# School Accountability, Test Scores, and Long-Run Outcomes

Joshua Hollinger[†]

November 1, 2021

Click here for most recent version.

**ABSTRACT**

While many education policies target test scores as a contemporaneous measure of student learning, a common concern is that these policies may generate higher test scores in a way that fails to translate to more important student outcomes in the long run. I use administrative data from North Carolina and two regression discontinuity designs to estimate the impact of school accountability pressure under No Child Left Behind on elementary students' test scores and their long-run outcomes at the end of high school. I find modest positive effects on elementary test scores and a significant increase in SAT scores years later. There is some evidence for a small increase in high school GPA, mixed evidence for an increase in students intending to attend a 4-year instead of a 2-year college, and no effect on high school graduation or intention to attend any college. Further evidence suggests the effect on SAT scores may be explained by persistent test-score effects in years after accountability exposure. Altogether, these results lend support to a mixed story for No Child Left Behind: while accountability pressure led to a long-run increase in skills captured by tests, these learning gains were not strong or broad enough to yield meaningful improvements in other long-run outcomes like educational attainment.

[†]University of Rochester, Department of Economics. jhollin3@ur.rochester.edu. I thank my advisors, Lisa Kahn, Ronni Pavan, and John Singleton for all of their support and suggestions. I also thank John Singleton for supporting my access to the data used in this paper. I thank Richard DiSalvo for encouraging and helpful conversations on this project, as well as Kegon Tan, Travis Baseler, Narayana Kocherlakota, Mark Bils, Michael Wolkoff, Joanna Venator, and Klint Mane for their valuable input.

# 1 Introduction

A long-standing fundamental question for policymakers is how to ensure quality provision of elementary and secondary public education. One approach is to focus on providing quality inputs for public schools[1]; another approach is measuring the output of public schools, generally with standardized tests, and attaching incentives to these measures. Though the task of public education in the United States has largely been relegated to state and local governments, the federal government has taken on a considerable role in overseeing public education, beginning most notably with the Elementary and Secondary Education Act of 1965. This law sought to address the plight of students in poorer school districts, and among other things, it created Title I funding aimed at helping these disadvantaged students. However, at the beginning of the 21st century, a bipartisan consensus had emerged on the need for the federal government to go beyond providing funding and start holding schools accountable for their students' learning. The No Child Left Behind Act of 2001 (NCLB) was the result, creating a federal system for school accountability based primarily on student test scores. NCLB created a set of standards for measuring student outcomes, creating targets for student outcomes, and meting out punishments for schools failing to meet these targets[2], which states had to comply with.

While a number of studies have estimated the effects of accountability pressure created by NCLB on student test scores[3], generally finding modest positive effects, the effects on students' longer-run outcomes are unknown. Since test scores are the measure targeted by the incentives, a major concern is that the test-score increases generated by the policy may not translate to more important longer-run outcomes, perhaps even making students worse off by shifting effort toward the narrow goal of increasing test scores. If educators respond to test-score-based accountability by shifting toward activities that solely increase skills measured by tests, rather than the broader set of skills important for long-run outcomes, this distortion of "teaching to the test" may render school accountability ineffective in the long-run. Alternatively, even in the absence of this type of distortion, it may be the case that test-score increases were too small to bring about meaningful improvements in long-run outcomes.

This paper evaluates the effect of No Child Left Behind accountability pressure on students'

[1] Some examples include providing sufficient funding for infrastructure and paying high-quality faculty and staff, creating standards that restrict who is allowed to teach, ensuring smaller class sizes, or reducing economic and racial segregation across schools.

[2] These punishments, or sanctions, are described in Section 2, and detailed in Appendix Table A1.

[3] Chakrabarti 2014, Ahn and Vigdor 2014, Dee and Jacob 2011

longer-run outcomes. Several barriers have hindered the evaluation of this question thus far in the literature. First, data linking variation in NCLB accountability pressure, student test score outcomes, and students' longer-run outcomes is scarce. I use administrative data from North Carolina containing elementary students' test scores and a number of important outcomes captured at the end of high school. Second, only recently has enough time passed for end-of-high-school outcomes to be observed for affected elementary students. I provide an early look at NCLB's effects on students' long-run outcomes by capturing important outcomes measured before adulthood but at the end of a student's time in the K-12 system. The North Carolina dataset contains information on whether a student graduates high school, their high school GPA, whether or not they take the SAT and what score they receive, and whether or not the student plans to attend a 4-year or 2-year college. These provide important proxies for even longer-run outcomes, such as college attainment and income upon entering the labor market.

I use two regression discontinuity designs (RDD) that capture different types of variation in accountability pressure created by NCLB. The first, which I call the subgroup RD, is based on a rule that excluded some groups of students from the targets set in the accountability regime. NCLB created targets for each school based on test score proficiency of the entire set of students in the school, as well as targets for nine subgroups of students within the school. Subgroups were based on race[4], as well as economic disadvantage[5], limited English proficiency, and disability status. However, schools were only held accountable for a subgroup if there were 40 or more students in the school in that subgroup for a given year. Thus, a student attending a school with 39 students in their subgroup took a test with no implications for the school's status under NCLB, while a student of the same subgroup attending a school with 40 students in their subgroup counted toward the incentivized NCLB objectives. To capture this variation, I use an RDD around this 40-student cutoff, similar to Gilraine (2018) and Farber (2016).

The second RD, which I call the school RD, strategy is based on the structure of NCLB sanctions for failing schools, where schools barely failing to meet their targets faced much more pressure than schools barely passing. Schools that failed for two consecutive years were deemed "in need of improvement" and were subjected to costly penalties that accumulated with each subsequent year of failure. Thus, a school that failed for the first time had strong incentives to improve and avoid the sanctions accompanying a second failure. A school that had not yet failed was under much less

[4]The racial subgroups were defined as white, black, Hispanic, Asian, Native American, and multi-racial.
[5]Economic disadvantage is defined by students' eligibility for free or reduced-price school lunch.

pressure, still being two consecutive years of failure away from facing sanctions. To capture this variation, I use an RDD to compare the outcomes of students in a school that barely passed in the prior year to those of students in a school that barely failed in the prior year. A similar strategy has been used in a number of studies, including Ahn and Vigdor (2014) and Chakrabarti (2014).

Similar to previous literature, I find modest but significant effects of NCLB accountability pressure on students' math and reading test scores in elementary school using both identification strategies. Being subjected to accountability results for a year of elementary school results in roughly 0.07 standard deviation (SD) higher test scores. Evaluating students' outcomes at the end of high school, I find economically and statistically insignificant effects on graduation and intention to attend any college. I find mixed evidence on high school GPA and a shift from 2-year to 4-year colleges[6]. However, I find a considerable positive effect on SAT scores. Given this positive effect on a different high-stakes standardized exam years after accountability pressure in elementary school, I investigate effects on student achievement over time. Evaluating effects on test scores in subsequent years, I find persistent increases even three years after exposure to accountability pressure. I also find positive effects on GPA in math and reading classes in 9th grade for the subgroup RDD, which continue but decrease through 10th, 11th, and 12th grade GPA. This lends support to the hypothesis that accountability pressure led to a persisting increase in math and reading skills, particularly the type measured by tests.

However, given the lack of an effect of accountability on other important long-run outcomes, such as high school graduation or intention to attend college, I consider one possible reason that accountability might increase test scores in the short run but fail to improve these long-run outcomes: the test score increases may be too small to create meaningful increases in long-run outcomes, given the limited relationship between test score gains and long-run outcomes. To do this, I do some "back-of-the-envelope" calculations to compare "commensurate" long-run effects and the actual long-run effects from my results. The commensurate long-run effect is the expected long-run effect given the test score effect and the relationship between test score gains and long-run outcomes. The commensurate long-run effects I calculate are generally somewhat small, and mostly within the confidence interval of the actual long-run effects from my estimates. One exception is the high school graduation effect, which is significantly lower than the commensurate effect. Thus, a larger increase in test scores may have been required for meaningful improvements across all of

[6] I find a small effect on high school GPA using the subgroup RDD, but no effect using the school RDD. I find evidence for increased intention to attend 4-year colleges and decreased intention to attend 2-year colleges with the school RDD, but no effect using the subgroup RDD.

these long-run outcomes. Also, the lack of any negative effects suggests that potential distortions from teaching to the test were not excessively harmful.

This paper is the first to estimate the long-run effects of accountability pressure under No Child Left Behind, the federal policy that is arguably the most important accountability legislation in the history of the United States. While a number of studies have explored various forms of distortion or gaming responses to the high-stakes placed on test scores[7], the question remains as to how these types of distortions translate to the long-run effectiveness of NCLB accountability. Relevant for economics more broadly, my paper provides an opportunity to empirically evaluate the multitasking model of Holmstrom and Milgrom (1991) in the context of test-based school accountability. I develop and discuss a version of this model in the Appendix. Deming et al. (2016) is the only paper in the literature on the long-run effects of accountability. They estimate the effects of an accountability scheme used in Texas prior to NCLB and find positive effects on test scores, as well as positive effects on long-run outcomes such as income and college attendance. Using a different identification strategy[8] and a different setting than Deming et al., I find a significant effect of accountability pressure in elementary school on SAT scores in high school, but little evidence of effects on other important long-run outcomes like high school graduation or intention to attend college.

The rest of the paper is structured as follows: Section 2 discusses the policy setting under No Child Left Behind. Section 3 describes the data. Section 4 presents my empirical strategy, using the subgroup RDD and the school RDD. Section 5 contains my main results on the effects of NCLB accountability pressure on test scores and long-run outcomes, as well as effects on test scores in subsequent years. Section 6 presents a discussion and analysis of several issues relevant for interpretation of the main results, including what size of long-run effects one should expect, heterogeneous effects across students and schools that may be more affected by accountability, the potential issue of attrition bias in the long-run estimates, and potential mechanisms. Section 8 concludes.

[7]Figlio and Wynicki (2005) show that schools increased caloric content of school lunches on test days to boost student performance. Jacob (2005) provides evidence that schools shifted effort from non-tested to tested subjects, and used special education classifications strategically to meet NCLB standards more easily. Koretz (2002), Figlio (2006), and Jacob and Levitt (2003) provide further evidence on the distortionary responses of schools to state-level accountability policies.

[8]Deming et al. estimate each school's probability of meeting their accountability target and identify effects from schools that go from zero to non-zero probability of failure due to targets increasing over time. This strategy is less desirable in my setting, partially due to the fact that targets under NCLB in North Carolina did not vary as much, and partially due to the difficulty of ascertaining how small changes in failure probability relate to the magnitude of accountability pressure.

## 2 No Child Left Behind

The No Child Left Behind Act of 2001 (NCLB) implemented school accountability at the federal level, requiring public schools in all states to track student performance via standardized tests. The accountability policies took effect for the 2002-2003 school year and remained in effect until its replacement by the Every Student Succeeds Act in 2015. A "proficiency score" determined the test score expected of all students, and schools were required to have a certain percentage of students meet this minimum score in order to reach "Adequate Yearly Progress" (AYP). Beginning in 2009, students who failed to reach proficiency were allowed to retake the test another time[9]. Schools also had to meet objectives for student attendance and percentages of students taking the tests. These objectives for percent of students proficient on tests, student attendance, and percent of students taking the tests were set for the entire school, as well as for nine subgroups based on student race and disadvantaged statuses: black, Hispanic, white, Asian, multi-racial, Native American, economically disadvantaged[10], limited English proficiency, and students with disabilities. If a school failed to meet any of these objectives, overall or for any subgroup, they failed to meet AYP. Failing AYP resulted in escalating sanctions for Title 1 schools who kept failing an objective in subsequent years. These sanctions, described in Appendix Table A1, began with requiring schools to allow students to choose another school and ended in restructuring of the school[11].

In order to prevent random fluctuations in test scores of a small number of students from causing schools to fail AYP, the law provided an exception such that if a school had fewer than 40 students in a subgroup they were not held accountable for that subgroup[12]. Thus, schools faced potentially strong incentives to ensure that a subgroup with 40 or more students met the proficiency target, but they had no direct incentive to worry about subgroups with fewer than 40 students. This creates a natural setting for an RDD, a strategy which has been used by Gilraine (2018) and Farber (2016), wherein a student in a subgroup of 40 students is treated and a student in a subgroup of 39 students is not.

However, several issues complicate the variation in incentives around the 40-student subgroup

[9] All students with achievement level 2, the level below proficient, were retested. Parents of students with achievement level 1, the lowest level, were notified that they could request a retake. See the Consolidated State Application Accountability Workbook from the State Board of Education of North Carolina for more information: https://www2.ed.gov/admins/lead/account/stateplans03/nccsa.pdf

[10] These are students eligible for free or reduced-price lunch at school.

[11] Appendix Table A1 is taken from Ahn and Vigdor (2014) and describes each of the sanctions. See Ahn and Vigdor (2014) for more details.

[12] The specific threshold varied by state, but 40 students was the threshold for North Carolina

threshold. First, in order to be counted for the school's AYP, a student had to attend the school for 75% of the school year. Given a school year of 180 days, this would be 135 days. This results in some ambiguity around the exact count of students to be included toward the subgroup threshold, particularly for the first eight or nine weeks of school, since students entering the school with more than 135 days left would add to the count, but those entering later would not. This ambiguity could potentially dampen the increase in incentives associated with exceeding the threshold.

Furthermore, due to the design of the accountability regime, the variation in incentives around the subgroup threshold is far from the only variation that should be expected to differentially motivate schools and teachers to increase students' test scores. Since the test score require for proficiency did not vary with students' prior achievement, schools had the strongest incentives to improve the marginal students with expected test scores near the proficiency cutoff. Also, schools failed AYP if they failed any one of their objectives, meaning that schools failing multiple objectives by large margins and schools passing all objectives with large margins both were unlikely to change their AYP status with marginal improvements[13]. Given these factors, incentive strength likely varied considerably across schools, specific subgroups within schools, and specific students within schools. While schools faced no direct incentive to increase the test scores of students in subgroups with fewer than 40 students, these students may be affected by school-level or classroom-level changes in inputs directed at other students. Thus, I intend to capture a relative effect of differential incentive strength, rather than the absolute size of the effect of accountability pressure. Conversely, incentive strength varied immensely across the subgroups with more than 40 students. To leverage this, I consider heterogeneous treatment effects corresponding to the types of variation in incentives discussed here.

# 3 Data

I use administrative data from the North Carolina Education Research Data Center (NCERDC), which includes extensive information on all public school students in North Carolina. To evaluate

[13] While there were fixed targets for the percentage of students within schools and subgroups of schools attaining a test score at or above proficiency, the rules allowed for two important adjustments that altered incentives: the confidence interval exception, and the "safe harbor" exception. The confidence interval rule allowed schools to pass the proficiency target for a subgroup if the percent proficient was within a confidence interval of the proficiency target, based on the number of students in the subgroup. This provided an effectively lower target for school subgroups with fewer students. The safe harbor rule allowed schools to pass the proficiency target as long as they reduced the percentage of failing students in a subgroup by ten percent relative to the prior year, providing a lower target for lower-performing school subgroups.

the effects of NCLB, I create a longitudinal student-level dataset in a manner following Ost (2014), but restrict to the school years ending in 2002 to 2008[14]. The data include test scores taken at the end of third through fifth grade, the pretest taken at the beginning of third grade, and student characteristics, including those relevant for the subgroup classifications under NCLB. NCLB was first effective in 2003, so the first year included in my dataset is 2002 to allow controlling for a student's prior-year test score. I exclude years after 2008 because the retesting of failing students in subsequent years changed the nature of the policy. I focus on elementary students for several reasons. First, this allows me to evaluate long-run effects with a greater time span between the education intervention and the long-run outcomes than would be possible for middle or high school students. Second, high school graduation rates, one of the long-run outcomes I evaluate, is included in the AYP objectives for high schools, complicating the interpretation of measured effects if I were to use high school students.

For long-run outcomes, I use data from the years 2010 to 2017 on students' outcomes at the end of high school, which are available in the NCERDC data. These outcomes include high school GPA, SAT scores, and indicators for taking the SAT, intention to attend a four-year college, intention to attend a two-year college, high school graduation, and dropping out of school[15]. The existence of a student identifier in the NCERDC data allows the elementary school students to be linked to their outcomes at the end of high school, but not all students can be linked to long-run outcomes[16].

Lastly, I use information from the AYP reports[17] for each school in order to pin down the number of students who were counted for AYP in each subgroup[18], as well as for information on well each school did on each of their AYP objectives. This is important for using the RDD based on school-subgroup student counts, since the counts of students in the end-of-grade test score data housed by the NCERDC likely vary from the actual counts used for AYP[19]. Following Gilraine, I drop

[14] These are the years included in Gilraine (2018)

[15] In theory dropping out is just the complement of graduating, but in practice it is more difficult to verify why a student leaves a school. The data includes indicators for reasons a student leaves a school, so the dropout variable is an indicator for when that reason is listed as dropping out of school.

[16] There could be a number of reasons students' outcomes at the end of high school cannot be found. If a student moves out of state or graduates from a private high school, they would no longer be in the data. The most likely explanation is that the identifying variables used by the NCERDC to match students across years contained missing values, errors, or variations that prevented matching. Examples include different spellings of names, nick names, school systems assigning their own SSNs, and parents not reporting SSNs. The NCERDC accounts for many sensible issues in their matching algorithm, but not all students could be reliably matched. For more information see the NCERDC report on creating a longitudinal student dataset: https://childandfamilypolicy.duke.edu/wp-content/uploads/2013/10/TECHREPT2.pdf.

[17] I gathered this data for 2003 to 2011 from the school-level reports from North Carolina's accountability website. See http://accrpt.ncpublicschools.org/docs/disag_datasets/. There is a separate webpage for each school for each year. The website is old and may no longer be available, but my dataset is available upon request.

[18] The count is based on students who attended the school for at least 75% of the school year.

[19] This could be due to students transferring schools during the school year, or missing data in the end-of-grade files.

from my analysis student-year observations where the student count for their subgroup-school-year is less than half of the count in the AYP report[20].

Table 1 shows descriptive statistics for the full sample as well as the subsamples used for the subgroup RDD and the school RDD. Column 1 contains the full sample described above, column 2 contains the subgroup RDD sample created by using a bandwidth of five around the cutoff of 40 students in a subgroup-school-year, column 3 contains the school RDD sample created by using a bandwidth of 0.08 around the AYP proficiency target. Math and reading scores are standardized within each grade-year to have a mean of zero and standard deviation of one. Due to the fact that white students make up a majority and school subgroups defined by minority students are more likely to contain around 40 students, the fraction of white students in the RD sample drops considerably from the full sample and the fractions of black and Hispanic students increase. Similarly, the fraction of economically disadvantaged students and English-language learners is greater in the RD sample. Given these demographics, it is unsurprising that elementary test scores are lower in the RD sample, by around 0.2 standard deviations. A lower fraction of these students take the SAT and those who do have lower SAT scores on average. They have a slightly lower average high school GPA, and they are slightly less likely to graduate high school or report an intention to attend college. The sample for the school RDD is much more similar to the main sample. There is a lot of within-school variation in student characteristics, such as test scores, and the school RDD sample is comprised of entire schools rather than subgroups within schools. These students have slightly lower test scores than students in North Carolina at large, and are slightly more likely to be a minority or be economically disadvantaged.

# 4 Empirical Strategy

## 4.1 Subgroup RDD

I first use a regression discontinuity design (RDD) that exploits the fact that schools were not held accountable for AYP proficiency targets for the NCLB-defined subgroups containing fewer than 40 students[21]. The basic idea is to compare, for example, a school with 40 white students to a school with 39 white students. Schools with 40 white students rather than 39 are unlikely to differ

[20]This amounts to about 7 percent of the observations.

[21]This design was used by Gilraine (2018), and a similar design was used in a different setting by Farber (2016).

systematically in terms of observed or unobserved characteristics. But the school with 40 white students will fail to meet AYP under NCLB if not enough of those students attain test scores at or above proficiency, while the school with 39 while students will be exempted from this criterion. To the extent a given subgroup has the potential to make the difference between the school meeting or not meeting AYP, which is the case if the school can pass all of their other AYP objectives, then the school faces a strong incentive to ensure an adequate percentage of students in this subgroup with 40 or more students attain proficient test scores. A school with fewer than 40 students in a subgroup has little incentive to worry about the test scores of the students in that subgroup[22]. This variation, while not a complete measure of the incentives under NCLB, provides clean identification of the effects of accountability pressure under standard assumptions required in an RDD.

