	*n.a.*	*n.a.*	–.077	–.083**	–1.558***
			(.769)	(.036)	(.315)
Grade 4	–.023	–.009	.397	–.024	–.272
	(.016)	(.013)	(.828)	(.038)	(.351)
Grade 5	.007	–.004	–1.192	.018	.005
	(.018)	(.013)	(.952)	(.040)	(.392)

* p<0.10, ** p<0.05, *** p<0.01

Note: IV estimates. Based on discontinuity sample with 10-point bandwidth. Dependent variable indicated in first row. Grade 3 refers to the retention year for students retained in grade 3. All IV estimations control for math scores, gender, age, race, free or reduced-price lunch status in grade 3. Standard errors clustered by third grade school and by third grade reading score in parentheses.

Supplementary Appendix

The Effects of Test-based Retention on Student Outcomes over Time: Regression Discontinuity Evidence from Florida

Guido Schwerdt* Martin R. West†§ Marcus A. Winters‡

For online publication only

This supplementary appendix provides additional information to the paper:

Section 1 provides evidence on the internal validity of the regression discontinuity design.

Section 2 provides a comparison with OLS estimation results and provides estimation results without covariates.

Section 3 provides estimation results by cohort.

Section 4 provides results of several robustness tests and explores potential qualitative differences in the effect of grade retention across subgroups.

*University of Konstanz, Department of Economics, Box D133, 78457 Konstanz, Germany. Email: guido.schwerdt@uni-konstanz.de

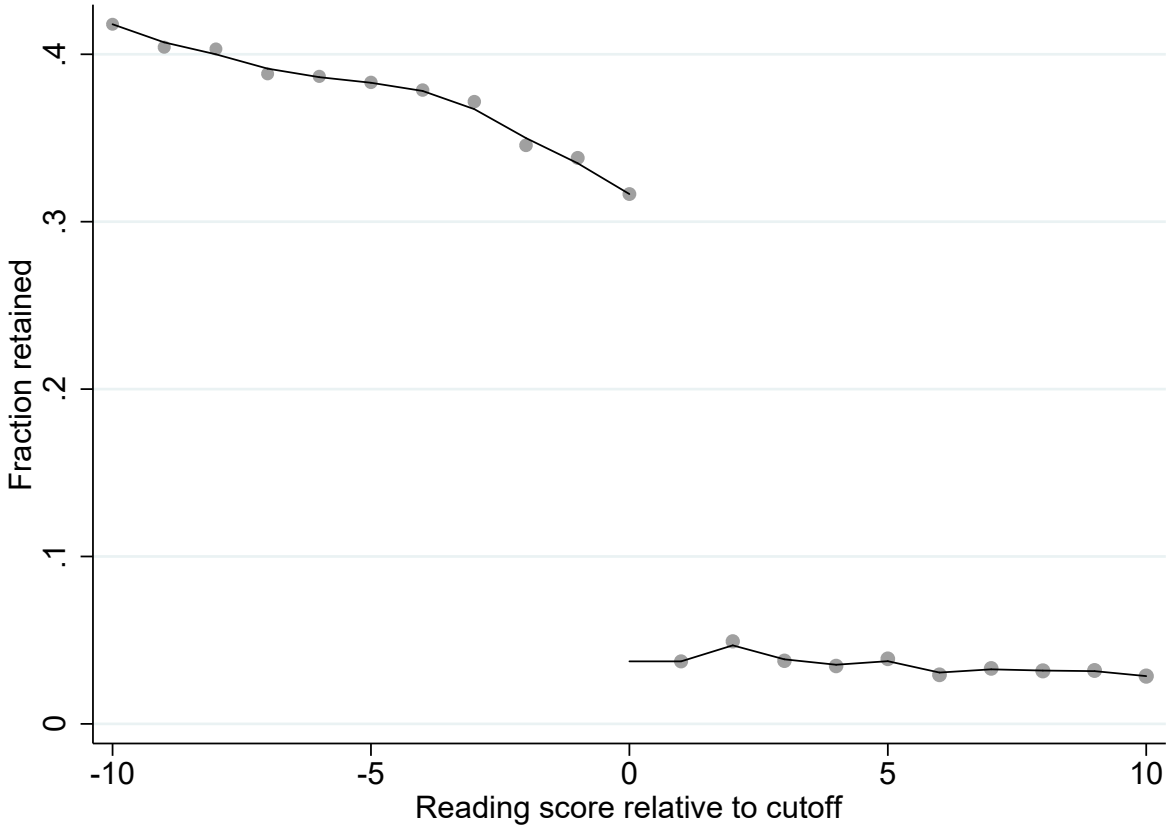
†Harvard Graduate School of Education, 6 Appian Way, Gutman 454, Cambridge, MA 02138, USA. Email: martin_west@gse.harvard.edu

‡Boston University, School of Education, 2 Silber Way, Boston, MA 02215, USA. Email: marcusw@bu.edu

§Corresponding author

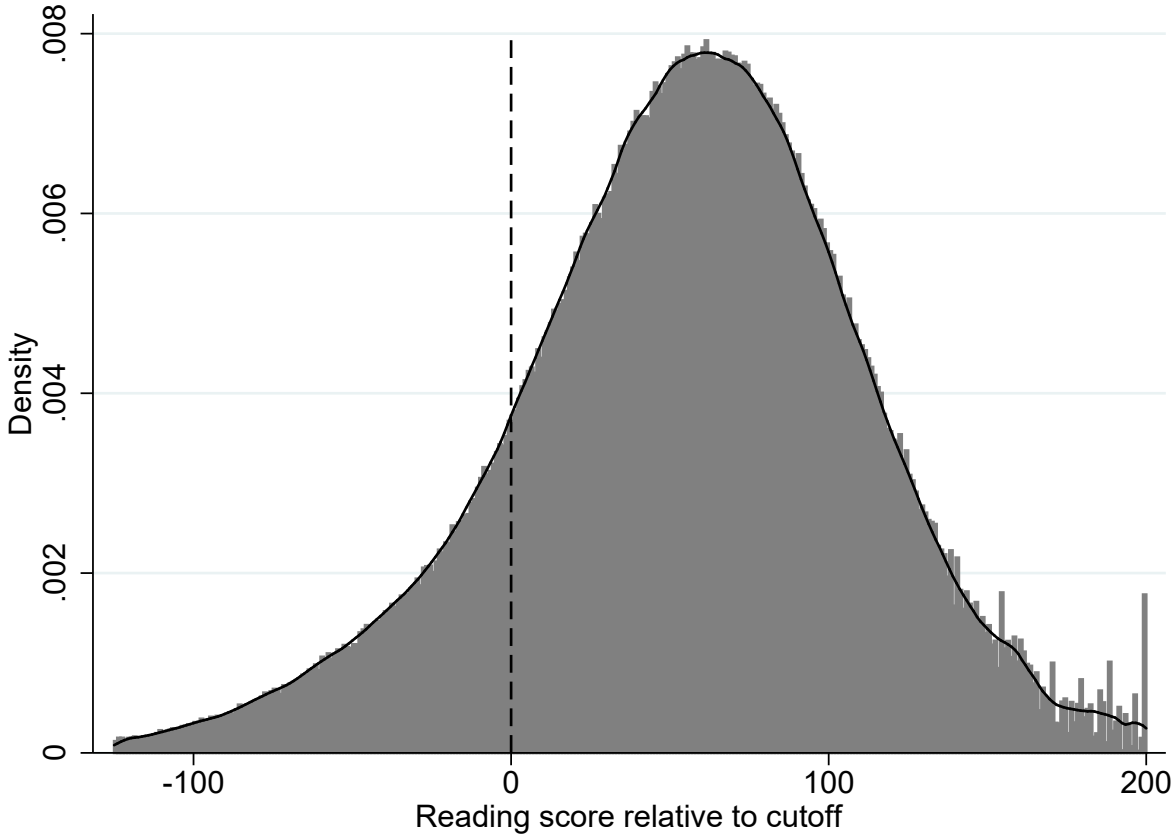
1. Evidence on the Validity of the Empirical Strategy