I implement this strategy using the following regression equation:

$$y_{isgt} = \tau T_{sgt} + \theta X_{sgt} + \phi(X_{sgt} \times T_{sgt}) + \lambda_g + \delta_t + \beta Z_{isgt} + \epsilon_{isgt} \tag{1}$$

The outcome variables I measure are math and reading test scores, as well as the outcomes measured at the end of high school previously mentioned, denoted by $y_{isgt}$, where $i$ indexes the student, $s$ the school they attend, $g$ the NCLB subgroup they are a part of, and $t$ the year. $X_{sgt}$ is the running variable, defined as the number of students in student $i$'s subgroup and school minus 40; this controls for differences in average student characteristics for schools with more or fewer students in a given subgroup. $T_{sgt}$ is an indicator for $X_{sgt} \geq 40$, which is an indicator for the "treatment" status of accountability pressure being applied to the student's test score; $\tau$ is thus the parameter of interest. The interaction parameter $\phi$ allows for differential linear effects of subgroup size on either side of the 40 student cutoff[23]. I include fixed-effects, $\lambda_g$ and $\delta_t$, for which subgroup the student is in and the year. To control for other characteristics that may vary randomly or in correlation with subgroup counts moving away from the discontinuity, I include other student characteristics in the control vector $Z_{isgt}$: the student's math and reading test scores in the prior year, gender, grade fixed-effects, and an indicator for limited English proficiency. The regressions

[22] A student may be in more than subgroup. They will be in one subgroup defined by their race, and could be in an additional subgroup if the student is classified as economically disadvantaged, limited English proficiency (LEP), or a student with a disability (SWD). In this case, the school may still have an incentive to worry about a student in a subgroup numbering 39 if the student is also in another subgroup numbering 40 or more. This is more likely to be a concern for economically disadvantaged students, since LEP and SWD are smaller subgroups that tend to number below 40. Any such accountability pressure on students in a subgroup between 35 and 39 would attenuate estimated effects of accountability pressure.

[23] I also run local linear specifications that allow for more flexible controls for the running variable around the discontinuity.

are restricted to a bandwidth of 5, so only students in a subgroup with 35 to 45 students are included.

In order to interpret $\tau$ as the causal effect of a school being held accountable for a subgroup on those students' test scores or long-run outcomes, two primary standard assumptions are required. First, after controlling for the variables included in equation 7, there should be no discontinuity in factors affecting student outcomes around the 40 student cutoff except for the variation in accountability pressure. Given that schools may desire to have fewer than 40 students counted in a subgroup to prevent being held accountable for that subgroup's test scores, manipulation of the running variable is a common concern that must be addressed. Most of the characteristics that define NCLB subgroups are not readily manipulable, such as a student's race or eligibility for free or reduced-price lunch. However, schools could potentially use their discretion to strategically classify a student as having a disability or not, which Jacob (2005) provides evidence for in the context of an accountability scheme in Chicago. I thus exclude students with disabilities, following Gilraine. Additionally, showing that there is no discontinuity in the observable characteristics of students reduces the concern of a discontinuity in unobservable factors. Second, the running variable is discrete in this setting and including some observations outside of the closest vicinity to the cutoff is required for statistical power. Thus, a causal interpretation requires the assumption of no difference in factors affecting student outcomes across subgroup counts included in the bandwidth, after conditioning on control variables, which include a separate linear effect of subgroup count on either side of the bandwidth in my baseline regressions. To provide evidence for this concern, I test multiple bandwidths and alternative specifications to control for the running variable.

Several additional threats to identification remain. First, the "sharp" RDD requires the maintained assumption that treatment probability is 0 on one side of the discontinuity and 1 on the other. Using the data from schools' AYP reports, I can ensure that this is the case. However, as also documented by Gilraine, the limited English proficiency subgroup is an exception, perhaps because North Carolina did not require students newly-added to this subgroup to be tested. Thus, this subgroup is dropped from the analysis. Second, there could be spillover effects, where school-subgroups designated in the "control" group are affected by efforts aimed primarily at the "treatment" group. This would arise in my empirical strategy if a school has multiple subgroups with around 40 students, resulting in students in one subgroup being affected by the accountability pressure on another subgroup in the same school. However, few schools have multiple subgroups within the bandwidth used in the RDD.

## 4.2 School RDD

An alternative identification strategy that has been used in the literature to estimate the effects of school accountability pressure on student outcomes is based on the fact that sanctions under No Child Left Behind really did not begin in earnest until after a school failed to make AYP for two years in a row. After failing a first time, the school had to create a school improvement plan. But the costly sanctions begin after failing a second time, where the school then had to offer students the opportunity to transfer to a better school in the district. Thus several studies, including Chakrabarti 2014 and Ahn and Vigdor 2014), used a regression discontinuity design comparing schools that barely passed AYP in the previous year to schools that barely failed AYP in the previous year. Restricting the analysis to schools that did not fail AYP in the year before the previous year, this allows exploitation of this added accountability pressure that comes after a school fails AYP for the first time. Schools that barely fail AYP for the first time are very similar on average to schools that barely make AYP, except for the fact that the failing school now faces a threat of costly future sanctions, starting with the mandated transfer offers, if they fail AYP again in the following year.

This regression discontinuity design is shown in the following specification:

$$y_{ist} = \tau F_{s,t-1} + \theta M_{s,t-1} + \phi(M_{s,t-1} \times F_{s,t-1}) + \beta Z_{ist} + \epsilon_{ist} \tag{2}$$

The outcome of interest for elementary student $i$ in school $s$ in year $t$, first their test scores at the end of the year, and second their outcomes at the end of high school, is $y_{ist}$. The parameter of interest, the effect of one year of accountability pressure, is $\tau$. $F_{s,t-1}$ is an indicator for the student's school failing AYP in the previous year. This creates the variation in accountability pressure on the students' test scores in the school, since having failed in the prior year creates pressure for the school to avoid the costly sanctions that begin if they fail a second time. The running variable in the RD specification is $M_{s,t-1}$, which is defined as the minimum distance between the school's AYP proficiency objectives and AYP proficiency counts in the prior year. The parameters $\theta$ and $\phi$ control for a school's distance from the AYP target they performed worst on, with a separate trend on either side of the discontinuity. $Z_{ist}$ includes the same controls used in the main analysis.

Each school has many objectives they have to meet in order to make AYP, and they fail AYP

if they fail any one of the objectives. Each school has a proficiency target for both math and reading for each subgroup of students numbering at 40 or more, as well as an overall proficiency target for the entire student body, and an objective for student attendance[24]. Thus, a minimum distance is calculated to determine which schools "barely passed" or "barely failed" AYP, based on the proficiency target the school performed worst on. Schools with a negative minimum distance had an objective for which their proficiency count was below the target required for AYP, indicating AYP failure. Schools with a positive minimum distance had no objectives for which their proficiency count was below the target required for AYP, indicating they should have made AYP. Whichever objective the school performs the worst on, relative to their target, this objective is the one used to determine where the school stands relative to AYP.

Two important exceptions used in NCLB make calculating this minimum distance more difficult. The first is the confidence interval exception. This exception states that if a school is within a 95% confidence interval of the proficiency target for a specific objective, they are counted as meeting the objective. The second is the safe harbor exception, which was implemented to protect schools starting off far below the required proficiency targets. This exception states that if a school reduces the fraction of students who failed to reach the proficiency test score in the previous year by ten percent, then the school is counted as meeting that specific AYP objective.

I calculate the minimum distance for each school in each year, following the rules specified by North Carolina in that year. After calculating this minimum distance, I test for accuracy in my calculation of minimum distance by running an RD analysis showing the relationship between minimum distance and an indicator for the school actually failing AYP in that year. The information on whether or not the school actually fails AYP in a given year comes from the AYP reports on the North Carolina Department of Education website referenced previously. Since the scale of the math and reading tests changed a few times in the time period of the analysis, and because accounting for the exact nature of the AYP exceptions is complex, I cannot perfectly calculate the minimum distance for each school in each year. Thus, there are a few schools with minimum distance greater than zero who actually failed AYP, and a few schools with minimum distance less than zero who actually passed AYP. Thus, following previous studies, I use a fuzzy RDD, with the first stage being the effect of the calculated minimum distance on an indicator for AYP failure, and the second stage being the effect of AYP failure on student outcomes in the following year.

[24]Essentially all schools met student attendance objectives, particularly in North Carolina, so following previous studies, I only evaluate math and reading proficiency objectives.

The first stage of the fuzzy RDD is plotted in Figure 1. The discontinuity estimate is around 0.75, meaning that having a calculated minimum distance just below the cutoff relates to a 75% greater probability of failing AYP. This is similar to the strength of the first stage shown in Ahn and Vigdor (2014).

### 4.3 Robustness

I test the robustness of these two RDDs using several standard tests. First, I check to make sure there is no bunching of observations to one side of the discontinuity, which may indicate strategic manipulation to keep subgroups below 40 students or keep proficiency counts high enough to avoid AYP failure. The density of the running variable and the p-value for testing a significant difference in density across the cutoff is shown in Appendix Figure A1 for both the subgroup RDD and school RDD. There is no significant evidence of this type of manipulation.

Second, I test for any discontinuities in observable student characteristics around each RD cutoff, which would may indicate non-random sorting of certain types of students to one side of the discontinuity. If there exists a discontinuity in an observable variable that is predictive of test score achievement or better long-run outcomes, this would bias my results if I did not control for it in the RD specification. More importantly, it may suggest that students differ in unobservable characteristics around the discontinuity as well, which would bias my results. I test for these discontinuities in observable characteristics in Appendix Table A2, finding no significant pattern of non-random sorting across the cutoff.

Third, I check for robustness of test score and long-run effect estimates to bandwidth selection. Results are shown in Appendix Figure A2 for test score effects for both the subgroup RDD and school RDD, in Appendix Figure A3 for long-run outcomes using the subgroup RDD, and Appendix Figure A4 for long-run outcomes using the school RDD. Effects are slightly larger in both RDDs when using smaller bandwidths, but significant effects persist across a broad range of potential bandwidths.

# 5 Results

## 5.1 Test Scores and Long-run Outcomes

Table 2 shows the results of estimating the regression for the subgroup RDD specified in Equation 1, followed by the results for the school RDD specified in Equation 2. The first panel of each table shows results with the students' math test scores as the outcome variable, and the second panel shows results for reading test scores. The coefficient is shown for the parameter of interest, $\tau$, which is the effect of one year of accountability pressure on a student's test score at the end of the year, identified at the discontinuity of either 40 students in a subgroup for the subgroup RDD or AYP failure in the prior year for the school RDD. The data used in the subgroup RDD is the subset of student-year observations in a subgroup with 35 to 44 students within the school in that year; for the school RDD, the regressions are restricted to students in a school close to failing for the first time in the prior year (within approximately 0.08 fraction proficient on the school's lowest performing criterion). Each column in each panel represents a separate regression, using a local linear estimation with a triangular kernel to flexibly control for the running variable near the discontinuity[25]. In all columns, I control for subgroup fixed-effects (e.g. an indicator for if the student is economically disadvantaged) as well as year fixed-effects. Regressions shown in columns 2 and 4 include additional student controls, including the student's math and reading test scores in the previous year, gender, limited English proficiency, and grade fixed-effects. While the validity of the RDD implies no systematic difference in observable characteristics, controlling for these additional variables may help eliminate noise due to random differences across the discontinuity in these student characteristics affecting achievement. Standard errors are clustered at the subgroup-by-count level in the subgroup RDD, following Lee and Card (2008)[26], and at the school level for the school RDD.

In line with previous literature identifying the causal effect of NCLB accountability pressure, I find modest statistically significant positive effects of accountability pressure on student test scores. Figures 2 and 3 show the RDD plots for the subgroup and school RDDs, respectively. In Figure 2,

[25] Results are similar using a simpler OLS estimation with a linear trend on either side of the discontinuity.

[26] This means that the error terms are allowed to be correlated within a given subgroup definition and number of students in the subgroup ($X_s gt$). With 7 subgroups included in the analysis and 11 subgroup counts within the bandwidth of 5, this results in 77 clusters, alleviating the concern of an insufficient number of clusters when clustering at the level of the discrete running variable. Kolesar and Rothe (2018) discuss issues with this type of clustering, which was proposed in Lee and Card (2008). Robust standard errors are moderately smaller than the ones shown in Table 2. Standard errors clustered at the school level are somewhat larger, but effects are still statistically significant.

there exists a negative relationship between subgroup size and test scores, which is sensible given that these students are lower-achieving on average and schools with more of these students may differ as a result. Gilraine (2018) estimates a similar RDD using the NCERDC data and finds very similar effect sizes. The test score effects are larger for math, where estimates range from roughly 0.06 to 0.08 test score standard deviations (SDs). This roughly equates to over half of the effect of having a 1 SD better teacher, as estimated in papers such as Chetty, Friedman, and Rockoff (2014). For reading test scores, the estimates range from roughly 0.04 to 0.06 SD. One plausible explanation for larger math score effects is that math skills measured by the standardized tests are more easily taught than reading skills, making the return to effort greater for math. Another explanation is that students failing to reach proficiency in math is more likely to alter a school's probability of meeting or failing AYP on the margin. Gilraine (2018) notes that about twice as many schools fail AYP for math relative to reading, lending support to this explanation. Effect sizes are similar for the school RDD, ranging from 0.05 to 0.09 SD for math and 0.04 to 0.10 SD for reading. While effects are somewhat smaller when adding controls for the student's prior-year test scores, these results come with the caveat that it is potentially problematic to control for a prior-test score that may have itself been affected by accountability pressure. All of these estimates taken together, I conclude that accountability pressure resulted in a modest to medium-sized positive effect on math and reading test scores.

The primary contribution of this paper is evaluating the long-run effects of accountability pressure under NCLB. Table 3 shows the results of estimating local linear regressions of the form shown in Equation 1 for the subgroup RDD, but now with the outcome variables being the long-run outcomes measured at the end of high school. There are two columns for each long-run variable, with the first showing the result of the regression including year fixed-effects and subgroup fixed-effects as controls, and the second adding the same additional student controls referenced in Table 2. The number of observations included in the regression varies across outcome variables, because the end-of-high-school outcomes come from different files and elementary school students in the dataset are linked to an end-of-high-school outcome with different degrees of attrition for different variables. Naturally, SAT scores contain the fewest students, since fewer than 30 percent of students in the sample take the SAT. For high school GPA, a little over half of the elementary students are included. About 65 percent have data for intention to attend a 4-year college. Roughly 80 percent have data on dropping out or graduating. All are included for SAT-taking, which is an indicator variable for having an SAT score in the data. The potential bias created by this attrition is discussed in Section

7.3.

The estimates in Table 3 show small and statistically insignificant effects of accountability pressure in a year of elementary school on long-run outcomes related to educational attainment at the end of high school. The estimates for dropping out / graduating high school, taking the SAT, and intention to attend any college, a 2-year college, or a 4-year college all show effects quite close to zero or perhaps even slightly in the negative direction. The top of the 90% confidence interval is 0.4 percentage points (pp) for the probability a student graduates high school, 0.3 pp for the probability a student intends to attend any college, and 1.5 pp for the probability a student intends to attend a 4-year college. Thus, I can rule out any large effects on these variables capturing educational attainment. However, there is a large and significant positive effect on SAT scores. Interestingly, SAT scores may involve the skills most closely related to those used for standardized tests in elementary school, relative to the other long-run outcomes. Given the mean and SD of SAT scores, this effect is close to a 0.10 SD effect. Lastly, I find that increased accountability pressure for students in subgroups just above the 40-student cutoff led to slightly higher end-of-high-school GPAs, around a 0.04 SD increase based on the distribution of GPA. The accompanying subgroup RDD plots are shown in Figure 4. There we see no noticeable discontinuity for most long-run outcomes, but a significant jump in SAT scores going from 39 to 40 students in a subgroup, and a smaller jump for high school GPA.

Effects on long-run outcomes using the variation in accountability captured by the school RDD are shown in Table 4. Broadly speaking, results are generally similar to those yielded from the subgroup RDD. There is again no effect on high school graduation, with the top of the 90% confidence interval at 0.2 pp. There is also a large observed effect on SAT scores, around 0.10 to 0.15 SD based on the SAT score distribution. However, in contrast to the subgroup RDD results, there is no effect on high school GPA (I can rule out an effect larger than 0.05 SD). This may be due to the school-level variation in accountability pressure captured by the school RDD. While variation captured in the subgroup RDD is accountability pressure targeted at a specific subgroup within a school (all other students in the school are excluded from the regressions), failing AYP potentially results in pressure on the entire school. Students in the same elementary school are very likely to end up in the same high school and GPA is largely a relative measure within a school. Thus, it may be unsurprising that GPA is unaffected by this variation in accountability pressure, even if cognitive skills improved (as captured by math and reading test scores). Lastly, another difference between the school RDD and subgroup RDD results is that I find a considerable, though noisy,

positive effect on intention to attend a 4-year college, around 0.04 pp. This is exactly offset by a decrease in intention to attend a 2-year school, indicating that students in a school under accountability pressure due to failing in the prior year shifted from 2-year to 4-year colleges. If students intending to attend a 2-year school out of high school are less likely to complete a bachelor's degree or eventually pursue graduate degrees, this could have a positive effect on educational attainment. One potential reason for this effect being found with the school RDD but not the subgroup RDD is that the school-level increase in accountability pressure may have increased skills and expectations for a student's entire peer group, leading to a shift from 2-year to 4-year colleges. In contrast, for the subgroup RDD, accountability pressure targeted at a smaller subset of students within a school may not have had the same effect.

## 5.2 Long-run effects on GPA in Specific Subjects

Given the positive effect of accountability pressure on math and reading scores in elementary school as well as SAT scores in high school, one might expect these increased math and reading skills to result in higher high school GPA as well, particularly in classes most related to math and reading. To assess this, I calculate high school GPA separately by subject based on transcript data, and repeat the subgroup and school RDD analysis with various types of subject-specific GPA as the outcome variables.

Results using the subgroup RDD are shown in Table 6. After finding a positive effect on total high school GPA in Table 4, I find positive effects on high school GPA across a range of subjects. Effects are largest for GPA in math classes, with students in a subgroup just above the 40-student cutoff having around 0.08 higher GPA in math classes. Effect sizes are smaller and roughly the same for GPA in other subjects, including reading, science, and social studies, around 0.04 to 0.05 grade points. Given the larger effects on elementary math scores relative to reading scores in Table 2, it may be expected that GPA in math classes should increase more than GPA in reading (language arts) or other classes. If accountability pressure on math and reading test scores in elementary school led to a long-run increase in math- and reading-specific skills, one might expect a larger GPA effect in math and reading classes in high school relative to other types of classes. However, there does not seem to be a difference between the increase in reading GPA and the increase in GPA in non math or reading classes. It may be the case that the cognitive skills increased as a result of accountability pressure, captured in math and reading scores, were broad enough to translate

to other academic classes in subjects like science or social studies. One might expect fundamental math and reading skills to be useful for these other subjects as well. It is also worth noting that the lack of negative effects in any subjects outside of math and reading may allay concerns that pressure on tested subjects (math and reading) may have lead to a deleterious effect on learning in non-tested subjects (social studies, science, or the arts).

Results using the school RDD are shown in Table 7. I find no effects on high school GPA within any subject. Although the overall lack of effects is expected given the null effect on GPA in Table 5, it is worth cautiously noting that the pattern of relative effects across subjects is consistent with the pattern found using the subgroup RDD. While there are statistically insignificant small negative effects on non-math subjects, the effect on GPA in math classes is close to zero. Thus, though the statistical precision of estimates precludes any strong conclusions based on Table 7 alone, the relatively higher GPAs for treated student in math classes lends some additional support to the hypothesis that math-related skills were more positively affected by accountability pressure.

### 5.3 Persistent Effects on Math and Reading Skills

The positive effects of accountability pressure on elementary math and reading test scores shown in Table 2, combined with the positive effects on SAT scores in high school shown in Tables 3 and 4, might suggest a persistent increase in math and reading skills captured by these standardized tests. To further test this hypothesis, I now consider the effects of accountability pressure on student skills in math and reading measured over time, between the contemporaneous test scores measured in elementary school and the long-run outcomes measured at the end of high school. In particular, I look at the effects of accountability pressure in a year of elementary school on math and reading test scores in subsequent years of school, as well as the effects on GPA in math and reading classes in each year of high school.

I first estimate Equations 1 and 2, but now with the outcomes $y_i sgt$ being the student's test score in the year of treatment, the next year, two years, and three years after the year in which they are treated by accountability pressure [27]. The effects of accountability pressure in a given

[27] Since the data includes third- through fifth-graders, this means I assess fourth-, fifth-, and sixth-grade standardized test scores for third-graders; I use fifth-, sixth-, and seventh-grade test scores for fourth-graders; and I use sixth-, seventh-, and eighth-grade test scores for fifth-graders. I stop at eighth-grade, the end of middle school, because that is the last year in which students take end-of-grade tests in math and reading. In high school, students take end-of-course exams, but students may choose to take different courses in high school, or may take a given course in different years of high school.

year of elementary school on students' math and reading test scores in the next three subsequent years are shown in Table 7. Each element in the table shows the estimated treatment effect for a local linear regression, where the outcome variable is either math or reading test score, measured in the same year, or one, two, or three years after exposure to accountability pressure. Even three years after the year in which the student is exposed to accountability pressure due to being in a subgroup just above the cutoff, students experience math and reading test score gains that are not significantly smaller than the contemporaneous test score effects shown in Table 2. Math test score effects persist one, two, and three years later, with an effect size around 0.06 SD with the subgroup RDD and around 0.06 to 0.08 SD with the school RDD[28]. Reading test score effects are persistent as well, with a persisting effect around 0.08 SD with the subgroup RDD and around 0.06 to 0.07 SD with the school RDD.

Next, I estimate effects of accountability pressure on GPA specific to math and reading classes in each year of high school to evaluate math and reading skills even further into the future. Estimates are shown in Table 8. Each element in the table shows the estimate for a local linear regression where the outcome is the GPA in a specific grade of high school (9th, 10th, 11th, 12th) in classes of a specific subject (math or reading). In line with the main results discussed previously, there is no effect on GPA in either math or reading classes using the school RDD, for any grade of high school. However, as we saw an increase in math and reading high school GPA in Table 6, here we see positive effects on both math and reading GPA through each grade of high school. Effects seem to be more persistent in math GPA, around 0.07 to 0.09 grade points through 11th grade, with a drop to around 0.04 grade points in 12th grade. Effects are positive for each grade in reading GPA as well, decreasing from roughly 0.07 grade points in 9th grade to around 0.03 grade points in 12th grade.

Figure 8 plots the effects on math and reading skills over time estimated in Tables 7 and 8. Test score effects are measured in terms of a fraction of standard deviation in the test score distribution, and GPA effects are measured in grade points (out of 4). Here we see a reasonable explanation for the positive effects of accountability on elementary test scores and SAT scores in high school: accountability pressure had a persistent effect on student achievement in math and reading. This is first seen in the persistent standardized math and reading test score effects in the three years following exposure to extra accountability pressure. Then for the subgroup RDD,

[28]Estimates are noisy for the school RDD: math effects in future years are not quite significant and reading effects in future years are only marginally significant.

capturing variation more specific to a small subset of students within a school, this also translates to higher GPA in math and reading classes through high school. While standardized test scores and GPA in high school are different measures that cannot be directly compared in magnitude, this shows the persistent effects of accountability pressure in elementary school on academic achievement outcomes through middle school and high school. Accountability pressure increased some forms of math and reading skills, particularly those captured by standardized tests (like the ones used for accountability or SAT exams), and perhaps captured by higher grades in math and reading classes in high school as well.

# 6 Discussion

## 6.1 Commensurate Long-run Effects

As shown in Section 2, a multitasking model suggests that accountability could lead to positive, negative, or zero effects on long-run outcomes. A reasonable interpretation of the null effects on long-run outcomes given positive effects on test scores is that these results are due to an offsetting of two factors: teachers increasing effort in a way that increases test scores and long-run outcomes, and teachers changing the focus of these efforts in a way that increases test scores but diminishes learning that matters for long-run outcomes. However, another explanation is that given the relationship between a student's test scores and long-run outcomes, some of the long-run effects that should be expected to accompany the test score increases shown in Table 2 are too small to be statistically distinguished from zero. To address this possibility, I do a back-of-the-envelope exercise to estimate "commensurate" long-run effects, and compare these commensurate effects to the actual long-run effects shown in Tables 3 and 4. Based on the confidence intervals around the actual long-run effect estimates, I can provide some evidence as to whether or not I can reasonably rule out commensurate long-run effects.

A commensurate long-run effect is the product of two components: the effect of accountability on tests scores, and the predicted improvement in long-run outcomes associated with a test score improvement of that magnitude. Explicitly, first I take the estimated treatment effect of accountability on math and reading scores, $(\tau_{math}, \tau_{read})$, from the subgroup RDD results and then the school RDD results. Then I estimate the relationship between math and reading test scores in

elementary school and long-run outcomes measured at the end of high school. I extend my full sample shown in column 1 of Table 1 to the years 2002 to 2011 and regress each of the long-run outcomes on student math and reading scores in elementary school, conditioning on the student characteristic controls used previously:

$$LR_i = \beta_1 Math_i + \beta_2 Read_i + \beta_3 Controls_i + \epsilon_i. \tag{3}$$

The commensurate long-run effect, $C(LR_i)$, is then defined by the product of the test score effects and the relationship between test score gains and long-run effect improvements:

$$C(LR_i) = \hat{\beta}_1 \tau_{math} + \hat{\beta}_2 \tau_{read}. \tag{4}$$

Appendix Table A3 shows the results of the regressions specified in Equation 3. The first panel shows results when only using basic controls for year, grade, and subgroup fixed-effects, showing the relationship between test scores and long-run outcomes. The second panel includes controls for students' test scores in the prior year, and thus shows the relationship between test score gains and long-run outcomes. Arguably, this relationship between test score gains and long-run outcomes is the most natural benchmark for thinking about commensurate long-run effects, since the effect of accountability test scores represents an increase in test scores. This shows us how large of an increase in long-run outcomes is associated with a one standard deviation increase in a student's math or reading scores. Test score levels are more likely to be a result of family inputs and other non-school inputs, which predict long-run outcomes.

Appendix Table A3 shows that while both elementary math and reading scores are strongly related with long-run outcomes, higher math scores are associated with larger improvements in long-run outcomes. For example, consider two students with the same test scores in the prior year, and the same reading score in the current year of elementary school, but one student has a 1 SD higher math test score in the current year. This student with the 1 SD math test score gain is predicted to be 6.6 pp more likely to take the SAT, 3.9 pp more likely to graduate high school, 3 pp more likely to intend to attend any college, and 10 pp more likely to intend to attend a 4-year college. They are also predicted to score 65 points (0.34 SD) higher on the SAT and to have a 0.2 higher high school GPA (0.27 SD).

These relationships line up well with previous literature. Rothstein (2017) uses the same North Carolina data and these end-of-high-school outcomes and replicates the main findings of Chetty, Friedman, Rockoff (2014), showing that an increase in teacher value-added is associated with significantly better long-run outcomes. Rothstein (2017) shows effect sizes commensurate with the relationships I document in Table A3. While Chetty, Friedman, and Rockoff were able to use better measures of long-run outcomes, including income in adulthood and college attainment, Rothstein provides evidence for the usefulness of these end-of-high-school variables as a proxy for longer-run outcomes. Additionally, Chetty, Friedman, and Rockoff run similar regressions to the ones I estimate in Table A3, using their long-run outcomes, and find similar relationships between test scores and long-run outcomes. In particular, the best comparison I can make is between college attendance at age 20 in their data and intentions to attend college in my data: they find that a 1 SD increase in test scores is associated with a 0.054 pp increase in the probability of being in college at age 20, conditional on controls. Given that some 20-year-olds have completed two years since high school and some have not, it makes sense that their coefficient for college attendance at age 20 is in between my coefficients for intention to attend a 4-year college and intention to attend any college.

Given the coefficients in the second panel of Table A3, showing the long-run improvements associated with increases in test scores, and the coefficients in Table 2, showing the effect of accountability pressure on test scores, I then calculate commensurate long-run effects using Equation 4. I compare these commensurate effects to the actual long-run effects from Tables 3 and 4, and the confidence intervals around the actual long-run effects. The results are shown in Appendix Table A4.

In general, the estimated commensurate long-run effects are within the confidence intervals of the estimated actual long-run effects. The subgroup RDD showed a large positive effect on SAT scores and a small positive effect on high school GPA, with no effect on other long-run outcomes like high school graduation or intention to attend college. The commensurate effect SAT scores is toward the lower end of the confidence interval around the actual effect on SAT scores, but still within it. For GPA, the actual effect is very close to the commensurate effect. For the other outcomes, commensurate effects are toward the top of the confidence intervals around actual effects, but still generally within them. Although for high school graduation and intention to attend any college, I can marginally rule out the commensurate effects. The school RDD showed similar results to the subgroup RDD, with two differences: no effect on high school GPA, and a shift from intention to attend 2-year colleges to intention to attend 4-year colleges. But similarly to the subgroup RDD,

this exercise suggests that the commensurate long-run effects are within the confidence intervals of the actual effects using the school RDD. The commensurate effects for SAT scores and intention to attend 4-year colleges are toward the lower end of the confidence intervals, but well within them. For the outcomes with null effects, commensurate effects are generally within the confidence intervals of actual effects, with the exception of high school graduation.

Altogether, this exercise suggests that it is difficult to rule out positive test score effects resulting in commensurate, though small, improvements in long-run outcomes. However, I do find large positive effects on SAT scores, indicating a particular increase in skills captured by the SAT relative to other skills that would lead to improvements in high school GPA, high school graduation, or intention to attend college. Additionally, for both the subgroup RDD and school RDD, I can marginally rule out commensurate effects on high school graduation, perhaps consistent with elementary test scores being increased in a way that was too narrowly focused to improve graduation rates. This could be due to either a lack of focus on skills that make students more likely to graduate, or due to a smaller effect on student more on the margin of graduating high school or not, which would be lower-achieving students.

## 6.2 Heterogeneous Effects

Accountability pressure created by NCLB may be expected to have more effects on some types of students and schools than others for a variety of potential reasons. First of all, the design of the policy may have put more pressure on certain schools or types of students. Schools may have an incentive to focus on students who are more on the margin of reaching the proficiency test score or not. Schools more on the margin of passing or failing AYP may have had a stronger incentive to improve test scores. Both of these dimensions of potential variation in pressure may also have led to greater pressure on certain demographics of students, either for certain racial groups or for economically disadvantaged students. Second, schools may have an easier time improving the outcomes of certain types of students, or certain types of students may benefit more from schools' responses to accountability pressure.

In this subsection I evaluate the heterogeneous effects of accountability pressure on test scores and long-run outcomes. I do this by restricting RDDs to the relevant subsets of students. First, given the large effect on SAT scores, I ask whether or not this can be explained by stronger effects

of accountability on the types of students who end up taking the SAT. Second, I use the subgroup RDD to evaluate whether or not schools were more on the margin or passing or failing AYP induced larger effects. Third, to try to bring together the different dimensions of variation captured by the subgroup RDD and school RDD, I use the school RDD to evaluate if failing AYP in the previous year leads schools to improve outcomes more for students in subgroups of 40 or more, since those students are the ones who matter for making AYP. Fourth, I test for effects on students more on the margin of proficiency. Lastly, I evaluate effects on the different subgroups defined by NCLB, including racial groups and economically disadvantaged students.

### 6.2.1 Effects on SAT-takers

I find a considerable effect of NCLB accountability pressure on test scores in elementary school and a large effect on SAT scores in high school. One explanation for this is that persistent math and reading skills were improved. The analysis in Section 6.3. showed increased test scores and perhaps increased math and reading GPA in the intermediate period connecting the immediate elementary test score effects and the SAT effects in high school, providing some evidence in favor of this explanation. However, another possible explanation for the large SAT score effects is that the accountability pressure had larger effects on the types of students who took the SAT. Since only around 30 percent of students in the elementary sample are observed taking the SAT in high school and taking the SAT is a step toward college attendance, one would expect that higher-achieving students are the ones taking the SAT. To consider this second possibility, I restrict both RDDs to the subset of students taking the SAT in high school and evaluate accountability effects on their test scores in elementary school and other long-run outcomes at the end of high school. While SAT taking is an endogenous variable, I do not find an effect of accountability pressure on SAT taking. Furthermore, these results are merely meant to assess whether stronger effects on these students taking the SAT are a mechanism for the large SAT score effects.

The effects of accountability pressure on elementary test scores, using both the subgroup RDD and school RDD, are shown in Appendix Table A5. Each element of the table is the treatment effect estimated from a local linear regression restricted to either SAT-takers or non SAT-takers, using the outcome variable on the left (math or reading test scores). Following the main analysis, all regressions include year and subgroup fixed-effects, and the second and fourth columns include additional student controls including prior-year test scores. Here we see test score effects on SAT-

takers similar to the effects on students not taking the SAT, around 0.07 SD. Appendix Tables A6 and A7 show the effects on long-run outcomes for SAT-takers and non SAT-takers. The outcome variables are listed on the left, estimated in local linear regressions restricted to SAT-takers or non SAT-takers. For the subgroup RDD results in Appendix Table A6, there are no effects on most outcomes for either group of students. However, it appears that the small positive effect on GPA, shown in Table 3, is more concentrated in the students who end up taking the SAT. This may indicate a small role for heterogeneous effects on SAT-takers to explain the large SAT score effects, but given the small difference in GPA effects and no difference in test score effects, it is unlikely to be a major factor. For the school RDD results in Appendix Table A7, we see no significant differences between SAT-takers and non SAT-takers. I conclude from all of these results that stronger effects on SAT-takers are unlikely to be a major reason for the large effects of accountability pressure under NCLB on SAT scores. Rather, it is more likely to have been the result of persistently increased math and reading skills for a broader set of students.

#### 6.2.2 Effects on Students in Schools on the Margin of Failing AYP

Although the safe harbor provision allowed low-achieving schools to pass AYP by reducing the fraction of students not proficient by ten percent, the design of NCLB still put much more pressure on schools with lower achieving students, whether or not the schools were the main causal factor for such low achievement. Given the discrete nature of the AYP criteria, where schools had to attain a certain fraction of math and reading proficiency for each subgroup and the student body as a whole, one might expect a stronger response to accountability for schools expecting to be close to failing or not failing AYP. Thus, using the minimum distance to AYP defined in Section 5.2., I group schools into two groups of equal size: those with a minimum distance closer to zero, and those with a minimum distance further from zero. Those closer to zero are more marginal, meaning improvements in test scores are more likely to result in making the difference between the school passing or failing AYP. The schools further from zero could either be performing much below their AYP target or much above it. Based on the school distribution of minimum distance to AYP, there are more schools far above AYP than far below it.

Since the school RDD captures variation between schools barely passing or barely failing AYP, all schools in the RD sample are on the margin of passing or failing AYP. Thus I use the subgroup RDD, restricting separately to schools close to and further from the AYP target, to evaluate whether

or not there was a larger effect for more marginal schools. The effects on elementary tests scores for marginal and not marginal schools are shown in Appendix Table A8. Interestingly, I find a slightly smaller math test score effect for students in schools on the margin of failing AYP, and a larger reading test score effect for students in schools on the margin of failing AYP. The precision of these estimates precludes strong conclusions, but it seems there is no clear pattern of a test score response concentrated amongst the schools more on the margin of failing AYP. Appendix Table A9 shows the effects on long-run outcomes. Effects are generally similar across marginal and not marginal schools, but there may be a small shift from 2-year to 4-year colleges for students in marginal elementary schools. Altogether, these results suggest the main subgroup RDD results are not mostly driven by the set of schools more on the margin of passing or failing AYP.

#### 6.2.3 Effects on Students in Subgroups of Fewer than 40

The subgroup RDD captures variation in accountability pressure across schools specific to students in a particular subgroup numbering around 40 within a school, since schools do not have to worry about test scores for subgroups fewer than 40. The school RDD captures variation in pressure across schools that barely pass or fail AYP in the prior year for the first time, since failing AYP results in increased pressure to avoid failing AYP again and incurring sanctions. To try to bring these two dimensions of variation together, I use the school RDD to test for stronger effects among subgroups of students above 40. If schools are able to target improvement efforts toward their students not in small subgroups of students fewer than 40, smaller effects in these subgroups below 40 would make sense. If schools respond to accountability pressure after failing AYP once by instituting school-wide changes, effects may spill over to subgroups below 40 as well.

Appendix Table A10 shows the effects of accountability pressure on elementary test scores, separately for students in subgroups below 40 and students in subgroups of 40 or more. Comparing columns 1 and 3, it appears there may be a slightly larger test score effect for the students in subgroups of 40 or more. The math estimate is 0.09 SD for subgroups of 40 or more and 0.06 SD for subgroups fewer than 40, and the reading estimates are 0.12 SD and 0.06 SD, respectively. However, estimates are somewhat noisy, particularly for subgroups below 40, due to the lower number of observations given by small subgroups, precluding a clear conclusion of differential effects. Furthermore, when including additional student controls (most notably prior-year test scores), the larger effects for subgroups of 40 or more disappear.

Appendix Table A11 shows the effects of accountability pressure on long-run outcomes, separately for students in subgroups below 40 and 40 or more. In general, effects are similar for students in both groups, with null effects on most outcomes, and a large increase in SAT scores. However, the shift from intending to attend 2-year colleges to intending to attend 4-year colleges in the main school RDD results, shown in Table 4, seem to be driven by students in subgroups of 40 or more. One potential reason for this might be the types of school responses to NCLB accountability pressure, which may be more school-level efforts and changes rather than measures targeted at specific students. Another possible reason might be smaller effects on the types of students represented in subgroups below 40, which are on average lower-achieving and higher-fraction minority and economically disadvantaged. These groups may be less affected by school-level changes, or less likely to be on the margin of deciding whether to attend a 2-year or 4-year college.

#### 6.2.4 Effects on Students on the Margin of Proficiency

NCLB set accountability targets for schools that were based on proficiency counts. As long as a student attained math and reading scores above the proficiency cutoff for each test, it did not matter for NCLB how high the student's test scores were. Similarly, if a student scored below proficiency, it did not matter how low the student's score was. An intuitive strategic response to this policy feature would be for schools to focus their efforts on improving the test scores of students who they expected to have test scores close to the proficiency cutoff, since improving the test scores of these students may make the difference between the student reaching the proficiency test score or not. Neal and Schanzenbach (2010) show that this was indeed the case in Chicago Public Schools, using the introduction of a district-level pre-NCLB accountability policy in 1996 as well as NCLB in 2002. Thus, I use the subgroup RDD and school RDD to estimate test score and long-run effects specifically on the more marginal students relative to the proficiency score.

To define marginal students, I use each student's demographic characteristics and prior-year test scores to predict their test scores in the current year. I then calculate the distance between their predicted test score and the test score required for proficiency, for math and reading. Then I divide students into two groups: those above and below the median in distance from the proficiency score. Those below the median in distance from the proficiency score are deemed marginal. Since the test scores required for proficiency are low, well over 1 SD below the mean, the students predicted to

be closer to the proficiency score are essentially all lower-achieving students[29].

The effects of accountability pressure on elementary test scores for students marginal and not marginal to proficiency are shown in Appendix Table A12. For the subgroup RDD, higher math and reading test score seem to be more driven by effects on marginal students. However, this pattern does not follow in using the school RDD. The second panel of the table shows that if anything, reading test scores were increased particularly for students not on the margin of proficiency. The type of variation in pressure captured by the subgroup RDD, with respect to a smaller group of students within a school, may be more likely to be responsive to incentive strength that varies at the student level. However, variation in pressure captured by the school RDD, which may be more of a school-level response, may result in effects that do not favor marginal students in the same way. On the contrary, non-marginal students may benefit more from these school-level responses.

Appendix Tables A13 and A14 show the effects on long-run outcomes for marginal versus non-marginal students. For the subgroup RDD, Appendix Table A13 shows mostly similar effects for both groups of students, with a few exceptions. The large SAT score effects shown in the main results, and the smaller effect on GPA, seem to be more driven by the higher-achieving students, those not on the margin of proficiency. This may be surprising given the larger test score effects for marginal students in the previous table. One possibility is that the extra accountability pressure for students in subgroups of 40 or more led to better short-run effects for lower-achieving students, and better long-run effects for higher-achieving students. For the school RDD, Appendix Table A13 shows a similar pattern of results. Effects are mostly similar between marginal and non-marginal students, with somewhat larger SAT score increases for the non-marginal students. Increased intention to attend a 4-year college may be larger for non-marginal students. All of these results must be interpreted cautiously due to statistical noise.

#### 6.2.5 Effects on Students in Different Subgroups

Lastly, I evaluate effects on test scores and long-run outcomes separately for the following demographic groups: white, black, Hispanic, minority[30], and economically disadvantaged. The effects of accountability on test scores are shown in Appendix Table A15. Here we see that the test score effects in the main results seem to be more driven by effects on white students, relative to minority

[29] As a result, the following results are essentially the same when grouping students into above- and below-median prior test scores.

[30] Defined as black or Hispanic.

students. This difference is more pronounced in the school RDD results than the subgroup RDD results, with no discernible effect on minority students. Appendix Table A16 shows the effects on long-run outcomes for each group, using the subgroup RDD. Here we see larger increases in GPA for white students. SAT score score increases are larger for minority and disadvantaged students. These results also point toward a shift from intending to intend a 2-year college to not intending to attend college for black students. In line with test score effects for school RDD being driven by effects on white students, Appendix Table A17 shows that the long-run effects shown in the main analysis in Table 4 are largely driven by effects on white students as well. Specifically, the large increase in SAT scores, and the shift from 2-year to 4-year colleges, is concentrated in white students. Taken together, these results may raise some questions about the efficacy of NCLB for minorities. Given the focus of the policy, which was designed to prevent schools from "leaving behind" students who were often neglected, including economically disadvantaged students, racial minorities, and lower-achieving students, my analysis of heterogeneous effects casts some doubt on the success of the policy in this regard.

### 6.3 Attrition Bias

For a variety of reasons mentioned previously, the NCERDC data do not allow matching high school outcomes to every elementary school student[31]. This presents the concern of attrition bias, in two primary ways. First, if certain types of students are more likely to be missing in the long-run outcome data, and these students' long-run outcomes were differentially affected by accountability pressure, then the estimates of long-run effects will be biased. Second, if there is selective attrition that makes the average long-run outcomes of students missing in the long-run outcome data different from those in the data, and there are more students missing from the data on one side of the regression discontinuity than the other, then the RD estimates will be biased. Around 30 percent of students in the sample take the SAT[32].

To address this potential bias, I run the same subgroup and school RD specifications in Equations

[31] One reason for missing long-run data would be if a student moves out of state or graduates from a private high school. However, the most likely explanation is that the identifying variables used by the NCERDC to match students across years contained missing values, errors, or variations that prevented matching. Examples include different spellings of names, nick names, school systems assigning their own SSNs, and parents not reporting SSNs. The NCERDC accounts for many sensible issues in their matching algorithm, but not all students could be reliably matched. For more information see the NCERDC report on creating a longitudinal student dataset: https://childandfamilypolicy.duke.edu/wp-content/uploads/2013/10/TECHREPT2.pdf.