Figure A-1: The Relationship between Reading Scores and Grade Retention around the Cutoff



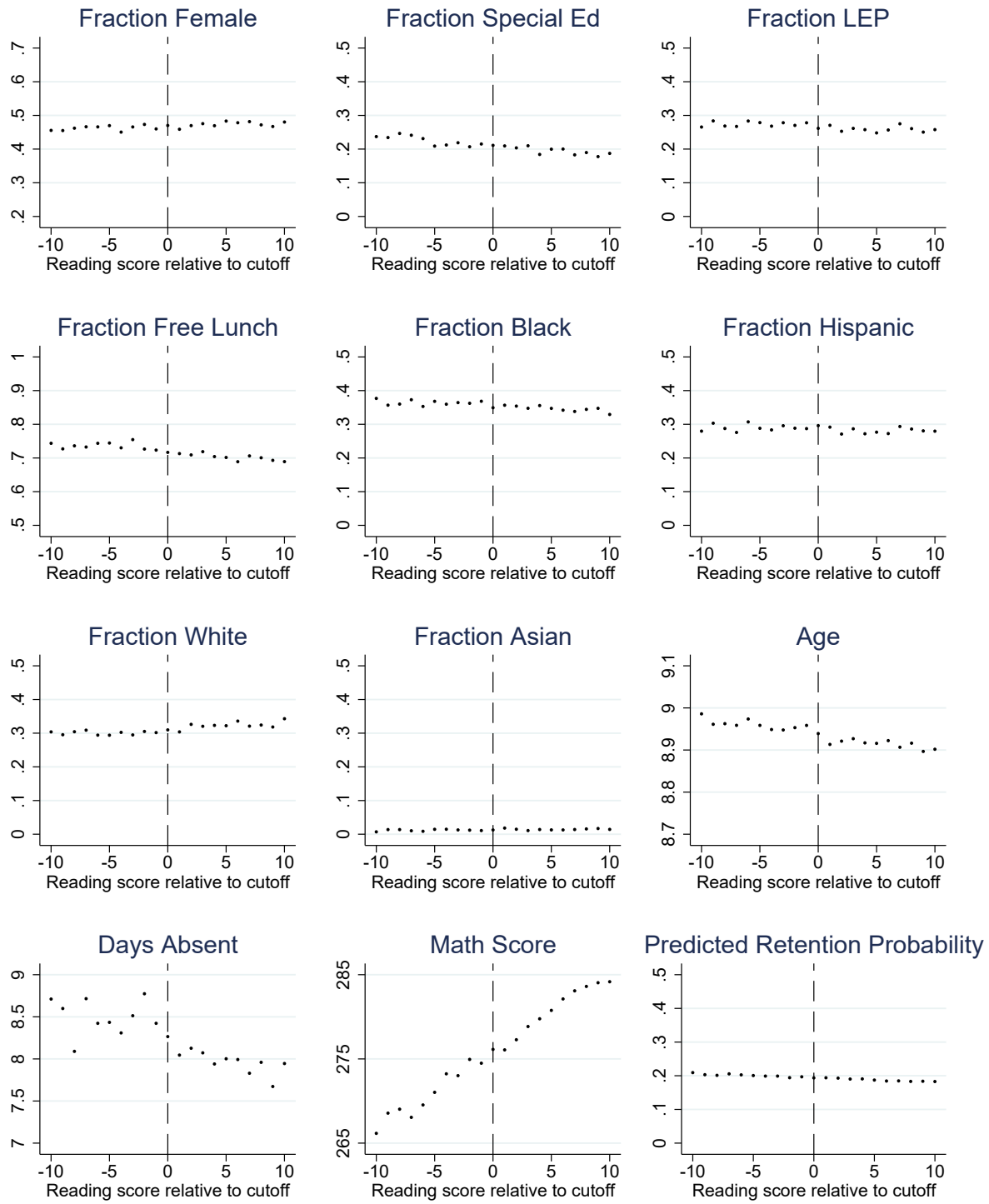
Note: Based on 2003-2008 cohorts. Discontinuity sample with 10-point bandwidth. Solid line represents predicted values from local linear regressions on both sides of the cutoff. Marker size represents relative group size.

Figure A-2: Distribution of Reading Scores in Grade 3



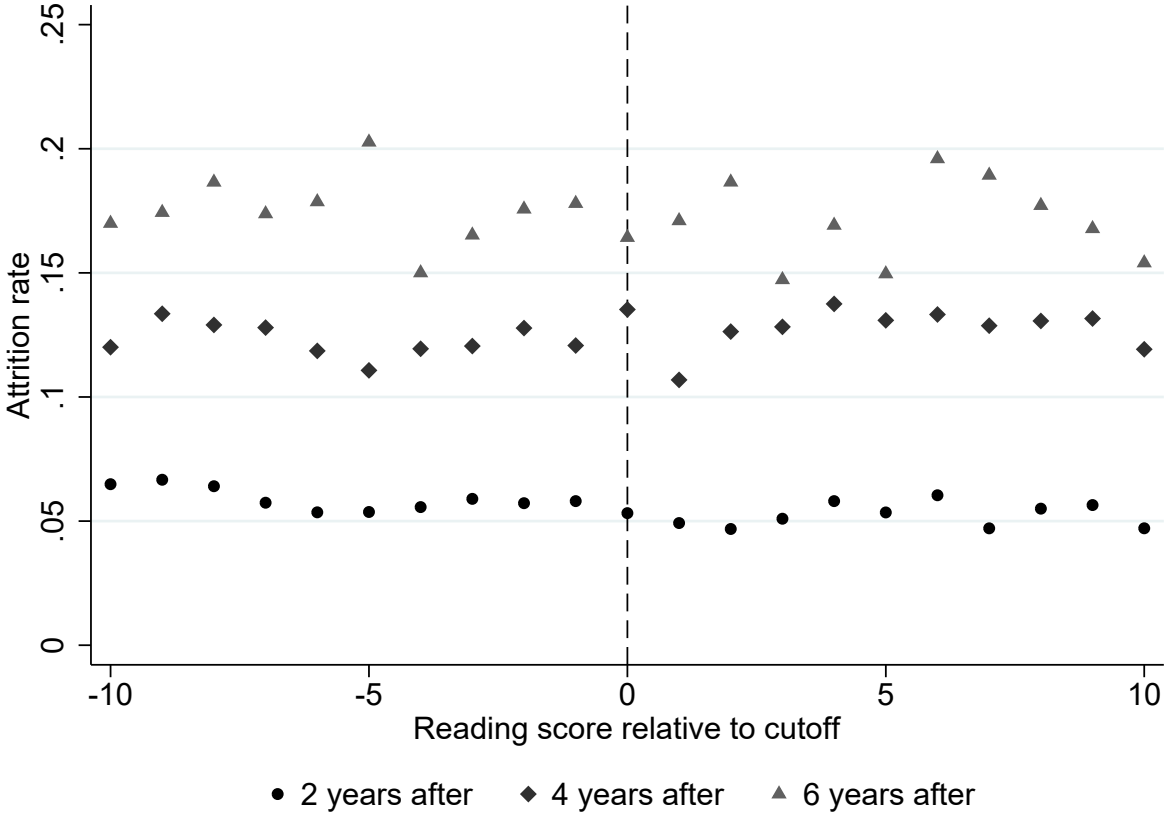
Note: Based on 2003-2008 cohorts. Full sample. Solid line represents kernel density estimates.