[32] For high school GPA, a little over half of the elementary students are included. About 65 percent have data for intention to attend a 4-year college. Roughly 80 percent have data on dropping out or graduating. All are included for SAT-taking, which is an indicator variable for having an SAT score in the data.

?? and 2, with the outcome being an indicator for the student missing a given long-run outcome variable. Results are shown in Table A18. Students in a subgroup just over the cutoff of 40 students are 1 percentage point less likely to be missing data on whether or not the student drops out or graduates from high school. However, this effect is only marginally significant, and small in magnitude, and for each of the other end-of-high-school variables, there is essentially no difference between students on one side of the cutoff versus the other. Using the school RDD, there is not significant change in probability of missing long run data for schools that barely fail AYP in the prior year relative to those that barely pass.

## 6.4 Analysis of Mechanisms

Educators may respond to accountability pressure on test score proficiency in many ways. Principals may strategically assign the best or most experienced teachers to classrooms with more marginal students in order to increase proficiency counts or meet an AYP objective for a marginal subgroup. Principals and administrators may change their hiring or retention practices. They may strategically assign students and teachers to classrooms to maximize the learning of tested material, leveraging student-teacher match effects, peer effects, or specific teachers who are willing and able to focus on improving test scores. Principals and other administrators may encourage teachers to adjust their teaching in order to improve test scores, or give guidance on which students should especially be targeted in order to increase proficiency counts. Teachers may decide to put extra effort toward improving their students' test scores, either by increasing effort in general to increase learning, or by shifting effort toward teaching tested material or test-taking strategies. This potential incentive response of teachers, either aided by exhortations of the principal or as an individual teacher response to the threats facing the school they work in, is discussed in Section 2. However, given the plausibility of a number of other potential mechanisms for the effects captured in the RDD results, this section uses a variety of measures available in the data to evaluate the empirical relevance of these alternative mechanisms.

Using the same RD specifications outlined in Equations ?? and 2 used in the main analysis, I search for discontinuities in several measures of inputs that may be used by schools to generate the positive test score effects seen in Table 2: teacher value-added, teacher experience, the prevalence of transferred teachers, and class size. First, I estimate each teacher's test score value-added for both math and reading, using data from 1997 to 2012. This is done by regressing each student's

test score on the set of student controls used in the main analysis, which include the student's test scores in the prior year, and estimating a fixed-effect for each teacher:

$$y_{ijt} = \mu_j + \beta Z_{it} + \epsilon_{ijt}, \tag{5}$$

Where $y_{ijt}$ is the test score of student $i$ who has teacher $j$ in year $t$, $\mu_j$ is the teacher fixed-effect, and $Z_{it}$ is the full set of controls used in the main analysis. After estimating each teacher's fixed-effect (value-added) in this manner, I test for whether or not students in a subgroup numbering just above the cutoff are assigned better teachers relative to those in a subgroup just below the cutoff. The results are in the first four columns of Appendix Table A19 for math value-added and reading value-added. Results are insignificant for the subgroup RDD, with point estimates suggesting that having a higher value-added teacher results in around 0.01 SD higher test scores for the students subjected to accountability pressure on the right side of the discontinuity. For the school RDD, there does seem to be a shift toward better teachers for schools barely failing AYP in the previous year, with around 0.05 SD higher value added in these schools. This may be due to the fact that this strategy captures school-level variation, whereby schools under pressure implement strategies for improvement across the whole school.

Second, I assess whether or not students just above the cutoff are assigned teachers with more years of experience. The coefficient is very close to zero and insignificant for both RDDs, though slightly bigger for the school RDD. Combined with estimates of the value-added returns to experience (for example, Wiswall (2013)), we can safely rule out the hypothesis that treated students being assigned more experienced teachers plays a meaningful role in the effects of accountability pressure on test scores.

Third, I use an indicator variable for the teacher having transferred into the school immediately preceding the year they teach a student, to evaluate whether more new teachers are being brought in to teach the students under accountability pressure. The estimates are near zero and insignificant for the subgroup RDD, and small, negative, and insignificant for the school RDD. The point estimate suggests that treated students are 0.2 pp less likely to be taught by a transferred teacher according to the subgroup RDD, and roughly 2.5 pp less likely to be taught by a transferred teacher according to the school RDD. These results indicate that increasing the teacher turnover rate is unlikely to be an important mechanism for the main effects. For the school RDD, given the value-added results,

it may be the case that schools try harder to keep good teachers.

Lastly, I assess whether schools respond to the accountability pressure on students in a subgroup with 40 or more students by placing these students in smaller classes. The estimates suggest that this is not the case. The coefficient is statistically insignificant, with the point estimate suggesting that, if anything, treated students are placed in very slightly smaller classes, around 0.08 fewer students per class for the subgroup RDD and around 0.46 fewer students per class for the school RDD.

Taken together, these results suggest that these alternative mechanisms to teacher effort, such as treated students being assigned better teachers, more experienced teachers, more or fewer transferred teachers, or smaller classrooms, do not play much of a role in explaining the effect of accountability pressure on student test scores or long-run outcomes as measured by the subgroup RDD. For the school RDD, the variation in pressure captured at the school level may have resulted in schools working to retain better teachers. However, the other alternative mechanisms are unlikely to have played a major role.

# 7 Conclusion

I use two complementary RD strategies to estimate the effect of subgroup accountability under NCLB on elementary student outcomes and long-run outcomes captured at the end of high school. The first captures across-school variation in pressure specific to a subgroup of students, since subgroups numbering fewer than 40 were excluded from determination of AYP. The second captures across-school variation in pressure relevant for the whole school, since schools barely failing AYP for the first time had strong incentives to improve and avoid the sanctions accompanying another failiure. I find a positive effect on math and reading scores in elementary school, which in the long-run leads to a large effect on SAT scores in high school. I find little evidence of improvement in other long-run outcomes, like high school graduation or intention to attend college. However, I find that the increased pressure associated with a subgroup of students just exceeding 40 resulted in a small positive effect on high school GPA, and the increased pressure associated with AYP failure led students to switch from 2-year to 4-year colleges.

The increase in elementary test scores and SAT scores later in high school may be explained by

a persistent increase in students' skills as captured by standardized tests, which is evidenced by a positive effect on students' test scores in subsequent grades. I also find some evidence that students under accountability pressure had higher GPA in math and reading-related classes in 9th grade, an effect which persists but diminishes through 10th, 11th, and 12th grade. This suggests that the high-stakes placed on these students' test scores may have led schools and teachers to improve the math and reading skills of these students that were captured by the tests, resulting in a persistent increase in this set of skills. Given the small magnitude of other long-run effects relative to the improvement in SAT scores, this lends some support to the hypothesis that NCLB resulted in an increase focus on test scores that did not translate to better long-run outcomes for students. However, given the modest size of the test score effects and the general relationship between students' test scores and long-run outcomes, another possibility is that the long-run effects are commensurate with a small increase in learning captured by test scores but statistically indistinguishable from zero. I analyze this issue in Section 7.1., finding that I can not rule out this possibility.

Several limitations necessitate caveats in the interpretation of these results and future research to supplement this analysis. First, since I use data from North Carolina, it would be useful to evaluate the question of NCLB's long-run effects in other states as well. Second, given the complex nature and broad scope of this nationwide accountability policy, isolating the effect of accountability pressure is difficult. If variation captured by each RD estimate only capture a portion of the effect of accountability on the students at hand, or effects were different for students outside the bandwidth of the RDDs, the overall effect of accountability may be different. Third, since the test score effects are fairly modest, the small long-run effects that would be expected given these effects and the ex ante relationship between test score gains and long-run outcomes may be undetectable. This merits caution against interpreting the near-zero long-run effects as the result of a distortionary shift toward "teaching to the test". Lastly, since this analysis relies on long-run outcomes captured at the end of high school, future research using data on students' income in adulthood would also provide a clearer picture of the long-run effects of No Child Left Behind.

This paper provides mixed evidence for the effectiveness of No Child Left Behind's accountability pressure on students' long-run outcomes. Test scores increased, but perhaps not enough to yield a large impact on student's long-run outcomes. However, students under accountability pressure had higher test scores even three years after the year in which they faced accountability pressure, as well as higher SAT scores in high school, suggesting that some persistent skill improvement did take place. The types of fundamental math and reading skills captured by these exams are important

for students' success in school and the labor market in the long-run, but other important skills that matter for students' long-run outcomes may not have benefited in the same way. Given the immense costs of implementing No Child Left Behind on schools, districts, states, and the federal government, the benefits may not have been large enough to justify such costs. As policymakers and educators grapple with the role of tests and accountability incentives going forward, the results of this paper and future research can help provide insight on the potential benefits and limitations of test-based accountability.

References

Ahn, Thomas and Jacob Vigdor (2014). "The impact of No Child Left Behind's accountability sanctions on school performance: Regression discontinuity evidence from North Carolina." Working Paper 20511, National Bureau of Economic Research, URL http://www.nber.org/papers/w20511.

Chakrabarti, Rajashri (2014). "Incentives and responses under No Child Left Behind: Credible threats and the role of competition." Journal of Public Economics, 110, 124-146.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff (2014). "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood." American Economic Review, 104 (9): 2633-79.

Dee, Thomas S. and Brian Jacob (2011). "The impact of No Child Left Behind on student achievement." Journal of Policy Analysis and Management, 30, 418-446.

Deming, David J., Sarah Cohodes, Jennifer Jennings, and Christopher Jencks (2016). "School accountability, postsecondary attainment and earnings." Review of Economics and Statistics, 98, 848-862.

Farber, Matthew S. (2016), "Essays on the Economics of Education of Underserved Populations." Ph.D. thesis, University of Texas at Austin.

Gilraine, Michael (2018). "School Accountability and the Dynamics of Human Capital Formation."

Holmstrom, Bengt, and Paul Milgrom (1991). "Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design." Journal of Law, Economics, and Organization 7: 24-52.

Jackson, C Kirabo, Rucker C. Johnson, Claudia Persico (2016), "The Effects of School Spending on Educational and Economic Outcomes: Evidence from School Finance Reforms." The Quarterly Journal of Economics, Volume 131, Issue 1, Pages 157–218.

Jacob, Brian A. (2005). "Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago Public Schools." Journal of Public Economics, 89, 761-796.

Kolesár, Michal, and Christoph Rothe. "Inference in regression discontinuity designs with a discrete running variable." American Economic Review 108.8 (2018): 2277-2304.

Krueger, Alan B. (1999), "Experimental estimates of education production functions." Quarterly Journal of Economics, 114, 497–532.

Lee, David S. and David Card (2008), "Regression discontinuity inference with specification error." Journal of Econometrics, 142, 655-674.

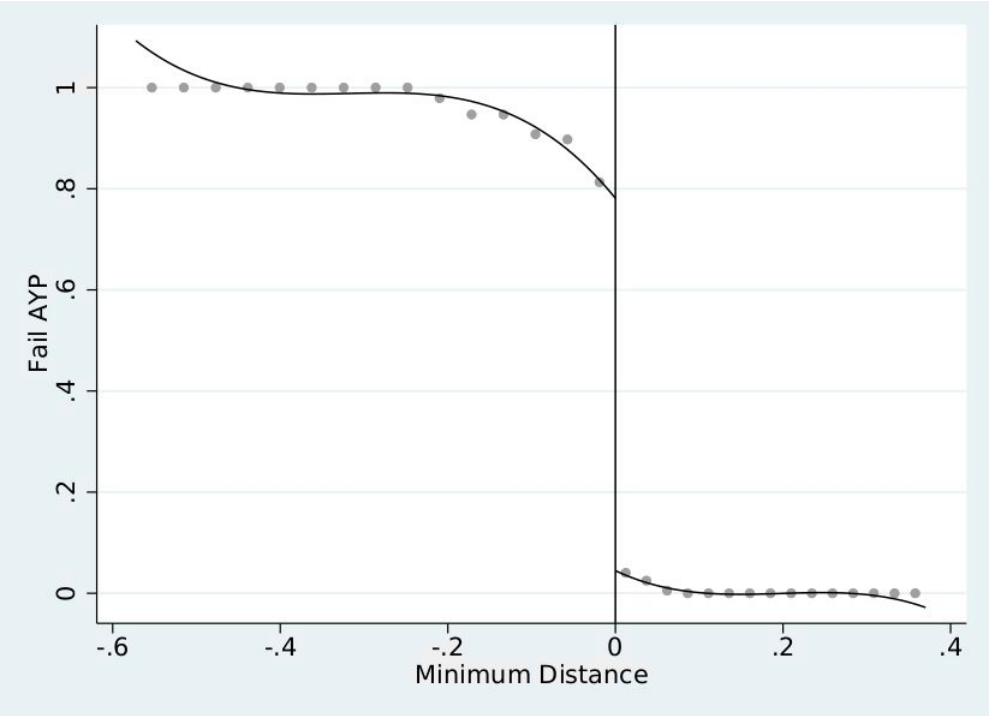
Neal, D. (2011). The design of performance pay in education. In Handbook of the Economics of Education (Vol. 4, pp. 495-550). Elsevier.

Neal, Derek and Diane Whitmore Schanzenbach (2010). "Left behind by design: Proficiency counts and test-based accountability." Review of Economics and Statistics, 92, 263-283.

Ost, B. (2014). "How do teachers improve? The relative importance of specific and general human capital." American Economic Journal: Applied Economics, 6(2), 127-51.

Wiswall, Matthew. "The dynamics of teacher quality." Journal of Public Economics 100 (2013): 61-78.

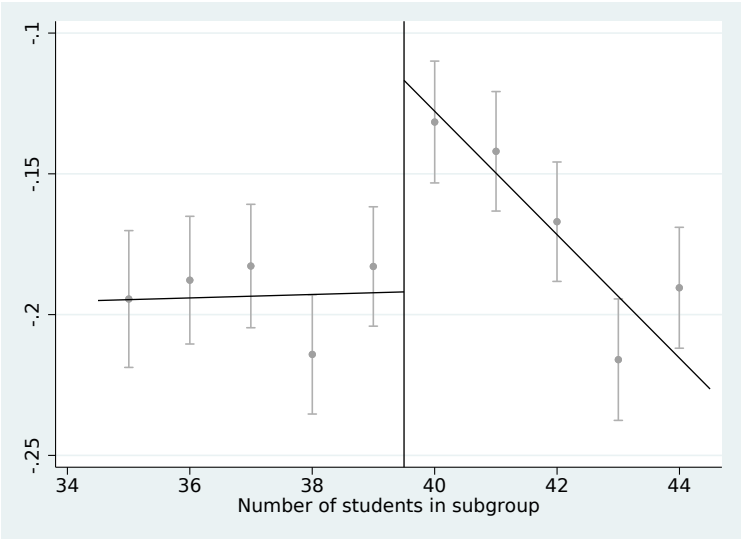
Figure 1: School RDD: Minimum Distance and AYP Failure



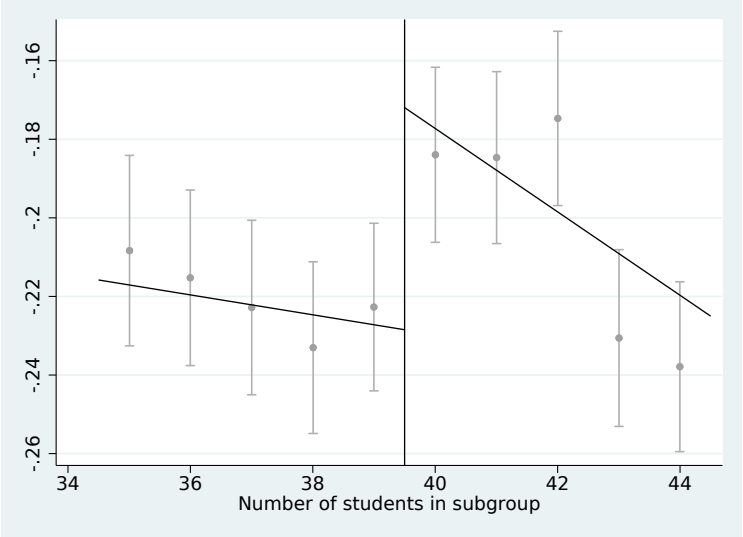

*Notes:* This figure plots the binned sample means of the probability that a school actually was deemed to fail AYP, conditional on the calculated minimum distance to the AYP cutoff. Minimum distance is based on the fraction of students reaching proficiency on the AYP criterion a school performed worst on relative to the fraction of proficiency required for that AYP criterion. The lines above show a third-order polynomial fitted to the binned sample means. The discontinuity in probability of AYP failure at the calculated AYP cutoff is 0.75, suggesting that my calculations replicate AYP determination fairly accurately.

Figure 2: Effect of Accountability Pressure on Test Scores: Subgroup RDD

(a) Math Test Scores



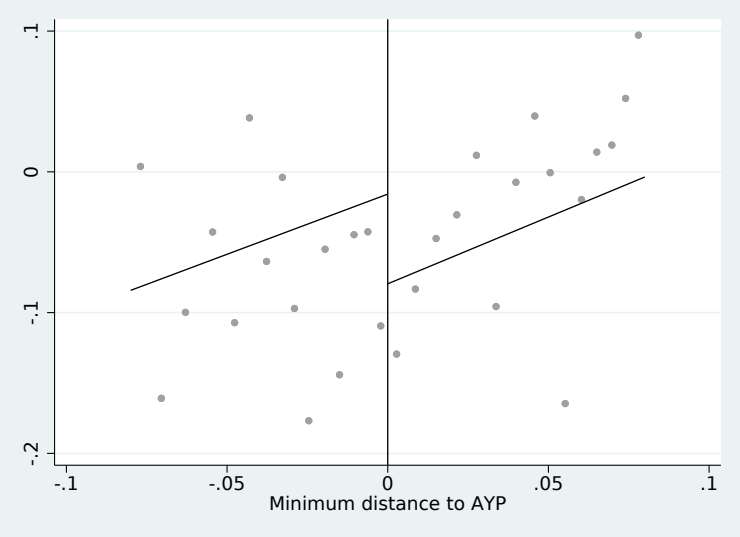

(b) Reading Test Scores



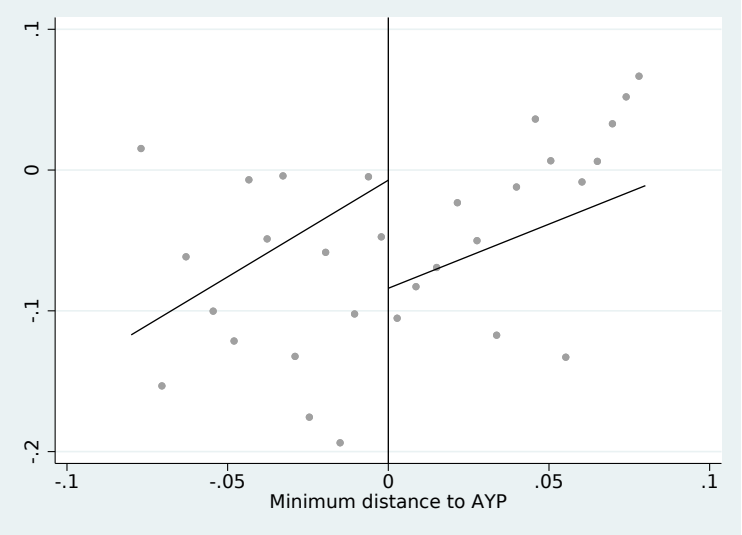

*Notes:* Based on 56,242 observations at the student-year level. Bandwidth is 5. Each plot shows the sample average within each value of the running variable, with a 90% confidence interval. The solid line shows a linear trend fit to these binned averages, with a separate trend on either side of the discontinuity.

Figure 3: Effect of Accountability Pressure on Test Scores: School RDD

(a) Math Test Scores



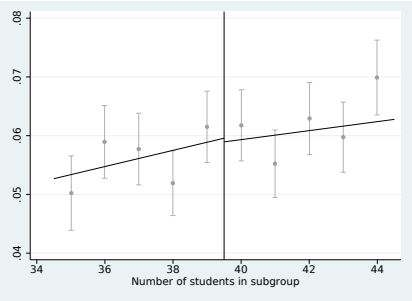

(b) Reading Test Scores



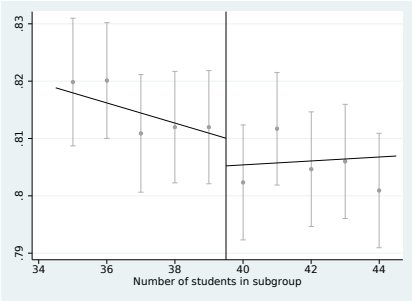

*Notes:* Based on 56,242 observations at the student-year level. Bandwidth is 5. Each plot shows the sample average within each value of the running variable, with a 90% confidence interval. The solid line shows a linear trend fit to these binned averages, with a separate trend on either side of the discontinuity.