Figure A-3: The Relationship between Reading Scores in Grade 3 and Student Characteristics



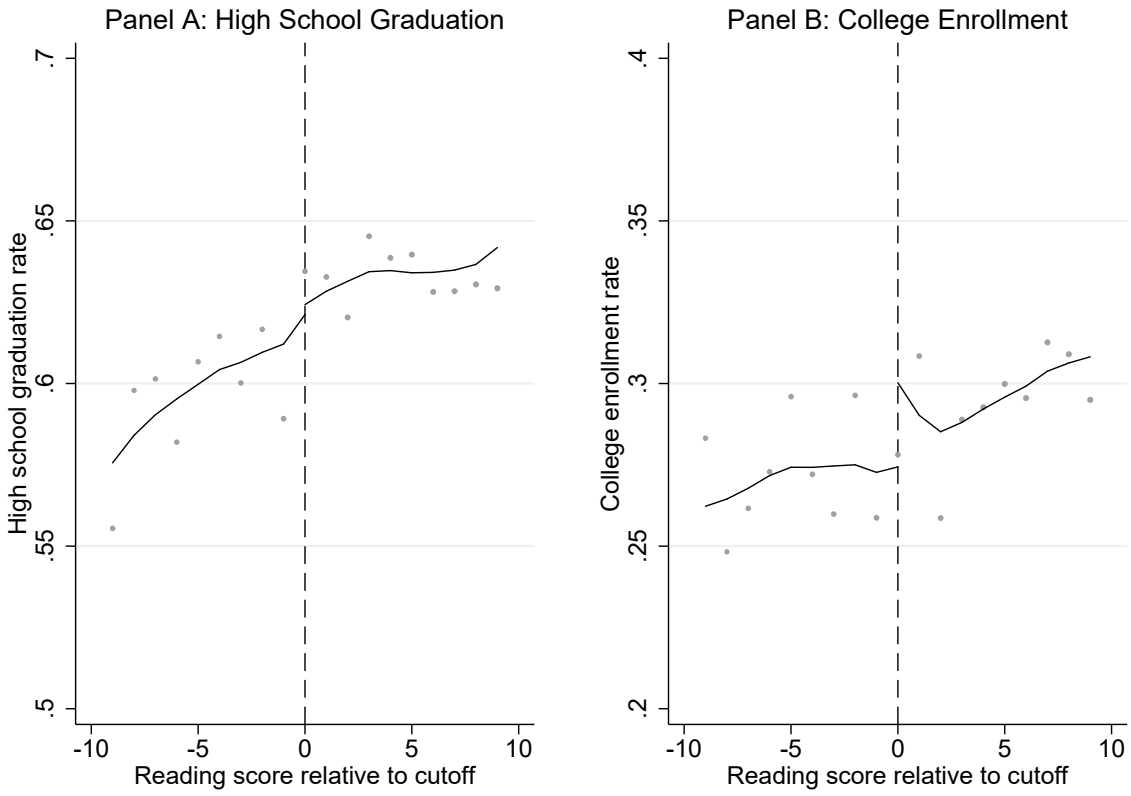
Note: Based on 2003-2008 cohorts. Discontinuity sample with 10-point bandwidth. Predicted retention probability displays predicted values after estimating a probit model that includes all student background variables except for reading scores as explanatory variables.

Figure A-4: The Relationship between Reading Scores in Grade 3 and Subsequent Attrition from the Data around the Cutoff



Note: Based on cohorts 2003-2008. Discontinuity sample with 10-point bandwidth.

Figure A-5: The Relationship between Grade 3 Reading Scores, High School Graduation and College Enrollment



Note: Panel A based on discontinuity sample with 10-point bandwidth for the 2003 and 2004 cohorts, panel B based on discontinuity sample with 10-point bandwidth for the 2003 cohort. Figure displays high school graduation rates and college enrollment rates by 3rd grade reading scores. Lines represent predicted values from local linear regressions on both sides of the cutoff.

Table A-1: Attrition Analysis: Reduced Form Estimates for 2003-2008 Cohorts

Outcome	Attrition from Flordia Public School Records in					
	1 year	2 years	3 years	4 years	5 years	6 years
Below cutoff	0.007	0.004	0.001	0.002	0.009	–0.013
	[0.004]	[0.004]	[0.005]	[0.006]	[0.007]	[0.009]
Reading	–0.001	–0.001	0.000	0.000	0.000	–0.002*
	[0.000]	[0.001]	[0.001]	[0.001]	[0.001]	[0.001]
Reading × Below cutoff	0.000	–0.001	0.000	–0.001	0.000	0.001
	[0.001]	[0.001]	[0.001]	[0.001]	[0.001]	[0.001]
Students	83,274	83,274	70,514	56,551	45,080	30,908
R^2	0.004	0.005	0.007	0.008	0.009	0.010

* p<0.10, ** p<0.05, *** p<0.01

Note: Based on discontinuity sample with 10-point bandwidth for the 2003-2008 cohorts. The table displays reduced form estimates for cohorts of students affected by the policy. Dependent variable is an indicator for missing test score information in a particular year. The top row indicates the distance in years between the year the outcome is measured and the first time students attended third grade. The table displays estimates with performance and demographic covariates of students. Performance and demographic covariates include math scores in grade 3, gender, age, race, special education status in grade 3, LEP status in grade 3, and free or reduced-price lunch status in grade 3. Robust standard errors in parentheses.