Figure 4: Effects on Long-run Outcomes: Subgroup RDD

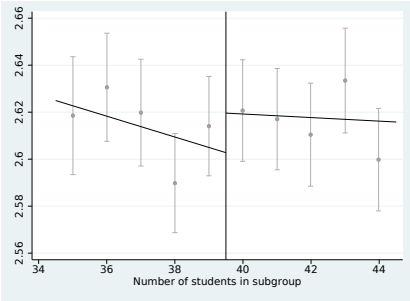
(a) Dropout



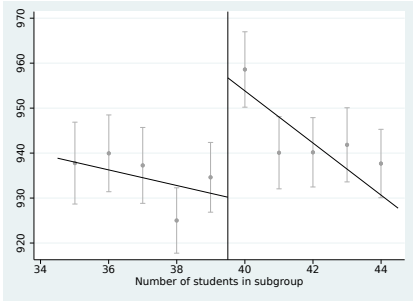

(b) Graduate



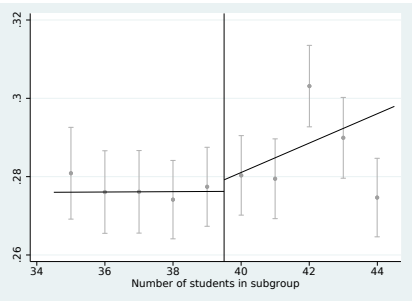

(c) GPA



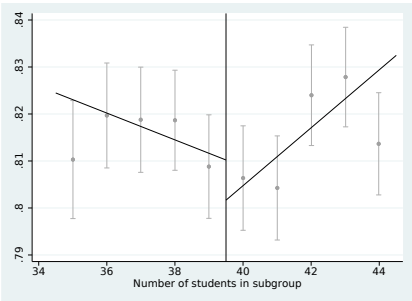

(d) SAT score



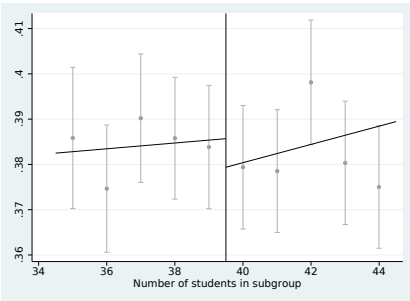

(e) Take SAT



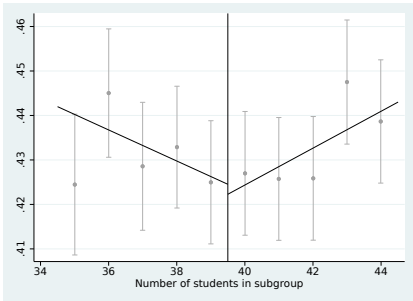

(f) Intend College




(g) Intend 4-year



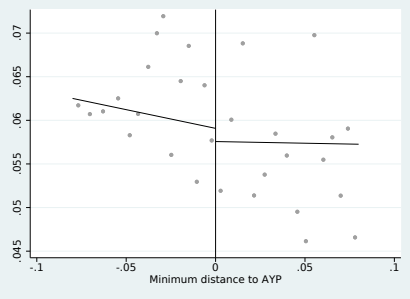

(h) Intend 2-year



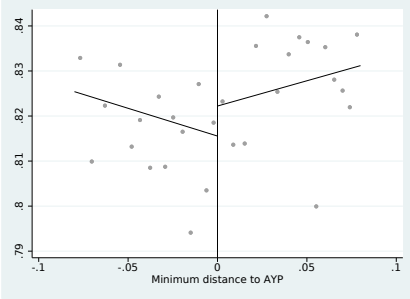

*Notes:* Bandwidth is 5. Each plot shows the sample average within each value of the running variable, with a 90% confidence interval. The solid line shows a linear trend fit to these binned averages, with a separate trend on either side of the discontinuity.

Figure 5: Effects on Long-run Outcomes: School RDD

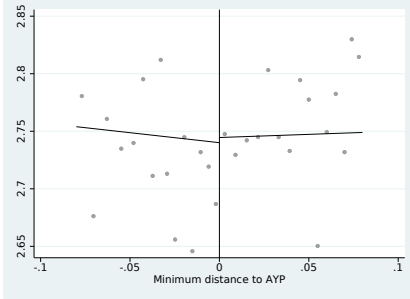
(a) Dropout



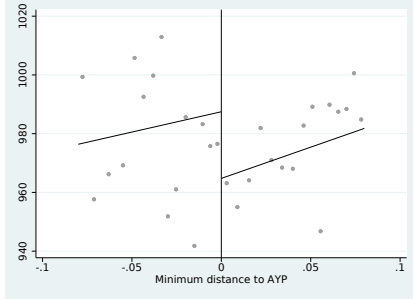

(b) Graduate



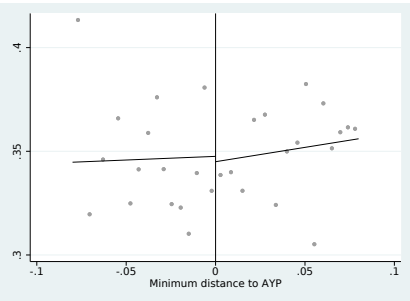

(c) GPA



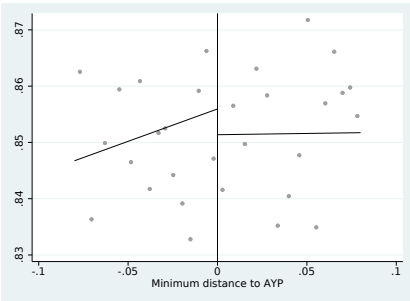

(d) SAT score



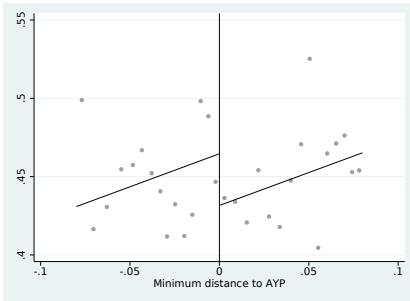

(e) Take SAT



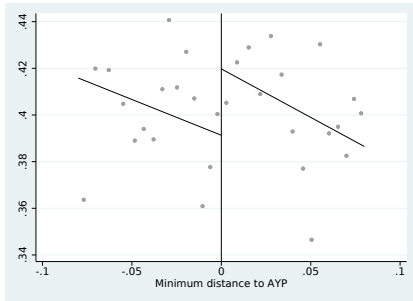

(f) Intend College




(g) Intend 4-year



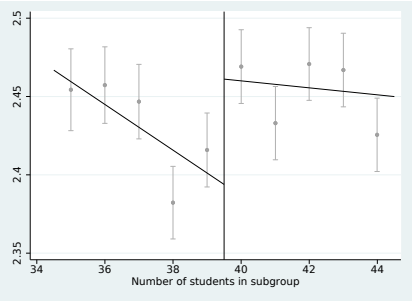

(h) Intend 2-year



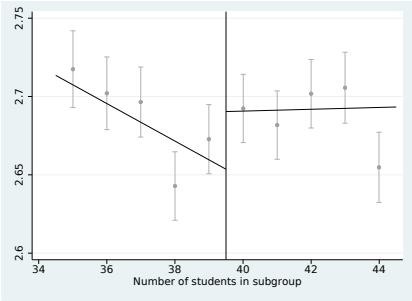

*Notes:* Bandwidth is MSE-optimal (around 0.08). Each plot shows the sample average within each value of the running variable, with a 90% confidence interval. The solid line shows a linear trend fit to these binned averages, with a separate trend on either side of the discontinuity.

Figure 6: Effects on Subject-Specific GPA: Subgroup RDD

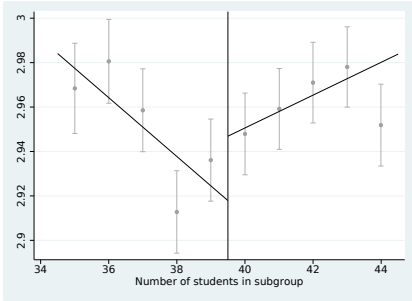
(a) Math



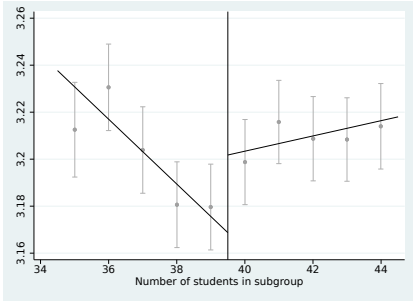

(b) Reading



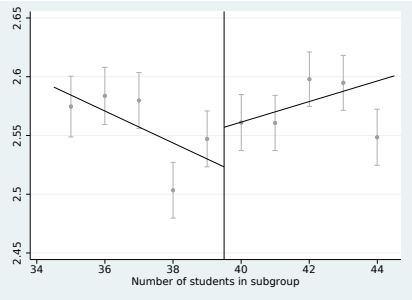

(c) Non Math/Reading



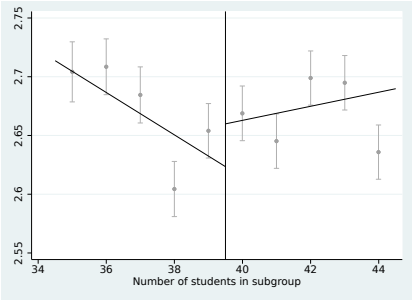

(d) Non Core Subjects



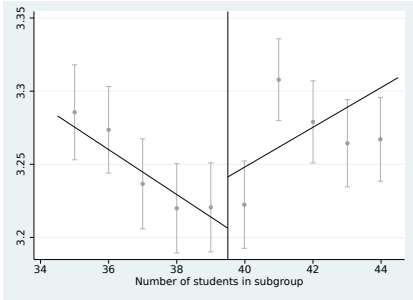

(e) Science




(f) Social Studies



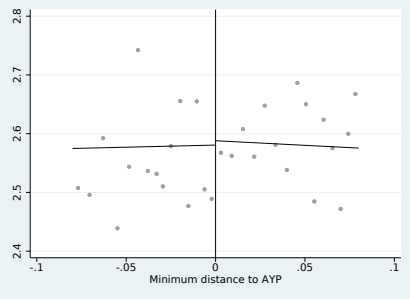

(g) Arts



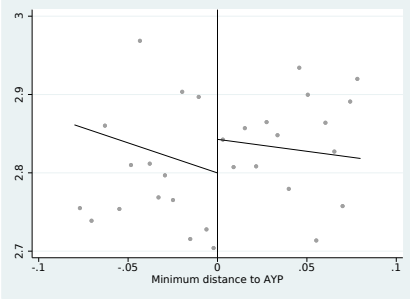

*Notes:* Bandwidth is 5. Each plot shows the sample average within each value of the running variable, with a 90% confidence interval. The solid line shows a linear trend fit to these binned averages, with a separate trend on either side of the discontinuity.

Figure 7: Effects on Subject-Specific GPA: School RDD

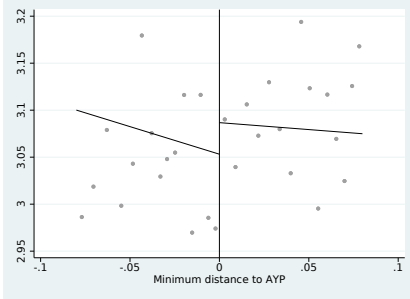
(a) Math



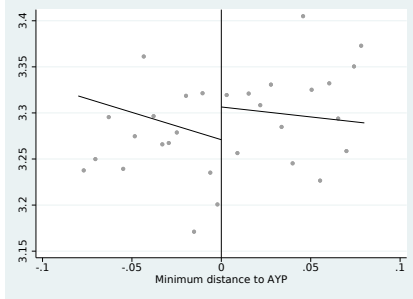

(b) Reading



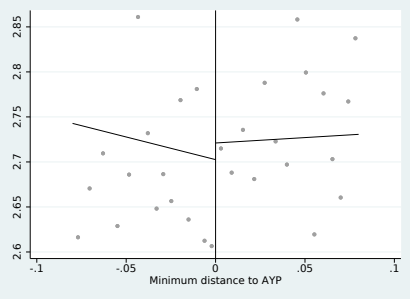

(c) Non Math/Reading



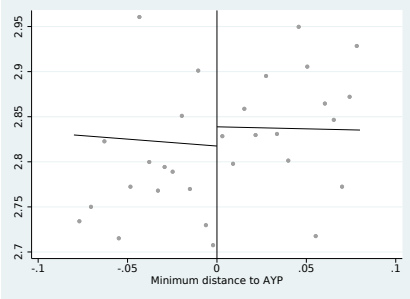

(d) Non Core Subjects



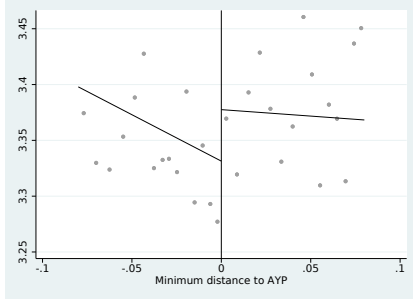

(e) Science




(f) Social Studies




(g) Arts

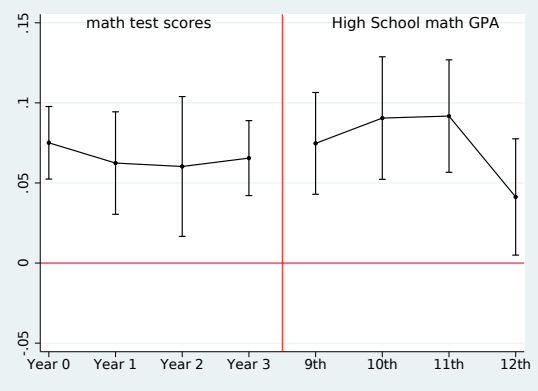



*Notes:* Bandwidth is MSE-optimal. Each plot shows the sample average within each value of the running variable, with a 90% confidence interval. The solid line shows a linear trend fit to these binned averages, with a separate trend on either side of the discontinuity.

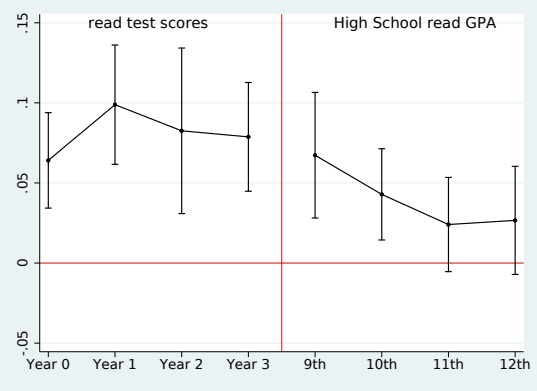
Figure 8: Effects of Accountability Pressure on Future Achievement

*Subgroup RDD*

(a) Math

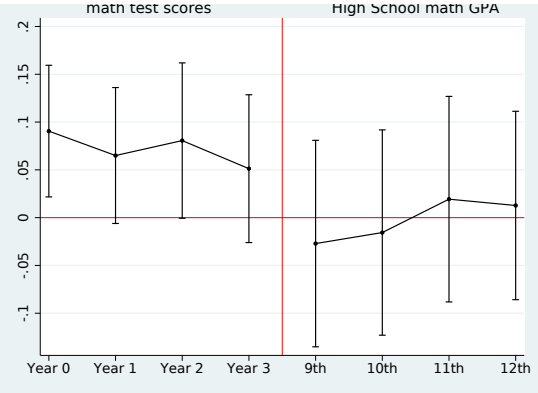



(b) Reading

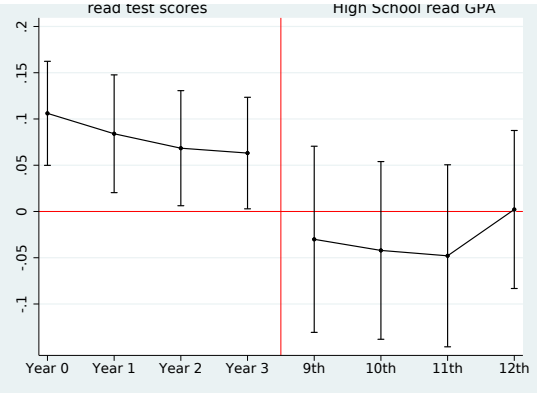



*School RDD*

(c) Math




(d) Reading




*Notes:* Each point in the figures above shows the point estimate and 90% confidence interval for the RD estimate of the outcome variables listed. Subfigures (a) and (c) show effects on math test scores in the year of accountability pressure in elementary school and the subsequent three years (into middle school), and then GPA in math classes in each year of high school, for the subgroup RDD and school RDD, respectively. Subfigures (b) and (d) show effects on reading test scores in the year of accountability pressure in elementary school and the subsequent three years (into middle school), and then GPA in reading classes in each year of high school, for the subgroup RDD and school RDD, respectively. Test score effects are in terms of fraction of a standard deviation in the student test score distribution. GPA effects are in terms of GPA (4 point scale). All student controls included in column 2 of Table 2 are included in these regressions.

Table 1: Summary Statistics

| | **Full Sample** | **Subgroup RD Sample** | **School RD Sample** |
|---|---|---|---|
| Math score | 0.021 | -0.183 | -0.040 |
| Reading score | 0.014 | -0.213 | -0.046 |
| White | 0.563 | 0.297 | 0.526 |
| Black | 0.274 | 0.378 | 0.317 |
| Hispanic | 0.084 | 0.237 | 0.088 |
| Asian | 0.021 | 0.042 | 0.020 |
| Economically Disadvantaged | 0.553 | 0.828 | 0.586 |
| English Language Learner | 0.053 | 0.143 | 0.057 |
| Female | 0.494 | 0.495 | 0.493 |
| SAT-taker | 0.345 | 0.281 | 0.348 |
| SAT score | 993.9 | 938.1 | 976.8 |
| High School GPA | 2.788 | 2.619 | 2.746 |
| Dropout | 0.049 | 0.060 | 0.058 |
| Graduate | 0.833 | 0.809 | 0.824 |
| Intend 4-year college | 0.453 | 0.380 | 0.448 |
| Intend any college | 0.852 | 0.816 | 0.852 |
| School size | 309.7 | 244.8 | 272.8 |
| Student-year observations | 1,304,301 | 56,246 | 184,004 |
| School-year observations | 7,714 | 1,349 | 1,135 |

*Notes:*This table shows the means of key variables used in the analysis. Math and reading scores are standardized within grade-year. Intention to attend 4-year or any college is only included for students graduating high school.

Table 2: Effect of Accountability Pressure on Test Scores

*Subgroup RDD*

| | Math | | Reading | |
|---|---|---|---|---|
| Treated ($\tau_{sgt}$) | 0.0814*** | 0.0615*** | 0.0625*** | 0.0389*** |
| | (0.0158) | (0.0110) | (0.0180) | (0.00799) |
| Observations | 50,702 | 50,702 | 50,814 | 50,814 |

*School RDD*

| | Math | | Reading | |
|---|---|---|---|---|
| Fail ($F_{s,t-1}$) | 0.0907** | 0.0493** | 0.107*** | 0.0424*** |
| | (0.0416) | (0.0247) | (0.0342) | (0.0162) |
| Observations | 165,859 | 199,679 | 175,057 | 238,880 |
| Year FE | YES | YES | YES | YES |
| Subgroup FE | YES | YES | YES | YES |
| Student controls | NO | YES | NO | YES |

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1
Student controls include lagged test scores, gender,
limited English proficiency, and grade FE.
The bandwidth is 5 for the subgroup RDD
and MSE optimal (around 0.08) for the school RDD.
Standard errors are clustered at the subgroup-by-count level
for the subgroup RDD and at the school level for the school RDD.

Table 3: Effect of Accountability Pressure on Long-Run Outcomes - Subgroup RDD

| | Dropout | | Graduate | | HS GPA | | SAT score | |
|---|---|---|---|---|---|---|---|---|
| Treated | -0.00132 | -0.00103 | -0.0053 | -0.00522 | 0.0308*** | 0.0224** | 19.67** | 17.99*** |
| | (0.00521) | (0.00509) | (0.00559) | (0.00538) | (0.0113) | (0.0104) | (8.508) | (6.212) |
| Mean Dep. Var. | 0.049 | | 0.833 | | 2.788 | | 993.9 | |
| SD Dep. Var. | 0.217 | | 0.373 | | 0.729 | | 192.4 | |
| Observations | 41,055 | 41,055 | 41,055 | 41,055 | 27,308 | 27,308 | 14,295 | 14,295 |

| | SAT-taking | | Intend college | | Intend 4-year | | Intend 2-year | |
|---|---|---|---|---|---|---|---|---|
| Treated | 0.00591 | 0.00341 | -0.00594 | -0.00711 | 0.00328 | -0.00233 | -0.00921 | -0.00478 |
| | (0.00851) | (0.00844) | (0.00552) | (0.00551) | (0.00948) | (0.00706) | (0.00996) | (0.00849) |
| Mean Dep. Var. | 0.345 | | 0.852 | | 0.453 | | 0.399 | |
| SD Dep. Var. | 0.475 | | 0.355 | | 0.498 | | 0.490 | |
| Observations | 50,747 | 50,747 | 33,230 | 33,230 | 33,230 | 33,230 | 33,230 | 33,230 |
| Year FE | YES | YES | YES | YES | YES | YES | YES | YES |
| Subgroup FE | YES | YES | YES | YES | YES | YES | YES | YES |
| Student controls | NO | YES | NO | YES | NO | YES | NO | YES |

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1
Student controls include lagged test scores, gender,
limited English proficiency, and grade FE.
The bandwidth is 5.
Standard errors are clustered at the subgroup-by-count level.

Table 4: Effect of Accountability Pressure on Long-Run Outcomes - School RDD

| | Dropout | | Graduate | | HS GPA | | SAT score | |
|---|---|---|---|---|---|---|---|---|
| Treated | 0.00369 | 0.00454 | -0.0111 | -0.0134* | 0.0043 | -0.02 | 30.59** | 20.69* |
| | (0.00388) | (0.00377) | (0.00787) | (0.00776) | (0.0271) | (0.0297) | (13.38) | (10.8) |
| Mean Dep. Var. | 0.049 | | 0.833 | | 2.788 | | 993.9 | |
| SD Dep. Var. | 0.217 | | 0.373 | | 0.729 | | 192.4 | |
| Observations | 196,179 | 199,261 | 202,922 | 202,922 | 136,637 | 123,121 | 44,458 | 41,999 |
| | SAT-taking | | Intend college | | Intend 4-year | | Intend 2-year | |
| Treated | 0.00267 | -0.0083 | 0.00491 | 0.00167 | 0.0490** | 0.0374* | -0.0428* | -0.0342* |
| | (0.017) | (0.0137) | (0.00894) | (0.00844) | (0.0248) | (0.0211) | (0.0219) | (0.0198) |
| Mean Dep. Var. | 0.345 | | 0.852 | | 0.453 | | 0.399 | |
| SD Dep. Var. | 0.475 | | 0.355 | | 0.498 | | 0.490 | |
| Observations | 183,451 | 213,279 | 164,411 | 164,891 | 87,595 | 87,797 | 93,583 | 93,583 |
| Year FE | YES | YES | YES | YES | YES | YES | YES | YES |
| Subgroup FE | YES | YES | YES | YES | YES | YES | YES | YES |
| Student controls | NO | YES | NO | YES | NO | YES | NO | YES |

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1
Student controls include lagged test scores, gender,
limited English proficiency, and grade FE.
The MSE-optimal bandwidth is used, which is approximately 0.08 (fraction of students proficient).
Standard errors are clustered at the school level.

Table 5: Effects of Accountability on Subject-specific HS GPA - Subgroup RDD

| | Math GPA | | Reading GPA | | Non Math/Reading GPA | | Non Core GPA | |
|---|---|---|---|---|---|---|---|---|
| Treated | 0.0798*** | 0.0754*** | 0.0449*** | 0.0419*** | 0.0461*** | 0.0419*** | 0.0484*** | 0.0449*** |
| | (0.0179) | (0.0175) | (0.0152) | (0.0144) | (0.0126) | (0.0116) | (0.0143) | (0.0132) |
| Observations | 32,046 | 32,046 | 32,070 | 32,070 | 32,273 | 32,273 | 32,264 | 32,264 |
| | Science GPA | | Social Studies GPA | | Arts GPA | | | |
| Treated | 0.0510*** | 0.0445*** | 0.0545*** | 0.0489** | 0.0462* | 0.0436* | | |
| | (0.0169) | (0.0159) | (0.0193) | (0.0197) | (0.0257) | (0.025) | | |
| Observations | 32,015 | 32,015 | 32,029 | 32,029 | 23,050 | 23,050 | | |
| Year FE | YES | YES | YES | YES | YES | YES | YES | YES |
| Subgroup FE | YES | YES | YES | YES | YES | YES | YES | YES |
| Student controls | NO | YES | NO | YES | NO | YES | NO | YES |

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1
Student controls include lagged test scores, gender,
limited English proficiency, and grade FE.
Non Core GPA includes all courses outside of math, language arts, science, and social studies.
The bandwidth is 5.
Standard errors are clustered at the subgroup-by-count level.

Table 6: Effects of Accountability on Subject-specific HS GPA - School RDD

| | Math GPA | | Reading GPA | | Non Math/Reading | | Non Core | |
|---|---|---|---|---|---|---|---|---|
| Treated | 0.00114 | -0.0127 | -0.0288 | -0.0437 | -0.0228 | -0.0337 | -0.0269 | -0.0393 |
| | (0.0586) | (0.0587) | (0.0524) | (0.0529) | (0.0435) | (0.0439) | (0.0379) | (0.0386) |
| Observations | 153,865 | 160,147 | 159,987 | 163,260 | 158,686 | 163,620 | 161,559 | 158,668 |
| | Science GPA | | Social Studies GPA | | Arts GPA | | | |
| Treated | -0.00797 | -0.0241 | -0.0114 | -0.0274 | -0.0413 | -0.0542 | | |
| | (0.0561) | (0.0573) | (0.0524) | (0.0532) | (0.0382) | (0.0384) | | |
| Observations | 155,574 | 160,046 | 154,667 | 160,200 | 119,103 | 118,321 | | |
| Year FE | YES | YES | YES | YES | YES | YES | YES | YES |
| Subgroup FE | YES | YES | YES | YES | YES | YES | YES | YES |
| Student controls | NO | YES | NO | YES | NO | YES | NO | YES |

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1
Student controls include lagged test scores, gender,
limited English proficiency, and grade FE.
Non Core GPA includes all courses outside of math, language arts, science, and social studies.
The MSE-optimal bandwidth is used.
Standard errors are clustered at the school level.

Table 7: Effects of Accountability on Test Scores in Subsequent Years

| | Same year | 1 year after | 2 years after | 3 years after |
|---|---|---|---|---|
| *Subgroup RDD* | | | | |
| Math | 0.0751*** | 0.0624*** | 0.0603** | 0.0655*** |
| | (0.0138) | (0.0194) | (0.0265) | (0.0142) |
| Reading | 0.0641*** | 0.0989*** | 0.0826*** | 0.0788*** |
| | (0.0181) | (0.0226) | (0.0314) | (0.0206) |
| Observations | 55,772 | 37,708 | 39,041 | 45,755 |
| *School RDD* | | | | |
| Math | 0.0905** | 0.065 | 0.0806 | 0.0512 |
| | (0.0419) | (0.0432) | (0.0494) | (0.047) |
| Reading | 0.106*** | 0.0840** | 0.0684* | 0.0632* |
| | (0.0342) | (0.0387) | (0.0378) | (0.0367) |
| Observations | 163,102 | 103,369 | 111,555 | 146,692 |

Standard errors are in parentheses,
clustered at the subgroub-by-count level for subgroup RDD
and the school level for school RDD
*** p<0.01, ** p<0.05, * p<0.1
Each entry in the table is the coefficient from a local linear regression
where the outcome variable is the test score on the left (math or reading)
measured at the time given at the top.
Year and subgroup FE are included in each regression.
The bandwidth is 5 for subgroup RDD, MSE-optimal for school RDD.

Table 8: Effects of Accountability on Math and Reading GPA in Subsequent Years

| | 9th grade | 10th grade | 11th grade | 12th grade |
|---|---|---|---|---|
| *Subgroup RDD* | | | | |
| Math | 0.0747*** | 0.0905*** | 0.0918*** | 0.0413* |
| | (0.0193) | (0.0232) | (0.0213) | (0.0221) |
| Reading | 0.0673*** | 0.0429** | 0.0241 | 0.0266 |
| | (0.0238) | (0.0173) | (0.0179) | (0.0205) |
| Observations | 41,736 | 40,085 | 36,490 | 34,610 |
| *School RDD* | | | | |
| Math | -0.0271 | -0.0156 | 0.0193 | 0.0127 |
| | (0.0657) | (0.0653) | (0.0654) | (0.0599) |
| Reading | -0.0301 | -0.0421 | -0.0479 | 0.00215 |
| | (0.0612) | (0.0584) | (0.0598) | (0.0519) |
| Observations | 195,258 | 178,285 | 162,702 | 153,445 |

Standard errors are in parentheses,
clustered at the subgroub-by-count level for subgroup RDD
and the school level for school RDD
*** p<0.01, ** p<0.05, * p<0.1
Each entry in the table is the coefficient from a local linear regression
where the outcome variable is the GPA for the subject on the left,
measured in the year of high school given at the top.
Year and subgroup FE are included in each regression.
The bandwidth is 5 for subgroup RDD, MSE-optimal for school RDD.

# 1 Appendix

## 1.1 A Theory of Multi-tasking for Test Scores and Long-Run Outcomes

I develop a principal-agent model with multi-tasking, based on the seminal model by Holmstrom and Milgrom (1991) and several applications to educator incentives by Neal (2011), and add several extensions to clarify how accountability incentives targeting student test scores might impact students' long-run outcomes. Assume one teacher is hired to teach one student (or a monolithic classroom). The student's learning outcomes are their test score $P$ and human capital $H$ (or the "true value" of their knowledge). These outcomes are a function of the student's ability level $\alpha$ and the teacher's contribution of effort in two dimensions, $t_1$ and $t_2$. Let the first task teachers spend effort on, $t_1$, stand for teaching deemed "best practices" in the absence of the accountability system; let the second task, $t_2$, stand for teaching aimed at producing a higher test score ("teaching to the test").

The student's test score $P$ and human capital $H$ are then:

$$P = \alpha + g_1 t_1 + g_2 t_2 + \nu, \qquad H = \alpha + f_1 t_1 + f_2 t_2 + \epsilon, \tag{6}$$

where $g_1$ is the test-score return to effort in "best practices" ($t_1$), $g_2$ is the test-score return to effort in "teaching to the test" ($t_2$), $f_1$ is the human capital return to effort for $t_1$, and $f_2$ is the human capital return to effort for $t_2$. It is generally assumed that $g_2 > g_1$ and $f_1 > f_2$, meaning that "teaching to the test" is at least marginally more effective at increasing test scores than "best practices", and "best practices" are at least marginally more effective at increasing human capital. This restriction is helpful for interpretation but not required. One possible exception to this assumption would be if the standardized tests used under accountability are crafted to focus the curriculum on the skills and knowledge that do in fact matter most for human capital, causing teachers to no longer spend time on activities that are low return for both test scores and human capital. $P$ and $H$ also include error components $\nu$ and $\epsilon$, which I assume have unimodal distributions with mean zero. The error component $\nu$ in $P$ has cumulative distribution function $\Phi$ and probability density function $\phi$.

The teacher cares about the student's human capital to the degree $\gamma$, a parameter with a value between 0 and 1. I include this component to account for the fact that teachers likely have an altruistic motive in teaching, and more generally to capture the degree to which teachers choose to focus on "best practices" for any number of reasons present before the accountability system is implemented[33]. The teacher faces a quadratic cost of total effort, $0.5(t_1 + t_2)^2$. There is no scaling parameter in the cost of effort function, as the marginal cost of effort is normalized to total effort level $t_1 + t_2$, and thus the other parameters are scaled in reference to the cost of effort. The teacher is punished with the cost $\pi$ if the student's test score falls below a "proficiency" level $\bar{p}$. Thus, the teacher maximizes the following objective function by choosing $t_1$ and $t_2$:

$$\gamma[f_1 t_1 + f_2 t_2] - \pi[\Phi(\bar{p} - \alpha - g_1 t_1 - g_2 t_2)] - 0.5(t_1 + t_2)^2. \tag{7}$$

The first order conditions for the solution to this maximization problem will set marginal benefit equal to marginal cost for both tasks, $t_1$ and $t_2$:

$$\gamma f_1 + \pi g_1 \phi(\bar{p} - \alpha - g_1 t_1 - g_2 t_2) = t_1 + t_2 \tag{8}$$

[33] Neal (2011) defines the cost of effort based on the gap between effort $t_2$ and an "effort norm" for $t_2$, which is the level of effort the teacher would choose absent an incentive created by the performance pay scheme. My strategy also creates a "default" level of effort in $t_2$ chosen by the teacher absent accountability pressure, but the default is based on the teacher caring directly about the student's human capital and factoring this into their optimization.

$$\gamma f_2 + \pi g_2 \phi(\bar{p} - \alpha - g_1 t_1 - g_2 t_2) = t_1 + t_2 \tag{9}$$

The marginal benefit of increasing $t_1$ is $\gamma f_1$, which is $t_1$'s return to human capital times the degree to which the teacher cares about the student's human capital, plus $\pi g_1 \phi(\bar{p} - \alpha - g_1 t_1 - g_2 t_2)$, which is the severity of the punishment for the student failing to meet proficiency times the reduction in failure probability caused by increasing $t_1$. A similar explanation applies to $t_2$. Since we assume $f_1 > f_2$ and $g_2 > g_1$, the marginal benefit of increasing the student's human capital should be larger for $t_1$ and the marginal benefit of reducing the probability of the student failing proficiency should be larger for $t_2$.

These equations also imply that the marginal benefit of increasing $t_1$ will be set equal to the marginal benefit of increasing $t_2$:

$$\gamma f_1 + \pi g_1 \phi(-) = \gamma f_2 + \pi g_2 \phi(-), \tag{10}$$

which can be rewritten as:

$$\gamma(f_1 - f_2) = \pi \phi(-)(g_2 - g_1). \tag{11}$$

Equation 11 shows that marginal benefit of focusing on $t_1$ ("best practices") rather than $t_2$ is driven by the teacher's value of the student's human capital and the greater human capital returns of $t_1$ relative to $t_2$. Similarly, the marginal benefit of focusing on $t_2$ ("teaching to the test") rather than $t_1$ is driven by the disutility of the punishment, the rate at which the failure probability is reduced, and the greater test-score returns of $t_2$ relative to $t_1$.

To illustrate the model's implications for the effects of school accountability pressure on test scores and long-run outcomes, first consider what happens in the absence of accountability pressure, when $\pi = 0$. In this case, the first order condition in Equation 8 simplifies to $\gamma f_1 = t_1$, and Equation 9 is irrelevant, since $t_2$ will be 0 and the solution is not an interior solution. The teacher has no incentive to spend effort on $t_2$, teaching to the test, because the teacher only cares about the student's human capital, $f_1 > f_2$, and the cost of effort is a function of total effort $t_1 + t_2$.

Now consider what happens under accountability pressure, when $\pi > 0$. The teacher then has an incentive to put effort into $t_2$, teaching to the test, in order to reduce the probability that the student's test score $P$ is below the proficiency score $\bar{p}$. Increasing $t_2$ reduces this probability insofar as $\bar{p} - \alpha$ is close enough to zero that the probability density function is significantly larger than zero. If the student's ability level makes them extremely unlikely to score below proficiency, or extremely unlikely to score above proficiency, teacher-induced test score increases have very low returns, because they will not change the failure probability much or at all. If $\bar{p} - \alpha$ is close enough to zero, then the teacher will put more effort in $t_2$ if the size of additional test-score returns to $t_2$ relative to $t_1$, $g_2 - g_1$, is larger. Lastly, all else equal, the teacher will invest more in $t_2$ if the cost of the punishment $\pi$ is larger. Conversely, the teacher will put more effort in $t_1$ if they care more about the student's human capital (scaled by $\gamma$), and if the additional human capital return to $t_1$ relative to $t_2$, $f_1 > f_2$, is larger.

Based on this model, what can we say about the effect of accountability pressure on test scores and human capital? Given the teacher's incentives as explained above, we see that accountability pressure should increase test scores, to the extent that these conditions are met: changing the student's test score significantly affects their probability of exceeding proficiency, "teaching to the test" has a larger test score return than "best practices", and the cost of punishment due to a test score below proficiency is large. Lastly, since the effort cost depends on total effort, $t_1 + t_2$, the incentive to increase test scores is also stronger if the teacher's incentive to invest in $t_2$ to increase the student's human capital is weaker. The effect of accountability pressure on the student's long-run outcomes (human capital) is more ambiguous. There could be a positive effect if the incentive makes the teacher invest much more in "teaching to the test" ($t_2$) and the human capital returns to doing so are not much lower than the human capital returns to the task they would invest in

without the incentive ($t_1$). There could be a negative effect if the incentive makes the teacher invest more in "teaching to the test" and invest less in "best practices" and "best practices" have a much higher return to human capital. This is more likely to happen if the teacher cares less about the student's human capital, and if the incentive to increase test scores is strong, implying that negative human capital effects should be accompanied by positive test score effects. Of course, all of these factors could result in intermediate outcomes, such that the effect of accountability pressure on human capital is small or near zero.

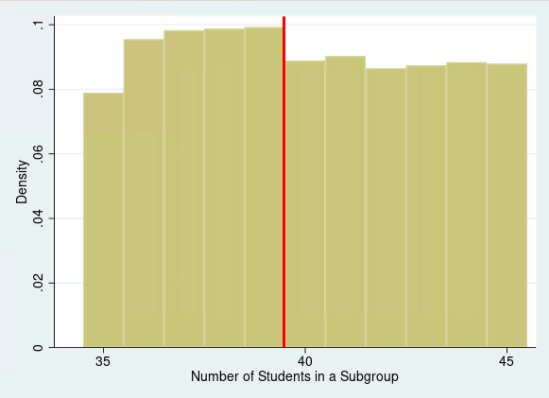
An important consideration beyond the scope of this model is the existence of many heterogeneous teachers allocating effort across many classrooms of heterogeneous students. If the test score and human capital effects vary systematically across students, teachers, and schools based on this heterogeneity and particular features of the accountability system, there is the potential for accountability pressure to substantially alter the relationship between test scores and long-run outcomes. Some students may be unaffected in either dimension, some may have test score increases but no change in long-run outcomes, some may have test score increases and improvements in long-run outcomes, and some may have test score increases but worse long-run outcomes. If teachers substitute time away from some students in the class toward others, there may be students negatively affected in both dimensions.

## 1.2 Appendix Figures

Figure A1: Density of the Running Variable

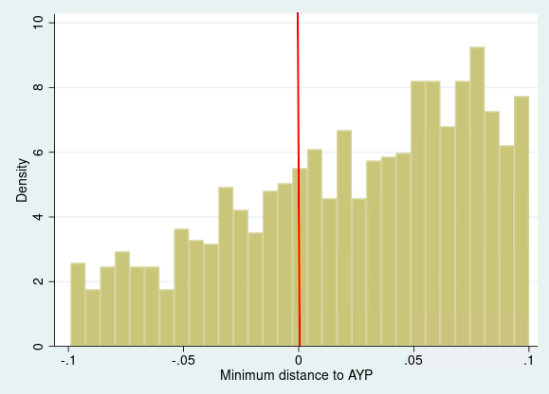
**Subgroup RD**

p-value for manipulation test: 0.83




**School RD**

p-value for manipulation test: 0.42

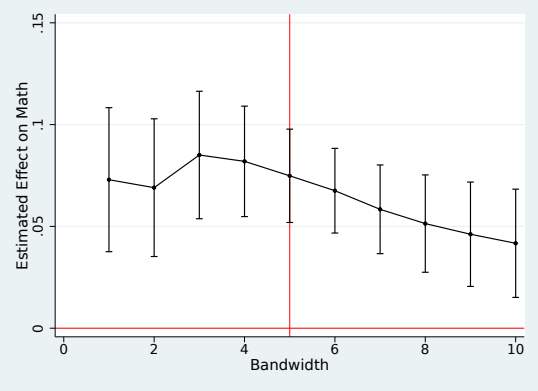



*Notes:* Based on 2,116 observations at the school-subgroup-year level. RD manipulation test at cutoff of 39.5 using local polynomial density estimation, adjusting for mass points in the running variable, is done based on Cattaneo, Jansson, and Ma (2017) using their rddensity package in Stata. Bandwidth is 5.

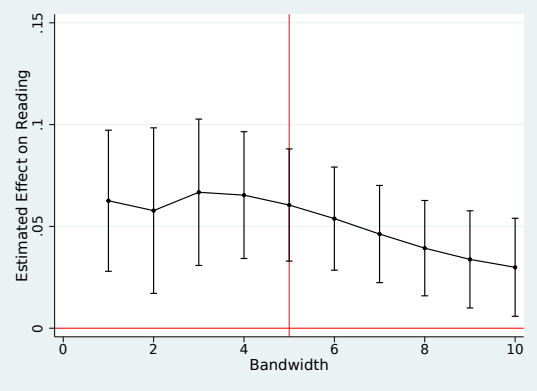
Figure A2: Sensitivity of Test Score Effects to Bandwidth Selection

*Subgroup RDD*

(a) Math

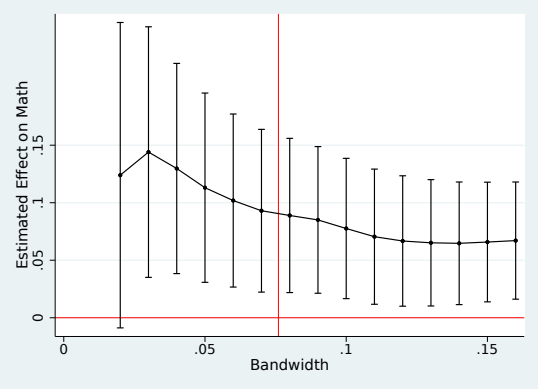



(b) Reading

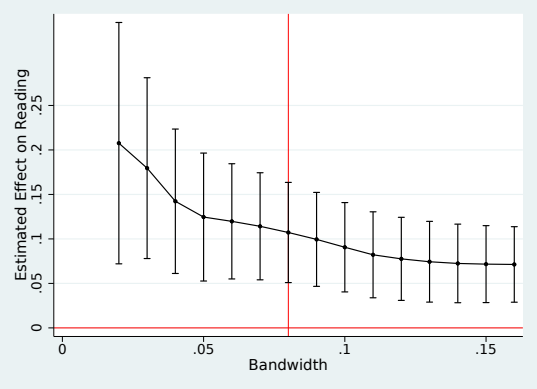



*School RDD*

(c) Math



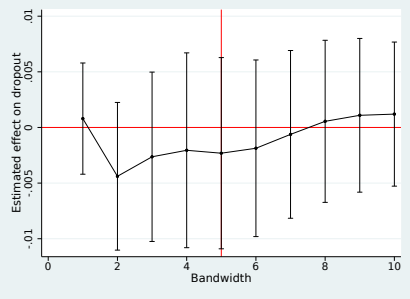

(d) Reading



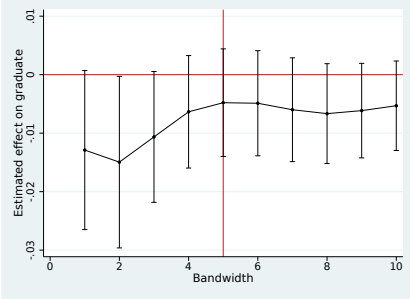

*Notes:* Each point in the figures above shows the point estimate and 90% confidence interval for the RD estimate using the given bandwidth on the x-axis. Each RD specification includes subgroup and year fixed-effects.