2. Comparison with OLS and Models without Controls

Table A-2: Effect of Grade Retention on Student Achievement
[Same Grade Comparison]

	Specification			
	OLS		IV	
Dependent Variable	(1)	(2)	(3)	(4)
Reading (SD= 370)				
grade 4 (n = 76,208)	115.04***	138.74***	253.48***	269.07***
	(2.85)	(2.69)	(10.22)	(12.98)
grade 5 (n = 59,562)	83.04***	106.45***	199.29***	204.48***
	(3.47)	(3.30)	(8.31)	(9.58)
grade 6 (n = 45,804)	56.73***	77.28***	159.18***	159.08***
	(3.41)	(2.97)	(16.68)	(13.93)
grade 7 (n = 35,051)	39.89***	53.46***	107.85***	102.43***
	(4.06)	(4.19)	(14.91)	(16.43)
grade 8 (n = 23,253)	21.71***	33.15***	73.49***	69.90***
	(3.55)	(3.45)	(11.72)	(9.67)
Math (SD= 306)				
grade 4 (n = 76,091)	92.05***	144.65***	182.14***	186.25***
	(2.87)	(1.60)	(8.26)	(8.56)
grade 5 (n = 59,334)	48.83***	98.68***	130.29***	133.21***
	(2.66)	(1.92)	(9.17)	(7.76)
grade 6 (n = 45,760)	38.53***	86.10***	118.18***	118.30***
	(3.34)	(2.27)	(14.35)	(12.04)
grade 7 (n = 35,057)	10.83***	45.27***	74.31***	65.88***
	(2.65)	(1.81)	(7.81)	(8.01)
grade 8 (n = 23,230)	6.47***	29.36***	49.49***	40.97***
	(2.16)	(1.60)	(10.69)	(8.22)
Performance and demographic covariates	No	Yes	No	Yes

Note: Based on discontinuity sample with 10-point bandwidth. Dependent variables are unadjusted developmental scale scores in reading and math. All estimations control for a linear function in grade 3 reading scores that allows for different trends on both sides of the cutoff and cohort dummies. Performance and demographic covariates include math scores in grade 3, gender, age, race, special education status in grade 3, LEP status in grade 3, and free or reduced-price lunch status in grade 3. Standard errors clustered by third grade school and by third grade reading score in parentheses.

Table A-3: Effect of Grade Retention on Student Achievement
[Same Age Comparison]

	Specification			
	OLS		IV	
Dependent Variable	(1)	(2)	(3)	(4)
Reading (SD= 370)				
1 year (n = 74,443)	–60.69***	–41.20***	81.67***	83.64***
	(3.54)	(3.94)	(8.67)	(8.67)
2 years (n = 59,554)	58.15***	76.40***	173.57***	174.14***
	(3.25)	(3.42)	(7.76)	(9.08)
3 years (n = 45,175)	–4.59	14.06***	90.53***	91.52***
	(3.69)	(3.56)	(13.45)	(10.73)
4 years (n = 35,001)	–53.22***	–35.89***	40.02***	40.96***
	(3.12)	(3.07)	(13.65)	(12.42)
5 years (n = 23,568)	–70.52***	–55.31***	–10.39	–8.19
	(4.71)	(4.83)	(15.10)	(15.09)
6 years (n = 12,912)	–30.21***	–14.74***	15.68	14.87
	(3.64)	(3.35)	(14.31)	(13.55)
Math (SD= 306)				
1 year (n = 74,327)	–1.46	47.84***	90.83***	92.51***
	(2.93)	(2.29)	(9.64)	(9.75)
2 years (n = 59,354)	–58.13***	–15.27***	24.55***	23.89***
	(3.08)	(2.14)	(9.30)	(7.07)
3 years (n = 45,093)	31.77***	73.81***	109.96***	110.34***
	(3.14)	(2.30)	(10.04)	(7.21)
4 years (n = 34,987)	–116.00***	–76.98***	–23.53**	–25.61**
	(4.05)	(2.97)	(11.92)	(11.75)
5 years (n = 23,563)	–77.69***	–48.61***	–25.33***	–25.53***
	(2.33)	(1.49)	(9.46)	(6.31)
6 years (n = 12,905)	–57.20***	–31.37***	–3.50	–8.59
	(3.45)	(2.44)	(6.94)	(5.90)
Performance and demographic covariates	No	Yes	No	Yes

Note: Based on discontinuity sample with 10-point bandwidth. Dependent variables are unadjusted developmental scale scores in reading and math; reported standard deviations are for grade 3. All estimations control for a linear function in grade 3 reading scores that allows for different trends on both sides of the cutoff and cohort dummies. Performance and demographic covariates include math scores in grade 3, gender, age, race, special education status in grade 3, LEP status in grade 3, and free or reduced-price lunch status in grade 3. Standard errors clustered by third grade school and by third grade reading score in parentheses.

Table A-4: Effect of Grade Retention on Student Achievement (rescaled)
[Same Age Comparison]

	Specification			
	OLS		IV	
Dependent Variable	(1)	(2)	(3)	(4)
Reading (SD= 370)				
1 year (n = 74,443)	70.58***	90.07***	212.94***	214.90***
	(3.54)	(3.94)	(8.67)	(8.67)
2 years (n = 59,554)	26.10***	44.74***	143.74***	144.26***
	(3.27)	(3.47)	(7.76)	(9.09)
4 years (n = 35,001)	–49.85***	–32.61***	42.87***	43.84***
	(3.11)	(3.05)	(13.66)	(12.46)
5 years (n = 23,568)	–57.41***	–42.52***	.63	2.77
	(4.72)	(4.83)	(15.27)	(15.26)
6 years (n = 12,912)	–73.64***	–56.91***	–21.07	–22.39*
	(3.75)	(3.48)	(14.08)	(13.44)
Math (SD= 306)				
1 year (n = 74,327)	36.00***	85.29***	128.29***	129.97***
	(2.93)	(2.29)	(9.64)	(9.75)
2 years (n = 59,354)	–17.51***	24.89***	62.64***	62.02***
	(3.09)	(2.15)	(9.27)	(7.04)
3 years (n = 45,093)	–25.59***	17.43***	58.54***	58.86***
	(3.14)	(2.31)	(10.04)	(7.03)
4 years (n = 34,987)	–83.38***	–45.12***	5.03	3.07
	(4.03)	(2.94)	(11.91)	(11.96)
5 years (n = 23,563)	–71.11***	–42.28***	–19.64**	–19.92***
	(2.30)	(1.45)	(9.45)	(6.34)
6 years (n = 12,905)	–87.89***	–61.14***	–29.53***	–34.99***
	(3.51)	(2.48)	(6.89)	(5.78)
Performance and demographic covariates	No	Yes	No	Yes

Note: Based on discontinuity sample with 10-point bandwidth. Dependent variables are rescaled developmental scale scores in reading and math; reported standard deviations are for grade 3. All estimations control for a linear function in grade 3 reading scores that allows for different trends on both sides of the cutoff and cohort dummies. Performance and demographic covariates include math scores in grade 3, gender, age, race, special education status in grade 3, LEP status in grade 3, and free or reduced-price lunch status in grade 3. Standard errors clustered by third grade school and by third grade reading score in parentheses.