Figure A3: Sensitivity of Long-run Effects to Bandwidth Selection: Subgroup RDD

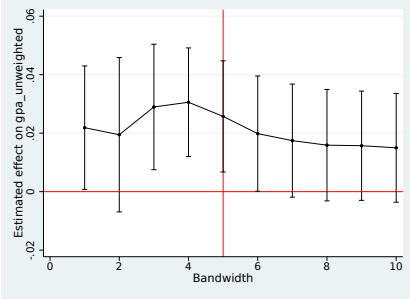
(a) Dropout



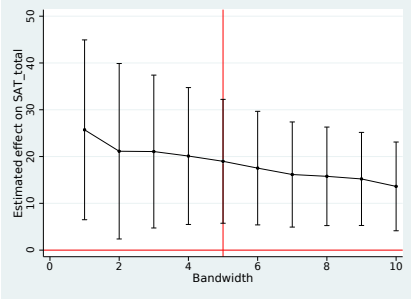

(b) Graduate



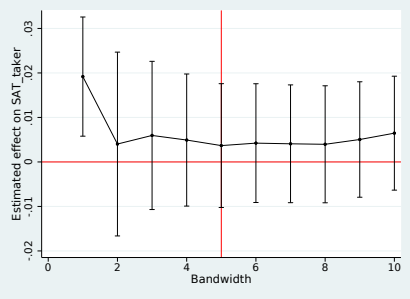

(c) GPA



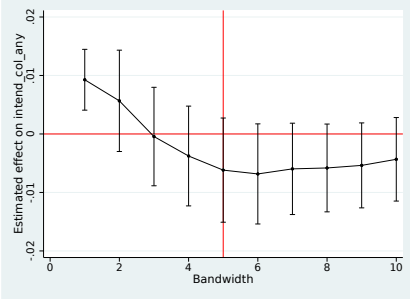

(d) SAT score



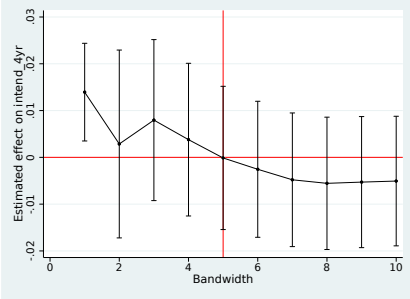

(e) Take SAT



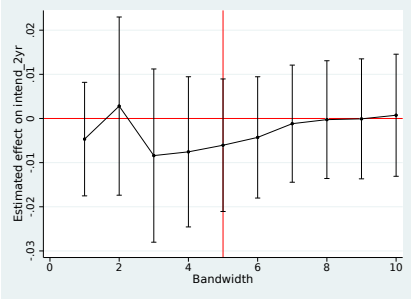

(f) Intend College




(g) Intend 4-year



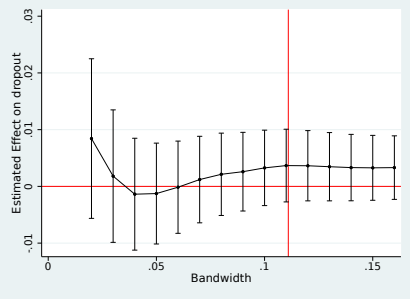

(h) Intend 2-year



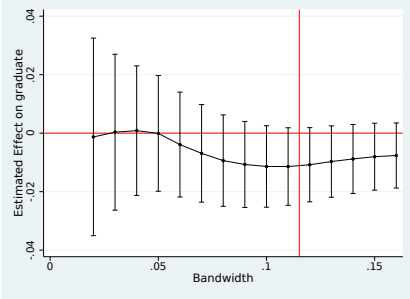

*Notes:* Each point in the figures above shows the point estimate and 90% confidence interval for the RD estimate using the given bandwidth on the x-axis. The vertical red line indicates the chosen bandwidth for the main analysis. Each RD specification includes subgroup and year fixed-effects.

Figure A4: Sensitivity of Long-run Effects to Bandwidth Selection: School RDD

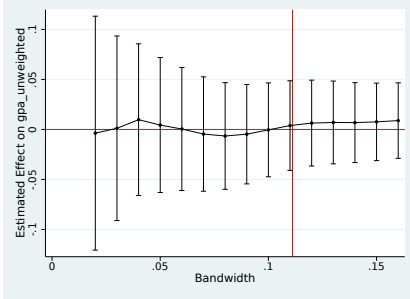
(a) Dropout



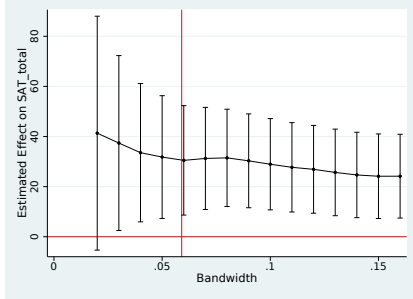

(b) Graduate



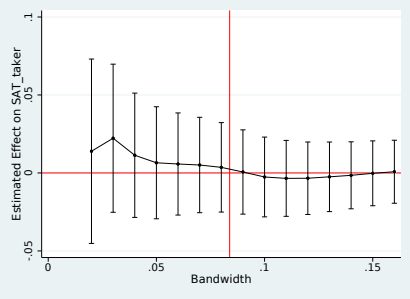

(c) GPA



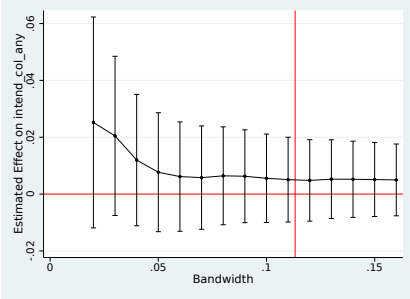

(d) SAT score



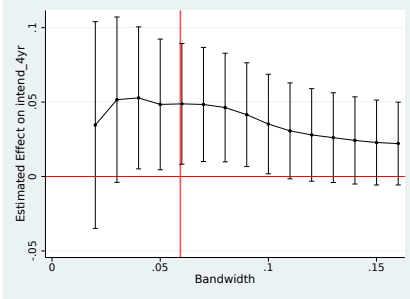

(e) Take SAT



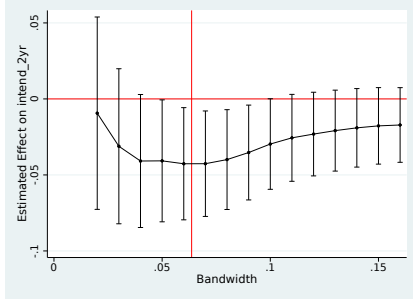

(f) Intend College




(g) Intend 4-year




(h) Intend 2-year




*Notes:* Each point in the figures above shows the point estimate and 90% confidence interval for the RD estimate using the given bandwidth on the x-axis. The vertical red line indicates the chosen bandwidth for the main analysis. Each RD specification includes subgroup and year fixed-effects.

## 1.3 Appendix Tables

Table A1: Sanctions for Consecutive AYP Failures under NCLB
From Ahn and Vigdor (2014)

| Number of consecutive years missed AYP in same subject | Sanction |
|---|---|
| 1 | None; placement on watch list, develop school improvement plan |
| 2 | District must offer transfers (with transportation) to higher-performing public schools in the same district. School listed as "needs improvement." |
| 3 | District must offer supplemental education services to students qualifying for free or reduced price lunch |
| 4 | School must undertake "corrective action." Corrective actions may include staff/leadership changes, curriculum changes, instructional time changes, or appointment of outside advisors. |
| 5 | School must formulate a restructuring plan. |
| 6 | School must implement the restructuring plan. Restructuring must involve either conversion to a charter school, replacement of the principal and most staff, state takeover, contracting with another entity to manage the school, or similar major change to school governance. |

Table A2: Balance tests

Subgroup RD

| | Premath | Prereading | Minority | EDS | Treated next year |
|---|---|---|---|---|---|
| Treated | 0.0210 | 0.0481 | 0.0231 | -0.00209 | 0.0335** |
| | (0.110) | (0.110) | (0.0734) | (0.216) | (0.0132) |
| Observations | 56,430 | 56,430 | 56,430 | 56,427 | 56,430 |

School RD

| | Premath | Prereading | Minority | EDS |
|---|---|---|---|---|
| Treated | -0.0398 | -0.0118 | 0.0711 | 0.0695 |
| | (0.0656) | (0.0660) | (0.0504) | (0.0544) |
| Observations | 450,358 | 450,358 | 450,358 | 450,280 |

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1
The bandwidth is 5 for the subgroup RDD
and MSE optimal (around 0.08) for the school RDD.
Standard errors are clustered at the subgroup-by-count level for the subgroup RDD and at the school level for the school RDD.

Table A3: Relationship between Test Scores and Long-Run Outcomes

Not controlling for prior-year test score

| | SAT-taker | SAT score | HS GPA | Dropout | Graduate | Intend 4-year | Intend 2-year | Intend college |
|---|---|---|---|---|---|---|---|---|
| Math | 0.102*** | 90.09*** | 0.254*** | -0.0187*** | 0.0454*** | 0.136*** | -0.0983*** | 0.0375*** |
| | (0.000465) | (0.227) | (0.000812) | (0.000260) | (0.000440) | (0.000593) | (0.000628) | (0.000455) |
| Reading | 0.0516*** | 70.98*** | 0.140*** | -0.0136*** | 0.0256*** | 0.0928*** | -0.0635*** | 0.0293*** |
| | (0.000466) | (0.235) | (0.000821) | (0.000261) | (0.000441) | (0.000599) | (0.000634) | (0.000460) |
| Observations | 2,004,023 | 657,542 | 1,149,794 | 1,591,885 | 1,591,885 | 1,330,660 | 1,330,660 | 1,330,660 |

Controlling for prior-year test score

| | SAT-taker | SAT score | HS GPA | Dropout | Graduate | Intend 4-year | Intend 2-year | Intend college |
|---|---|---|---|---|---|---|---|---|
| Math | 0.0660*** | 65.04*** | 0.197*** | -0.0153*** | 0.0386*** | 0.103*** | -0.0718*** | 0.0308*** |
| | (0.000536) | (0.261) | (0.000966) | (0.000309) | (0.000523) | (0.000704) | (0.000746) | (0.000542) |
| Reading | 0.0291*** | 50.04*** | 0.101*** | -0.0105*** | 0.0198*** | 0.0658*** | -0.0428*** | 0.0230*** |
| | (0.000519) | (0.257) | (0.000934) | (0.000298) | (0.000504) | (0.000683) | (0.000724) | (0.000526) |
| Observations | 2,004,023 | 657,542 | 1,149,794 | 1,591,885 | 1,591,885 | 1,330,660 | 1,330,660 | 1,330,660 |

Table A4: Commensurate vs. Actual Long-Run Effects

Using relationship between test score gains and long-run outcomes

*Subgroup RD*

| | Commensurate effect | Actual effect | Confidence Interval |
|---|---|---|---|
| Take SAT | 0.006 | 0.004 | (-0.010, 0.018) |
| SAT score | 7.56 | 18.98 | (5.75, 32.22) |
| GPA | 0.020 | 0.026 | (0.007, 0.045) |
| Dropout | -0.002 | -0.002 | (-0.011, 0.006) |
| Graduate | **0.004** | -0.005 | (-0.014, **0.004**) |
| Intend 4-year | 0.011 | 0.000 | (-0.015, 0.015) |
| Intend 2-year | -0.008 | -0.006 | (-0.021, 0.009) |
| Intend college | **0.004** | -0.006 | (-0.015, **0.003**) |

*School RD*

| | Commensurate effect | Actual effect | Confidence Interval |
|---|---|---|---|
| Take SAT | 0.009 | 0.003 | (-0.025, 0.031) |
| SAT score | 10.86 | 30.60 | (8.57, 52.6) |
| GPA | 0.028 | 0.004 | (-0.040, 0.049) |
| Dropout | -0.002 | 0.004 | (-0.003, 0.010) |
| Graduate | **0.006** | -0.011 | (-0.024, **0.002**) |
| Intend 4-year | 0.016 | 0.049 | (0.008, 0.090) |
| Intend 2-year | -0.011 | -0.043 | (-0.079, -0.007) |
| Intend college | 0.005 | 0.005 | (-0.010, 0.020) |

Table A5: Effects on Test Scores for SAT-takers

| | SAT-taker | | Not SAT-taker | |
|---|---|---|---|---|
| *Subgroup RDD* | | | | |
| Math effect | 0.0746*** | 0.0748*** | 0.0780*** | 0.0509*** |
| | (0.0263) | (0.0107) | (0.0191) | (0.0138) |
| Reading effect | 0.0649** | 0.0493*** | 0.0512*** | 0.0295*** |
| | (0.0295) | (0.0151) | (0.0193) | (0.0109) |
| Observations | 14,291 | 14,291 | 36,411 | 36,411 |
| *School RDD* | | | | |
| Math effect | 0.0757 | 0.0252 | 0.0900** | 0.0616** |
| | (0.0462) | (0.0266) | (0.0386) | (0.0266) |
| Reading effect | 0.100*** | 0.0309* | 0.0890*** | 0.0496*** |
| | (0.0320) | (0.0181) | (0.0318) | (0.0185) |
| Observations | 50,874 | 65,000 | 121,695 | 132,181 |
| Year FE | YES | YES | YES | YES |
| Subgroup FE | YES | YES | YES | YES |
| Student controls | NO | YES | NO | YES |

Standard errors are in parentheses,
clustered at the subgroub-by-count level for subgroup RDD
and the school level for school RDD
*** p<0.01, ** p<0.05, * p<0.1
Each entry in the table is the coefficient from a local linear regression
restricted to SAT-takers for the first two columns,
restricted to non SAT-takers for the last two columns,
using the outcome variable listed on the left.
Student controls include lagged test scores, gender,
limited English proficiency, and grade FE.
The bandwidth is 5 for subgroup RDD, MSE-optimal for school RDD.

Table A6: Effects on Long-run Outcomes for SAT-takers - Subgroup RDD

| | SAT-taker | | Not SAT-taker | |
|---|---|---|---|---|
| Dropout | 0.000171 | 0.000161 | -0.00174 | -0.00162 |
| | (0.000415) | (0.000411) | (0.00722) | (0.00694) |
| Graduate | -0.00627 | -0.00617 | -0.00659 | -0.00527 |
| | (0.00393) | (0.00383) | (0.00682) | (0.00664) |
| GPA | 0.0426*** | 0.0387*** | 0.0139 | 0.0108 |
| | (0.0136) | (0.0122) | (0.0141) | (0.0125) |
| Intend college | -0.00357 | -0.00279 | -0.00854 | -0.00863 |
| | (0.00499) | (0.00500) | (0.0109) | (0.0110) |
| Intend 4-year | -0.00168 | -0.00429 | 0.00446 | 0.00128 |
| | (0.0123) | (0.0115) | (0.00755) | (0.00664) |
| Intend 2-year | -0.00189 | 0.00149 | -0.0130 | -0.00992 |
| | (0.0117) | (0.0105) | (0.0102) | (0.00988) |
| Year FE | YES | YES | YES | YES |
| Subgroup FE | YES | YES | YES | YES |
| Student controls | NO | YES | NO | YES |

Standard errors are in parentheses,
clustered at the subgroub-by-count level .
*** p<0.01, ** p<0.05, * p<0.1
Each entry in the table is the coefficient from a local linear regression
restricted to SAT-takers for the first two columns,
restricted to non SAT-takers for the last two columns,
using the outcome variable listed on the left.
Student controls include lagged test scores, gender,
limited English proficiency, and grade FE.
The bandwidth is 5.

Table A7: Effects on Long-run Outcomes for SAT-takers - School RDD

| | SAT-taker | | Not SAT-taker | |
|---|---|---|---|---|
| Dropout | 0.000241 | 0.000270 | 0.00348 | 0.00316 |
| | (0.000612) | (0.000599) | (0.00624) | (0.00617) |
| Graduate | -0.00250 | -0.00143 | -0.0118 | -0.0119 |
| | (0.00337) | (0.00341) | (0.0134) | (0.0131) |
| GPA | -0.00200 | -0.0319 | -0.00649 | -0.0193 |
| | (0.0274) | (0.0321) | (0.0363) | (0.0374) |
| Intend college | -0.00388 | -0.00447 | 0.0149 | 0.0113 |
| | (0.00518) | (0.00514) | (0.0144) | (0.0142) |
| Intend 4-year | 0.0349* | 0.0245 | 0.0453** | 0.0393** |
| | (0.0187) | (0.0177) | (0.0196) | (0.0182) |
| Intend 2-year | -0.0418** | -0.0312* | -0.0306 | -0.0293 |
| | (0.0173) | (0.0163) | (0.0222) | (0.0215) |
| Year FE | YES | YES | YES | YES |
| Subgroup FE | YES | YES | YES | YES |
| Student controls | NO | YES | NO | YES |

Standard errors are in parentheses,
clustered at the school level.
*** p<0.01, ** p<0.05, * p<0.1
Each entry in the table is the coefficient from a local linear regression
restricted to SAT-takers for the first two columns,
restricted to non SAT-takers for the last two columns,
using the outcome variable listed on the left.
Student controls include lagged test scores, gender,
limited English proficiency, and grade FE.
The bandwidth is MSE-optimal.

Table A8: Effects on Test Scores for Students in Schools on the Margin of Failing AYP

| | Marginal schools | | Not marginal schools | |
|---|---|---|---|---|
| *Subgroup RDD* | | | | |
| Math effect | 0.0518** | 0.0400* | 0.0930*** | 0.0733*** |
| | (0.0249) | (0.0238) | (0.0250) | (0.0145) |
| Reading effect | 0.0824*** | 0.0585*** | 0.0340 | 0.0204 |
| | (0.0255) | (0.0159) | (0.0299) | (0.0138) |
| Observations | 21,855 | 21,855 | 28,847 | 28,847 |
| Year FE | YES | YES | YES | YES |
| Subgroup FE | YES | YES | YES | YES |
| Student controls | NO | YES | NO | YES |

Standard errors are in parentheses,
clustered at the subgroub-by-count level.
*** p<0.01, ** p<0.05, * p<0.1
Each entry in the table is the coefficient from a local linear regression
restricted to students in marginal schools for the first two columns,
restricted to students in not-marginal schools for the last two columns,
using the outcome variable listed on the left (math or reading score).
Marginal schools are those below the median in absolute value
of minimum distance from AYP target.
Student controls include lagged test scores, gender,
limited English proficiency, and grade FE.
The bandwidth is 5.

Table A9: Effects on Long-run Outcomes for Students in Schools on the Margin of Failing AYP

| | Marginal schools | | Not marginal schools | |
|---|---|---|---|---|
| *Subgroup RDD* | | | | |
| Dropout | 0.00113 | 0.00183 | -0.00461 | -0.00484 |
| | (0.00543) | (0.00525) | (0.00534) | (0.00545) |
| Graduate | -0.00852 | -0.00915 | -0.00233 | -0.00156 |
| | (0.00727) | (0.00715) | (0.00813) | (0.00812) |
| GPA | 0.0239 | 0.0139 | 0.0342* | 0.0300 |
| | (0.0146) | (0.0157) | (0.0198) | (0.0196) |
| SAT score | 15.99** | 15.40*** | 21.20 | 20.43** |
| | (7.633) | (5.204) | (14.04) | (9.946) |
| Take SAT | 0.0134 | 0.0115 | -0.00135 | -0.00410 |
| | (0.00935) | (0.00953) | (0.0153) | (0.0150) |
| Intend college | -0.0134 | -0.0137 | 0.000367 | -0.000946 |
| | (0.00967) | (0.00892) | (0.0103) | (0.0109) |
| Intend 4-year | 0.0249* | 0.0188 | -0.0191 | -0.0229 |
| | (0.0128) | (0.0134) | (0.0178) | (0.0147) |
| Intend 2-year | -0.0383*** | -0.0324*** | 0.0195 | 0.0219 |
| | (0.0126) | (0.0125) | (0.0176) | (0.0148) |
| Year FE | YES | YES | YES | YES |
| Subgroup FE | YES | YES | YES | YES |
| Student controls | NO | YES | NO | YES |

Standard errors are in parentheses,
clustered at the subgroub-by-count level.
*** p<0.01, ** p<0.05, * p<0.1
Each entry in the table is the coefficient from a local linear regression
restricted to students in marginal schools for the first two columns,
restricted to students in not-marginal schools for the last two columns,
using the outcome variable listed on the left.
Marginal schools are those below the median in absolute value
of minimum distance from AYP target.
Student controls include lagged test scores, gender,
limited English proficiency, and grade FE.
The bandwidth is 5.