3. Cohort-specific Results

Table A-5: Achievement Results by Cohort: Same Age Comparison

Cohort	2003	2004	2005	2006	2007	2008
Reading						
Based on vertically scaled FCAT scores (SD= 370)						
after 1 year	51.83***	85.10***	109.09***	86.84***	165.84***	19.84
	(8.05)	(20.52)	(24.66)	(16.08)	(49.79)	(26.80)
after 2 years	215.22***	158.93***	116.72***	158.17***	199.17***	237.95***
	(15.61)	(16.74)	(18.74)	(17.86)	(48.39)	(39.60)
after 3 years	76.72***	146.91***	71.81***	84.06***	132.95***	
	(11.47)	(44.17)	(26.35)	(18.58)	(48.79)	
after 4 years	32.63*	47.96**	50.05**	26.00		
	(19.31)	(22.93)	(22.57)	(26.13)		
after 5 years	–30.83*	33.26**	25.00*			
	(17.92)	(16.36)	(14.73)			
after 6 years	14.87	84.91***				
	(13.55)	(25.75)				
Based on FCAT 2.0 scores (SD= 21)						
after 7 years			3.33**	–.89	–2.44	
			(1.36)	(.89)	(2.53)	
Math						
Based on vertically scaled FCAT scores (SD= 306)						
after 1 year	63.99***	114.86***	83.23***	105.68***	123.01***	85.95***
	(10.27)	(18.52)	(13.49)	(18.28)	(29.50)	(24.58)
after 2 years	–4.46	20.51	15.04	43.73**	77.96***	106.01***
	(10.95)	(13.08)	(14.93)	(17.97)	(24.61)	(22.34)
after 3 years	104.49***	106.46***	109.24***	126.74***	107.50***	
	(9.17)	(16.25)	(22.62)	(20.45)	(33.53)	
after 4 years	–41.08**	–36.40***	4.87	–17.13		
	(20.06)	(11.88)	(14.29)	(12.62)		
after 5 years	–30.93***	–17.00**	–15.38			
	(7.95)	(8.65)	(15.98)			
after 6 years	–8.59	–7.19				
	(5.90)	(14.12)				
Students	15,687	12,037	12,434	9,981	12,995	11,536

* p<0.10, ** p<0.05, *** p<0.01

Note: Based on discontinuity sample with 10-point bandwidth. Dependent variables are unadjusted developmental scale scores in reading and math. The table displays IV estimates with performance and demographic covariates by cohort of students. Performance and demographic covariates include math scores in grade 3, gender, age, race, special education status in grade 3, LEP status in grade 3, and free or reduced-price lunch status in grade 3. A cohort is defined by the school year students attended third grade for the first time. The last row indicates the number of students by cohort in the first stage regression for outcomes after 1 year. Standard errors clustered by third grade school and by third grade reading score in parentheses.

Table A-6: Achievement Results by Cohort: Same Grade Comparison

Cohort	2003	2004	2005	2006	2007	2008
Reading						
Based on vertically scaled FCAT scores (SD= 370)						
in grade 4	268.54***	257.61***	252.04***	288.82***	410.85***	215.49***
	(13.78)	(22.49)	(30.62)	(25.74)	(63.34)	(46.66)
in grade 5	245.68***	253.80***	186.17***	216.01***	197.04***	
	(20.39)	(33.04)	(22.80)	(21.84)	(70.25)	
in grade 6	146.56***	226.56***	193.64***	193.31***		
	(15.97)	(43.38)	(25.57)	(61.15)		
in grade 7	120.11***	121.47***	119.21***			
	(31.06)	(21.48)	(33.90)			
in grade 8	61.15***	104.68***				
	(8.64)	(27.51)				
Based on FCAT 2.0 scores (SD= 21)						
in grade 9			8.92***	6.64***	6.52**	
			(1.59)	(1.20)	(2.75)	
in grade 10		4.23***	6.64***	3.09***		
		(1.49)	(1.47)	(1.12)		
Math						
Based on vertically scaled FCAT scores (SD= 306)						
in grade 4	176.29***	234.31***	170.15***	185.70***	227.66***	146.21***
	(14.69)	(27.12)	(17.46)	(20.17)	(40.32)	(24.01)
in grade 5	135.99***	121.07***	141.98***	176.51***	121.61***	
	(11.75)	(16.93)	(16.84)	(21.19)	(36.74)	
in grade 6	140.63***	126.10***	149.51***	102.06**		
	(14.33)	(17.81)	(32.92)	(47.62)		
in grade 7	89.13***	30.88*	108.48***			
	(19.33)	(17.42)	(28.04)			
in grade 8	42.42***	39.49**				
	(7.30)	(19.34)				
Students	16,093	12,448	12,744	10,263	13,193	11,693

Note: Based on discontinuity sample with 10-point bandwidth. Dependent variables are unadjusted developmental scale scores in reading and math. The table displays IV estimates with performance and demographic covariates by cohort of students. Performance and demographic covariates include math scores in grade 3, gender, age, race, special education status in grade 3, LEP status in grade 3, and free or reduced-price lunch status in grade 3. A cohort is defined by the school year students attended third grade for the first time. The last row indicates the number of students by cohort in the first stage regression for outcomes in grade 4. Standard errors clustered by third grade school and by third grade reading score in parentheses.

Table A-7: Grade Progression Results by Cohort

Cohort	2003	2004	2005	2006	2007	2008
after 2 years	–.884***	–.806***	–.881***	–.898***	–.875***	–.944***
	(.012)	(.025)	(.021)	(.024)	(.032)	(.024)
after 3 years	–.845***	–.735***	–.816***	–.871***	–.824***	–.879***
	(.021)	(.038)	(.027)	(.027)	(.038)	(.021)
after 4 years	–.784***	–.641***	–.771***	–.836***	–.769***	–.817***
	(.020)	(.048)	(.027)	(.030)	(.041)	(.023)
after 5 years	–.740***	–.569***	–.750***	–.836***	–.661***	–.758***
	(.021)	(.063)	(.027)	(.037)	(.048)	(.035)
Students	15,687	12,040	12,435	9,981	12,995	11,536

Note: Based on discontinuity sample with 10-point bandwidth. Dependent variable is the student's grade level. The table displays IV estimates with performance and demographic covariates by cohort of students. Performance and demographic covariates include math scores in grade 3, gender, age, race, special education status in grade 3, LEP status in grade 3, and free or reduced-price lunch status in grade 3. A cohort is defined by the school year students attended third grade for the first time. The last row indicates the number of students by cohort in the first stage regression for outcomes after 1 year. Standard errors clustered by third grade school and by third grade reading score in parentheses.

4. Sensitivity Analyses

In this section, we provide additional details on the sensitivity analyses presented in section 4.6 of the paper. Figure A-6 examines the robustness of our results to the use of alternatives to our preferred ten test-score-point bandwidth ranging from five to 25 points on either side of the cutoff. To consolidate presentation, we combine the data on each outcome across multiple years – achievement after 1-3 years and after 4-6 years as well as retention rates in years 2-5. The achievement results are based on the unadjusted DSS scores. In each panel the baseline estimate using the ten test-score-point bandwidth for the specification with outcomes combined across multiple years is reported explicitly. Achievement impacts in both subjects are consistently more positive using wider bandwidths, but the differences are modest in size. No consistent pattern with respect to bandwidth choice is evident in the results for future retention, high school graduation, and college enrollment.

Table A-8 examines the consequences of other changes to our specification and sample restrictions. It confirms that our results are not influenced by the exclusion of students at or within one test score point of the promotion cutoff, are essentially unchanged when we use school fixed effects to restrict comparisons to students attending the same school in third grade, and are robust to the use of quadratic terms in modeling the relationship between third grade reading scores and the probability of retention on either side of the cutoff.

As discussed in the text, a potential concern with interpreting our results as the causal effect of test-based retention is the possibility of labeling effects (Papay et al., 2016). To test whether labeling effects bias our estimates of test-based retention, we conduct a placebo test using the two cohorts of students in our data that entered third grade before 2003 and therefore were unaffected by the promotion policy. The results in Table A-9 confirm that being labeled a level 1 reader had no effect on future achievement for these students.

Tables A-10 and A-11 report estimation results for subgroups defined based on their own characteristics or those of the schools they attended in third grade. Lacking strong theoretical expectations about which students would benefit most from retention, we primarily regard these subgroup results as an additional robustness check to confirm that our primary results are not driven by large positive effects for specific groups of students. However, they also serve as exploratory analyses of potential effect heterogeneity intended to guide future research.

Table A-10, which presents results for several key subgroups over the same time periods displayed in Figure A-6, provides little evidence of qualitative differences in the effect of

grade retention across subgroups based on gender, ethnicity, or free/reduced-price lunch eligibility. The short-term and longer-term achievement effects of retention appear to be modestly less positive for black students than for whites or Hispanics. The achievement gains from retention also appear to be larger and more persistent for students who were absent from school more often in grade 3.[1] This suggests that test-based retention may be particularly beneficial for low-achieving students whose initial third grade year was disrupted by repeated absences.

The remaining rows in Table A-10 examine whether our estimates of retention effects are local to students at a specific achievement level, exploiting the fact that there is considerable variation in the math achievement of Florida students who are retained on the basis of their reading test scores. Among students in our preferred bandwidth, 20,537 (27 percent) were classified as performing at level one (of five) based on the third grade math test, 26,357 (35 percent) performed at level two, and 29,253 (29 percent) performed at level three or higher. The first-stage results in column (1) show that the increase in the probability of retention at the promotion cutoff was more than twice as large for students performing at level one in math as for students performing at level three or above, suggesting that students' math performance influenced whether they were granted an exemption from the retention requirement. The estimated effects on reading and math achievement are quite similar across all three groups, however, providing at least suggestive evidence that the short-term benefits of test-based retention are not limited to students achieving at a specific level.

The results also confirm that the increase in retention probabilities for students just missing the cutoff was smaller for special education students. This is as expected given that students in this group were eligible for additional good cause exemptions from the retention requirement.

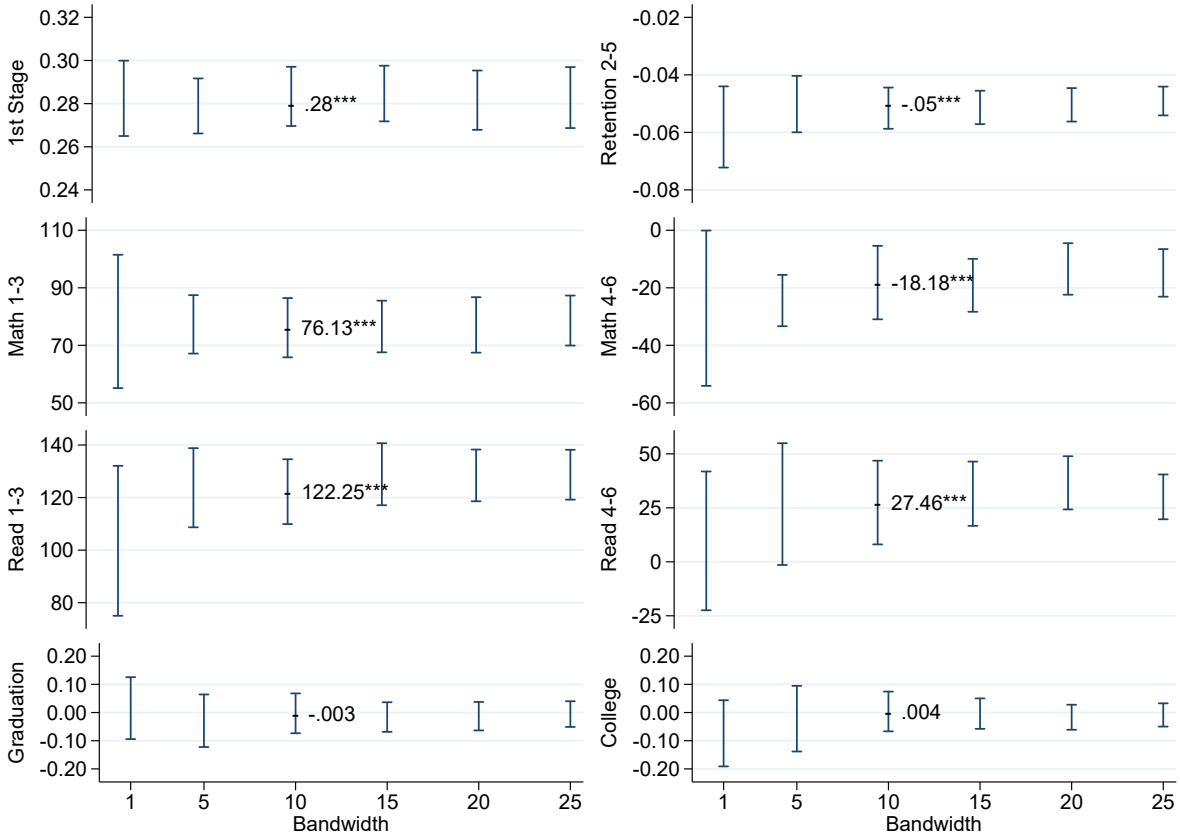
Similar to the subgroup analysis by student characteristics, Table A-11 examines whether the effects of retention vary according to the characteristics of the school attended in third grade. For simplicity, we split the discontinuity sample into two subgroups at the median of each available school characteristic. The results provides little evidence of systematic heterogeneity in the effects of test-based retention by pupil/teacher ratio, expenditure per student, average teacher experience, and average teacher salary. There is some evidence, however, that the positive effects of test-based retention are more pronounced in schools with below-median retention rates. This could indicate that retained students receive more attention when there are fewer of them, which may reinforce any

[1] Among students in our preferred bandwidth, 28 percent were absent 10 days or more and 47 percent were absent fewer than 5 days during their initial third grade year.

beneficial impact of test-based retention.

Although we find little evidence of qualitative differences in retention effects across subgroups, policymakers may nonetheless be interested in which students subgroups are most impacted by the introduction of a test-based promotion policy. Because some retained students (i.e., always-takers) would have been retained regardless of whether they scored below the promotion cutoff and other students (i.e., never-takers) would never be retained, students complying with the policy cannot be individually identified. We can, however, use the first-stage estimates of the effect of scoring below the promotion cutoff on retention probabilities for students with various characteristics to describe the distribution of these characteristics among compliers (c.f. Angrist and Pischke, 2009). Table A-12, which provides results of a standard complier analysis based on the estimates reported in Column (1) of Table A-10, suggests that compliers are on average not very different in their observable student characteristics from the average student in the discontinuity sample, with one notable exception. For students with different math achievement levels we see substantial differences in compliance rates. For example, a complier is 36 percent more likely to score at level one in math than the average student in the discontinuity sample.

Figure A-6: Estimated Effects for Different Bandwidths



Note: Based on discontinuity samples with varying bandwidths all cohorts in panels 1-6, for the 2003 and 2004 cohorts in panel 7, and for the 2003 cohort in panel 8. Figure displays 95% confidence bands for robustness checks with different bandwidths reported in Table A-8. The estimate displayed in each panel indicates the estimate from our preferred specification.

Table A-8: Robustness Checks

Outcomes:	1st Stage	Reading		Math		Retention	HS Grad.	College
Years:		1-3	4-6	1-3	4-6	2-5		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Baseline	.28***	122.25***	27.46***	76.13***	–18.18***	–.05***	–.00	.00
	(.01)	(6.29)	(9.89)	(5.26)	(6.52)	(.00)	(.04)	(.04)
Bandwidth								
25	.28***	128.70***	30.07***	78.61***	–14.81***	–.05***	–.01	–.01
	(.01)	(4.84)	(5.30)	(4.44)	(4.22)	(.00)	(.02)	(.02)
20	.28***	128.44***	36.59***	77.10***	–13.44***	–.05***	–.01	–.02
	(.01)	(5.03)	(6.29)	(4.92)	(4.57)	(.00)	(.03)	(.02)
15	.28***	128.88***	31.54***	76.55***	–19.13***	–.05***	–.02	–.00
	(.01)	(6.02)	(7.60)	(4.58)	(4.70)	(.00)	(.03)	(.03)
5	.28***	123.75***	26.73*	77.29***	–24.42***	–.05***	–.03	–.02
	(.01)	(7.68)	(14.39)	(5.18)	(4.55)	(.01)	(.05)	(.06)
1	.28***	103.55***	9.69	78.30***	–27.09**	–.06***	.02	–.07
	(.01)	(14.56)	(16.40)	(11.83)	(13.78)	(.01)	(.06)	(.06)
w/o cutoff	.28***	120.68***	31.36**	76.33***	–19.01**	–.05***	.00	.03
	(.01)	(7.55)	(12.69)	(6.77)	(8.04)	(.00)	(.04)	(.03)
School fe	.29***	117.20***	13.23	73.24***	–29.04***	–.05***	–.01	–.00
	(.01)	(6.86)	(8.62)	(5.57)	(7.17)	(.00)	(.03)	(.04)
Quadratic	.29***	121.33***	27.03***	75.45***	–18.78***	–.05***	–.01	.01
	(.01)	(6.06)	(9.98)	(5.31)	(6.10)	(.00)	(.04)	(.04)

Note: Based on discontinuity sample for all cohorts in columns (1)-(6), for the 2003 and 2004 cohorts in column (7), and for the 2003 cohort only in column (8). Top row indicates dependent variable. Second row indicates years after potential grade 3 retention. Column (1) shows first stage estimates, while columns (2)-(8) report the corresponding IV estimates. All estimations control for special education status in grade 3, LEP status in grade 3, a linear function in grade 3 reading scores that allows for different trends at both sides of the cutoff, cohort dummies, grade 3 math scores, gender, age, race, and free or reduced-price lunch status in grade 3. Estimated effects on achievement are based on unadjusted developmental scales scores. Standard errors clustered by third grade school and by third grade reading score in parentheses.

Table A-9: Placebo Test: Reduced Form Estimates for 2001 and 2002 Cohorts

Panel A	Outcome: Reading Scores in					
	1 year	2 years	3 years	4 years	5 years	6 years
Below cutoff	–3.711	1.137	4.146	–2.623	0.799	–8.490*
	(4.248)	(4.395)	(4.605)	(4.600)	(4.039)	(4.406)
Reading	4.024***	3.677***	4.102***	3.110***	2.707***	1.864***
	(0.498)	(0.519)	(0.554)	(0.539)	(0.481)	(0.513)
Reading	–0.647	–0.278	–0.962	–0.648	–0.572	–0.397
× Below cutoff	(0.712)	(0.739)	(0.773)	(0.770)	(0.686)	(0.732)
Additional covariates	Yes	Yes	Yes	Yes	Yes	Yes
Students	34,028	31,800	30,237	29,713	28,937	26,804
R^2	0.109	0.088	0.108	0.105	0.120	0.109

Panel B	Outcome: Math Scores in					
	1 year	2 years	3 years	4 years	5 years	6 years
Below cutoff	0.049	–1.654	–3.357	–4.571	–4.419	–4.362
	(3.519)	(3.656)	(3.811)	(3.912)	(3.405)	(3.096)
Reading	1.322***	1.371***	0.746*	0.904**	0.621	0.934***
	(0.416)	(0.430)	(0.453)	(0.459)	(0.397)	(0.355)
Reading	–0.678	–0.901	–0.494	–0.765	–0.546	–1.591***
× Below cutoff	(0.589)	(0.610)	(0.637)	(0.652)	(0.571)	(0.510)
Additional covariates	Yes	Yes	Yes	Yes	Yes	Yes
Students	34,022	31,830	30,220	29,699	28,816	26,801
R^2	0.365	0.346	0.303	0.279	0.286	0.277

* p<0.10, ** p<0.05, *** p<0.01

Note: Based on discontinuity sample with 10-point bandwidth for the 2001 and 2002 cohorts. The table displays reduced form estimates for cohorts of students not affected by the policy. Dependent variables are unadjusted developmental scale scores in reading in panel A and math in panel B. The top row indicates the distance in years between the year the outcome is measured and the first time students attended third grade. Additional covariates include math scores, gender, age, race, free or reduced-price lunch status in grade 3. Robust standard errors in parentheses.

Table A-10: Subgroup Results by Student Characteristics

Outcomes:	1st St.	Reading		Math		Retention	HS Grad.	College
Years:		1-3	4-6	1-3	4-6	2-5		
Subgroup	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Baseline	.28***	122.25***	27.46***	76.13***	–18.18***	–.07***	–.00	.00
	(.01)	(6.29)	(9.89)	(5.26)	(6.52)	(.01)	(.04)	(.04)
Girls	.28***	126.45***	30.82**	74.36***	–15.86*	–.05***	–.05	–.02
	(.01)	(10.48)	(15.36)	(9.04)	(8.76)	(.00)	(.06)	(.06)
Boys	.29***	117.92***	23.75**	76.80***	–21.90**	–.06***	.05	.02
	(.01)	(8.16)	(10.43)	(6.77)	(8.77)	(.01)	(.05)	(.03)
White	.25***	142.12***	50.94***	91.34***	–4.43	–.06***	.07	–.04
	(.01)	(11.63)	(16.75)	(12.78)	(12.86)	(.01)	(.05)	(.06)
Black	.31***	107.11***	2.45	73.41***	–31.10***	–.06***	–.06	.06
	(.01)	(8.76)	(11.86)	(5.29)	(9.08)	(.01)	(.06)	(.04)
Hispanic	.29***	121.95***	33.31**	62.38***	–12.02	–.03***	–.02	–.07
	(.01)	(10.34)	(14.17)	(12.72)	(9.98)	(.00)	(.06)	(.07)
Math	.38***	94.41***	12.01	68.60***	–38.98***	–.06***	–.01	.02
Level 1	(.01)	(12.12)	(11.86)	(5.33)	(9.77)	(.00)	(.06)	(.06)
Math	.31***	132.10***	41.93***	95.81***	–5.08	–.05***	.04	–.05
Level 2	(.01)	(11.44)	(15.36)	(8.78)	(10.32)	(.00)	(.05)	(.05)
Math	.19***	149.74***	13.00	54.14***	–16.73**	–.04***	–.08	.06
Level 3+	(.01)	(12.22)	(17.25)	(13.32)	(8.26)	(.01)	(.06)	(.08)
Age 9	.27***	125.49***	37.92***	79.78***	–7.04	–.05***	.01	.06
or above	(.01)	(8.13)	(12.61)	(6.89)	(8.59)	(.00)	(.04)	(.04)
Age 8	.33***	113.82***	8.28	69.10***	–39.42***	–.06***	–.03	–.13**
or below	(.01)	(10.28)	(14.60)	(9.14)	(8.72)	(.01)	(.06)	(.05)
Free or	.31***	113.92***	23.51*	75.42***	–12.28	–.05***	–.01	–.01
red. lunch	(.01)	(6.39)	(12.15)	(7.73)	(9.09)	(.00)	(.05)	(.03)
LEP	.29***	134.46***	35.24	74.21***	–21.98	–.04***	–.01	.01
Students	(.01)	(11.85)	(21.46)	(10.37)	(15.52)	(.00)	(.05)	(.06)
Special Ed	.22***	127.87***	67.79***	70.51***	–16.65	–.05***	–.01	.01
Students	(.01)	(15.17)	(25.54)	(15.37)	(15.86)	(.01)	(.06)	(.08)
Days absent	.29***	140.16***	69.75***	97.74***	8.57	–.06***	.05	.03
> 10	(.01)	(15.13)	(24.21)	(16.12)	(13.85)	(.01)	(.04)	(.07)
Days absent	.29***	129.65***	37.54***	79.38***	–8.64	–.04***	.03	.04
$5 - 10$	(.01)	(13.80)	(13.01)	(14.63)	(12.02)	(.01)	(.06)	(.05)
Days absent	.28***	108.87***	–.96	64.97***	–35.72***	–.06***	–.05	–.03
< 5	(.01)	(16.25)	(15.74)	(5.91)	(9.78)	(.00)	(.05)	(.06)

Note: Based on discontinuity sample with 10-point bandwidth for all cohorts in columns (1)-(6), for the 2003 and 2004 cohorts in column (7), and for the 2003 cohort only in column (8). Top row indicates dependent variable. Second row indicates years after potential grade 3 retention. Column (1) shows first stage estimates. Columns (2)-(8) report IV estimates with performance and demographic covariates. Estimated effects on achievement are based on unadjusted developmental scales scores. Standard errors clustered by third grade school and by third grade reading score in parentheses.

Table A-11: Subgroup Results by Third Grade School Characteristics

Outcomes:	1st St.	Reading		Math		Retention	HS Grad.	College
Years:		1-3	4-6	1-3	4-6	2-5		
Subgroup	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Baseline	.28***	122.25***	27.46***	76.13***	–18.18***	–.05***	–.00	.00
	(.01)	(6.29)	(9.89)	(5.26)	(6.52)	(.00)	(.04)	(.04)
Pupil teacher ratio								
≥ median	.31***	120.76***	30.28***	69.57***	–20.20**	–.06***	.00	.02
	(.01)	(6.16)	(10.98)	(7.56)	(7.99)	(.01)	(.03)	(.06)
< median	.26***	123.68***	22.12	83.73***	–14.97**	–.05***	–.01	–.01
	(.01)	(11.32)	(14.29)	(3.05)	(7.35)	(.00)	(.05)	(.04)
Expenditure per student								
≥ median	.30***	104.62***	23.50**	69.24***	–22.83***	–.05***	–.02	–.00
	(.01)	(6.18)	(11.71)	(8.09)	(5.60)	(.01)	(.03)	(.04)
< median	.27***	137.51***	33.84**	78.59***	–10.27	–.06***	.02	.03
	(.01)	(9.67)	(15.30)	(9.34)	(12.31)	(.01)	(.05)	(.07)
Teacher experience								
≥ median	.27***	116.61***	21.27	68.49***	–27.02***	–.06***	–.02	–.03
	(.01)	(6.88)	(13.08)	(8.82)	(9.94)	(.01)	(.06)	(.05)
< median	.30***	126.28***	35.08***	83.92***	–9.92	–.05***	.01	.05*
	(.01)	(8.63)	(11.78)	(7.14)	(7.46)	(.00)	(.04)	(.03)
Teacher salary								
≥ median	.31***	112.66***	15.55	69.53***	–33.94***	–.04***	–.04	–.01
	(.01)	(8.52)	(15.57)	(8.42)	(8.17)	(.01)	(.03)	(.06)
< median	.26***	130.97***	45.02***	83.80***	5.80	–.06***	.05	.04
	(.01)	(8.82)	(11.64)	(6.22)	(9.21)	(.01)	(.07)	(.05)
Retention rate								
≥ median	.36***	102.71***	16.33**	70.20***	–20.65***	–.05***	.01	–.01
	(.01)	(8.78)	(8.03)	(6.44)	(7.18)	(.00)	(.04)	(.04)
< median	.21***	149.90***	49.88**	82.68***	–15.56	–.06***	–.03	.03
	(.01)	(7.92)	(20.57)	(8.77)	(17.41)	(.01)	(.07)	(.08)
Failure rate								
≥ median	.31***	112.57***	13.47	76.80***	–24.05***	–.05***	.01	–.02
	(.01)	(11.98)	(11.90)	(5.89)	(7.54)	(.01)	(.05)	(.04)
< median	.26***	130.08***	43.95*	71.98***	–13.30	–.06***	–.03	.08
	(.01)	(11.42)	(22.84)	(7.67)	(9.65)	(.00)	(.07)	(.10)

Note: Based on discontinuity sample with 10-point bandwidth for all cohorts in columns (1)-(6), for the 2003 and 2004 cohorts in column (7), and for the 2003 cohort only in column (8). Top row indicates dependent variable. Second row indicates years after potential grade 3 retention. Column (1) shows first stage estimates. Columns (2)-(8) report IV estimates with performance and demographic covariates. Estimated effects on achievement are based on unadjusted developmental scales scores. Standard errors clustered by third grade school and by third grade reading score in parentheses.

Table A-12: Characterizing Compliers

Variable	All students	Compliers	Relative likelihood that compliers have the characteristic indicated in each row
	(1)	(2)	(3)
Girl	0.47	0.46	0.98
Boy	0.53	0.54	1.02
White	0.32	0.28	0.88
Black	0.35	0.39	1.11
Hispanic	0.28	0.29	1.01
Age 9 or above	0.75	0.71	0.95
Age 8 or below	0.25	0.29	1.15
Free/reduced lunch	0.71	0.77	1.08
Days absent > 10	0.27	0.28	1.01
Days absent 5-10	0.25	0.26	1.01
Days absent < 5	0.47	0.47	0.98
Math Level 1	0.28	0.38	1.36
Math Level 2	0.35	0.38	1.08
Math Level ≥ 3	0.37	0.25	0.67

Note: The table reports an analysis of complier characteristics. Column (1) reports the shares of students with the characteristic indicated in each row among all students in the discontinuity sample. Column (2) reports the shares of students with the characteristic indicated in each row among compliers. Column (3) reports the ratio of the first stage estimate for individuals with that characteristic to the first stage for the discontinuity sample as a whole, a statistic which can be interpreted as the relative likelihood that compliers have this characteristic. Based on discontinuity sample with 10-point bandwidth for the 2003-2008 cohorts.

References

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

Papay, J. P., Murnane, R. J., and Willett, J. B. (2016). The impact of test score labels on human-capital investment decisions. *Journal of Human Resources*, 51(2):357–388.