Table A10: Effects on Test Scores for Students in Subgroups Below 40 AYP

| | Subgroup below 40 | | Subgroup 40 or more | |
|---|---|---|---|---|
| *School RDD* | | | | |
| Math effect | 0.0640 | 0.108** | 0.0935** | 0.0416 |
| | (0.0531) | (0.0421) | (0.0446) | (0.0253) |
| Reading effect | 0.0591 | 0.0906*** | 0.117*** | 0.0378** |
| | (0.0411) | (0.0260) | (0.0367) | (0.0174) |
| Observations | 29,837 | 23,879 | 137,339 | 173,582 |
| Year FE | YES | YES | YES | YES |
| Subgroup FE | YES | YES | YES | YES |
| Student controls | NO | YES | NO | YES |

Standard errors are in parentheses,
clustered at the school level.
*** p<0.01, ** p<0.05, * p<0.1
Each entry in the table is the coefficient from a local linear regression,
restricted to students in a school with fewer than 40 students
in their subgroup for the first two columns,
restricted to students in a school with 40 or more students
in their subgroup for the last two columns,
using the outcome variable listed on the left (math or reading score).
Student controls include lagged test scores, gender,
limited English proficiency, and grade FE.
The bandwidth is MSE-optimal.

Table A11: Effects on Long-run Outcomes for Students in Subgroups Below 40 AYP

| | Subgroup below 40 | | Subgroup 40 or more | |
|---|---|---|---|---|
| *School RDD* | | | | |
| Dropout | 0.00297 | 0.00291 | 0.00332 | 0.00437 |
| | (0.00954) | (0.00958) | (0.00390) | (0.00378) |
| Graduate | 0.00284 | 0.00308 | -0.0126 | -0.0151* |
| | (0.0165) | (0.0167) | (0.00803) | (0.00773) |
| GPA | -0.00381 | -0.00456 | 0.00945 | -0.0187 |
| | (0.0416) | (0.0388) | (0.0282) | (0.0313) |
| SAT score | 31.65 | 22.02 | 30.25** | 20.43* |
| | (20.17) | (14.91) | (12.86) | (10.83) |
| Take SAT | -0.00487 | 0.00123 | 0.000904 | -0.0102 |
| | (0.0193) | (0.0191) | (0.0172) | (0.0140) |
| Intend college | 0.0193 | 0.0196 | 0.00337 | -0.000123 |
| | (0.0187) | (0.0188) | (0.00918) | (0.00865) |
| Intend 4-year | 0.0108 | 0.0103 | 0.0544** | 0.0428* |
| | (0.0271) | (0.0243) | (0.0275) | (0.0239) |
| Intend 2-year | 0.0117 | 0.0123 | -0.0497** | -0.0408* |
| | (0.0299) | (0.0282) | (0.0239) | (0.0219) |
| Year FE | YES | YES | YES | YES |
| Subgroup FE | YES | YES | YES | YES |
| Student controls | NO | YES | NO | YES |

Standard errors are in parentheses,
clustered at the school level.
*** p<0.01, ** p<0.05, * p<0.1
Each entry in the table is the coefficient from a local linear regression,
restricted to students in a school with fewer than 40 students
in their subgroup for the first two columns,
restricted to students in a school with 40 or more students
in their subgroup for the last two columns,
using the outcome variable listed on the left.
Student controls include lagged test scores, gender,
limited English proficiency, and grade FE.
The bandwidth is MSE-optimal.

Table A12: Effects on Test Scores for Students on the Margin of Proficiency

| | Marginal students | | Not marginal students | |
|---|---|---|---|---|
| *Subgroup RDD* | | | | |
| Math effect | 0.0527*** | 0.0531*** | 0.0513** | 0.0117 |
| | (0.0189) | (0.0134) | (0.0261) | (0.0116) |
| Reading effect | 0.0321** | 0.0471*** | 0.0168 | -0.00494 |
| | (0.0137) | (0.0101) | (0.0224) | (0.0110) |
| Observations | 23,668 | 23,668 | 15,439 | 15,439 |
| *School RDD* | | | | |
| Math effect | 0.00309 | 0.000655 | 0.0581 | 0.0120 |
| | (0.0325) | (0.0291) | (0.0490) | (0.0278) |
| Reading effect | -0.00780 | 0.00729 | 0.121*** | 0.0715*** |
| | (0.0293) | (0.0235) | (0.0327) | (0.0196) |
| Observations | 86,279 | 78,854 | 85,425 | 82,221 |
| Year FE | YES | YES | YES | YES |
| Subgroup FE | YES | YES | YES | YES |
| Student controls | NO | YES | NO | YES |

Standard errors are in parentheses,
clustered at the subgroub-by-count level for subgroup RDD
and the school level for school RDD
*** p<0.01, ** p<0.05, * p<0.1
Each entry in the table is the coefficient from a local linear regression
restricted to students in marginal students for the first two columns,
restricted to students in not-marginal students for the last two columns,
using the outcome variable listed on the left.
Marginal students are those below the median in absolute value
of predicted distance from the proficiency test score.
Student controls include lagged test scores, gender,
limited English proficiency, and grade FE.
The bandwidth is 5 for subgroup RDD, MSE-optimal for school RDD.

Table A13: Effects on Long-run Outcomes for Marginal Students - Subgroup RDD

| | Marginal students | | Not marginal students | |
|---|---|---|---|---|
| Dropout | -0.00454 | -0.00591 | -0.000570 | 0.000832 |
| | (0.00443) | (0.00459) | (0.00656) | (0.00620) |
| Graduate | -0.00546 | -0.00438 | -0.0113 | -0.0142 |
| | (0.00660) | (0.00649) | (0.00893) | (0.00880) |
| GPA | -0.00136 | -0.000319 | 0.0353* | 0.0244 |
| | (0.0155) | (0.0134) | (0.0209) | (0.0192) |
| SAT score | 5.811 | 7.143 | 26.28* | 16.00* |
| | (4.549) | (5.418) | (13.93) | (9.697) |
| Take SAT | 0.0117 | 0.0121 | -0.0179 | -0.0228 |
| | (0.00850) | (0.00843) | (0.0164) | (0.0155) |
| Intend college | -0.0151 | -0.0157 | -0.0231*** | -0.0230*** |
| | (0.0110) | (0.0111) | (0.00530) | (0.00610) |
| Intend 4-year | -0.0174** | -0.0170** | 0.000976 | -0.00873 |
| | (0.00870) | (0.00769) | (0.0186) | (0.0170) |
| Intend 2-year | 0.00231 | 0.00121 | -0.0241 | -0.0143 |
| | (0.0156) | (0.0145) | (0.0162) | (0.0150) |
| Year FE | YES | YES | YES | YES |
| Subgroup FE | YES | YES | YES | YES |
| Student controls | NO | YES | NO | YES |

Standard errors are in parentheses,
clustered at the subgroub-by-count level .
*** p<0.01, ** p<0.05, * p<0.1
Each entry in the table is the coefficient from a local linear regression
restricted to students in marginal students for the first two columns,
restricted to students in not-marginal students for the last two columns,
using the outcome variable listed on the left.
Marginal students are those below the median in absolute value
of predicted distance from the proficiency test score.
Student controls include lagged test scores, gender,
limited English proficiency, and grade FE.
The bandwidth is 5.

Table A14: Effects on Long-run Outcomes for Marginal Students - School RDD

| | Marginal students | | Not marginal students | |
|---|---|---|---|---|
| Dropout | 0.00185 | -0.000821 | 0.00497 | 0.00767 |
| | (0.00655) | (0.00655) | (0.00611) | (0.00606) |
| Graduate | -0.00448 | -0.00246 | -0.0230** | -0.0276*** |
| | (0.0130) | (0.0130) | (0.0107) | (0.0103) |
| GPA | -0.0384 | -0.0411 | 0.0345 | 0.0107 |
| | (0.0354) | (0.0346) | (0.0376) | (0.0371) |
| SAT score | 19.67 | 19.92* | 34.67** | 23.02** |
| | (12.26) | (10.71) | (14.73) | (10.99) |
| Take SAT | -0.00848 | -0.00892 | 0.0121 | 0.00612 |
| | (0.0156) | (0.0164) | (0.0257) | (0.0234) |
| Intend college | -0.00002 | -0.00363 | 0.0109 | 0.00630 |
| | (0.0143) | (0.0140) | (0.0118) | (0.0110) |
| Intend 4-year | 0.0222 | 0.0207 | 0.0669* | 0.0504 |
| | (0.0233) | (0.0225) | (0.0375) | (0.0327) |
| Intend 2-year | -0.0316 | -0.0341 | -0.0463 | -0.0329 |
| | (0.0249) | (0.0245) | (0.0315) | (0.0282) |
| Year FE | YES | YES | YES | YES |
| Subgroup FE | YES | YES | YES | YES |
| Student controls | NO | YES | NO | YES |

Standard errors are in parentheses,
clustered at the school level.
*** p<0.01, ** p<0.05, * p<0.1
Each entry in the table is the coefficient from a local linear regression
restricted to students in marginal students for the first two columns,
restricted to students in not-marginal students for the last two columns,
using the outcome variable listed on the left.
Marginal students are those below the median in absolute value
of predicted distance from the proficiency test score.
Student controls include lagged test scores, gender,
limited English proficiency, and grade FE.
The bandwidth is MSE-optimal.

Table A15: Effects on Test Scores for Subgroups of Students

| | White | | Black | | Hispanic | | Minority | | Disadvantaged | |
|---|---|---|---|---|---|---|---|---|---|---|
| *Subgroup RDD* | | | | | | | | | | |
| Math effect | 0.128*** | 0.0929*** | 0.0367 | 0.0362* | 0.0507** | 0.0243* | 0.0430** | 0.0296* | 0.0691*** | 0.0517*** |
| | (0.0372) | (0.0184) | (0.0309) | (0.0198) | (0.0201) | (0.0142) | (0.0209) | (0.0160) | (0.00985) | (0.00724) |
| Reading effect | 0.0829 | 0.0604* | 0.0323 | 0.0336** | 0.0521* | 0.00718 | 0.0383** | 0.0180 | 0.0576*** | 0.0356*** |
| | (0.0549) | (0.0318) | (0.0229) | (0.0157) | (0.0305) | (0.0304) | (0.0194) | (0.0172) | (0.00748) | (0.00446) |
| Observations | 14,641 | 14,641 | 19,478 | 19,478 | 11,918 | 11,918 | 31,396 | 31,396 | 41,923 | 41,923 |
| *School RDD* | | | | | | | | | | |
| Math effect | 0.163*** | 0.0654** | 0.00450 | 0.0105 | -0.0273 | 0.00888 | -0.00810 | 0.00585 | 0.0228 | 0.0289 |
| | (0.0616) | (0.0314) | (0.0415) | (0.0288) | (0.0634) | (0.0472) | (0.0372) | (0.0266) | (0.0370) | (0.0276) |
| Reading effect | 0.221*** | 0.101*** | -0.0162 | -0.0136 | 0.000164 | 0.0506 | -0.00485 | 0.00196 | 0.0265 | 0.0256 |
| | (0.0588) | (0.0277) | (0.0312) | (0.0229) | (0.0537) | (0.0406) | (0.0314) | (0.0216) | (0.0277) | (0.0182) |
| Observations | 95,104 | 104,322 | 62,557 | 66,669 | 18,005 | 19,168 | 87,332 | 93,282 | 115,107 | 104,823 |
| Year FE | YES | YES | YES | YES | YES | YES | YES | YES | YES | YES |
| Subgroup FE | YES | YES | YES | YES | YES | YES | YES | YES | YES | YES |
| Student controls | NO | YES | NO | YES | NO | YES | NO | YES | NO | YES |

Standard errors are in parentheses, clustered at the subgroub-by-count level for subgroup RDD estimates and the school level for school RDD estimates.

*** p<0.01, ** p<0.05, * p<0.1

Each entry in the table is the coefficient from a local linear regression restricted to students in the subgroup listed on the top, using the outcome variable listed on the left (math or reading test score).

Student controls include lagged test scores, gender, limited English proficiency, and grade FE.

The bandwidth is 5 for the subgroup RDD and MSE optimal (around 0.08) for the school RDD.

Table A16: Effects on Long-run Outcomes for Subgroups of Students - Subgroup RDD

| | White | | Black | | Hispanic | | Minority | | Disadvantaged | |
|---|---|---|---|---|---|---|---|---|---|---|
| Dropout | 0.00749 | 0.00893 | -0.00978*** | -0.0105*** | -0.00935 | -0.00861 | -0.00959* | -0.00947* | -0.00383 | -0.00374 |
| | (0.00624) | (0.00551) | (0.00239) | (0.00286) | (0.00969) | (0.00972) | (0.00496) | (0.00517) | (0.00524) | (0.00528) |
| Graduate | -0.0139 | -0.0154 | 0.0151 | 0.0181** | -0.0121* | -0.0146** | 0.00406 | 0.00449 | 0.00393 | 0.00410 |
| | (0.0129) | (0.0126) | (0.00937) | (0.00853) | (0.00679) | (0.00656) | (0.00435) | (0.00409) | (0.00469) | (0.00473) |
| GPA | 0.0705* | 0.0679* | 0.0449* | 0.0432 | -0.0419*** | -0.0535*** | 0.00990 | 0.00429 | 0.0393*** | 0.0323*** |
| | (0.0424) | (0.0367) | (0.0258) | (0.0282) | (0.00931) | (0.00853) | (0.0150) | (0.0178) | (0.00435) | (0.00556) |
| SAT score | 10.79 | -1.833 | 0.785 | 6.620* | 28.71*** | 23.10*** | 8.532*** | 11.60*** | 14.78*** | 16.31*** |
| | (23.63) | (15.12) | (3.116) | (3.837) | (5.745) | (6.256) | (3.169) | (3.884) | (3.400) | (3.353) |
| Take SAT | 0.0307** | 0.0269** | -0.0187* | -0.0172 | 0.0356*** | 0.0292*** | 0.00433 | 0.00186 | 0.0130** | 0.0103* |
| | (0.0136) | (0.0116) | (0.0106) | (0.0113) | (0.00822) | (0.00882) | (0.00936) | (0.00953) | (0.00560) | (0.00599) |
| Intend college | 0.0319*** | 0.0292*** | -0.0523*** | -0.0534*** | 0.0330* | 0.0340* | -0.0191* | -0.0190* | -0.00309 | -0.00393 |
| | (0.00798) | (0.00954) | (0.00582) | (0.00596) | (0.0170) | (0.0179) | (0.0103) | (0.0106) | (0.00599) | (0.00581) |
| Intend 4-year | 0.00743 | -0.00152 | -0.00791 | -0.00492 | 0.00441 | -0.00656 | -0.00201 | -0.00476 | 0.00203 | -0.00275 |
| | (0.0259) | (0.0192) | (0.0112) | (0.0123) | (0.0152) | (0.0122) | (0.0103) | (0.0103) | (0.00585) | (0.00526) |
| Intend 2-year | 0.0245 | 0.0307 | -0.0444*** | -0.0484*** | 0.0286 | 0.0405* | -0.0171 | -0.0142 | -0.00512 | -0.00117 |
| | (0.0242) | (0.0192) | (0.0100) | (0.0114) | (0.0219) | (0.0221) | (0.0122) | (0.0118) | (0.00603) | (0.00547) |
| Year FE | YES | YES | YES | YES | YES | YES | YES | YES | YES | YES |
| Subgroup FE | YES | YES | YES | YES | YES | YES | YES | YES | YES | YES |
| Student controls | NO | YES | NO | YES | NO | YES | NO | YES | NO | YES |

Standard errors are in parentheses, clustered at the subgroub-by-count level.
*** p<0.01, ** p<0.05, * p<0.1
Each entry in the table is the coefficient from a local linear regression restricted to students in the subgroup listed on the top, using the outcome variable listed on the left.
Student controls include lagged test scores, gender, limited English proficiency, and grade FE.
The bandwidth is 5.

Table A17: Effects on Long-run Outcomes for Subgroups of Students - School RDD

| | White | | Black | | Hispanic | | Minority | | Disadvantaged | |
|---|---|---|---|---|---|---|---|---|---|---|
| Dropout | -0.000406 | 0.00288 | 0.00914 | 0.00473 | 0.00471 | 0.00119 | 0.00789 | 0.00383 | 0.00814 | 0.00617 |
| | (0.00470) | (0.00475) | (0.00737) | (0.00790) | (0.0136) | (0.0133) | (0.00713) | (0.00754) | (0.00588) | (0.00577) |
| Graduate | -0.00260 | -0.00913 | -0.0269 | -0.0206 | -0.00794 | -0.00359 | -0.0229 | -0.0167 | -0.0224** | -0.0198* |
| | (0.00932) | (0.00910) | (0.0181) | (0.0178) | (0.0222) | (0.0213) | (0.0157) | (0.0155) | (0.0114) | (0.0111) |
| GPA | 0.0567* | 0.0125 | -0.0643 | -0.0604 | -0.103 | -0.0943 | -0.0811* | -0.0731 | -0.0520* | -0.0549* |
| | (0.0343) | (0.0361) | (0.0449) | (0.0455) | (0.0741) | (0.0767) | (0.0443) | (0.0448) | (0.0307) | (0.0326) |
| SAT score | 52.47*** | 28.39** | 3.253 | -0.554 | -10.55 | -6.776 | 0.932 | -1.088 | 10.10 | 5.210 |
| | (19.13) | (14.43) | (9.379) | (8.228) | (19.28) | (12.46) | (9.035) | (7.805) | (8.644) | (6.739) |
| Take SAT | 0.0201 | 0.00131 | -0.0157 | -0.0136 | -0.0277 | -0.0212 | -0.0184 | -0.0148 | -0.0225* | -0.0220* |
| | (0.0247) | (0.0193) | (0.0252) | (0.0239) | (0.0226) | (0.0229) | (0.0206) | (0.0202) | (0.0124) | (0.0119) |
| Intend college | 0.0103 | 0.00317 | -0.00207 | -0.00281 | -0.0164 | -0.0136 | -0.00555 | -0.00624 | 0.00740 | 0.00728 |
| | (0.0110) | (0.0101) | (0.0166) | (0.0160) | (0.0327) | (0.0337) | (0.0162) | (0.0160) | (0.0129) | (0.0129) |
| Intend 4-year | 0.0953** | 0.0659** | -0.0120 | -0.0118 | -0.0204 | -0.0141 | -0.0144 | -0.0131 | -0.00195 | -0.00254 |
| | (0.0371) | (0.0297) | (0.0249) | (0.0248) | (0.0292) | (0.0291) | (0.0220) | (0.0224) | (0.0166) | (0.0155) |
| Intend 2-year | -0.0821*** | -0.0590** | 0.00993 | 0.00978 | 0.00549 | 0.00255 | 0.0107 | 0.00820 | 0.00818 | 0.00863 |
| | (0.0318) | (0.0264) | (0.0262) | (0.0262) | (0.0348) | (0.0344) | (0.0243) | (0.0243) | (0.0169) | (0.0163) |
| Year FE | YES | YES | YES | YES | YES | YES | YES | YES | YES | YES |
| Subgroup FE | YES | YES | YES | YES | YES | YES | YES | YES | YES | YES |
| Student controls | NO | YES | NO | YES | NO | YES | NO | YES | NO | YES |

Standard errors are in parentheses, clustered at the school level.
*** p<0.01, ** p<0.05, * p<0.1
Each entry in the table is the coefficient from a local linear regression restricted to students in the subgroup listed on the top, using the outcome variable listed on the left.
Student controls include lagged test scores, gender, limited English proficiency, and grade FE.
The bandwidth is MSE optimal (around 0.08).

Table A18: Effect of Accountability on Missing Long-run Outcomes

*Subgroup RDD*

| Missing: | SAT | | GPA | | Graduation | | College Intention | |
|---|---|---|---|---|---|---|---|---|
| Treated | -0.00591 | -0.00341 | -0.00161 | -0.00315 | -0.0148*** | -0.0171*** | -0.00436 | -0.00606 |
| | (0.00851) | (0.00844) | (0.00932) | (0.00976) | (0.00533) | (0.00571) | (0.00678) | (0.00688) |
| Observations | 50,747 | 50,747 | 50,747 | 50,747 | 50,747 | 50,747 | 50,747 | 50,747 |

*School RDD*

| Missing: | SAT | | GPA | | Graduation | | College Intention | |
|---|---|---|---|---|---|---|---|---|
| Treated | -0.00269 | 0.00829 | -0.0250 | -0.0197 | 0.00219 | 0.00240 | 0.0109 | 0.0130 |
| | (0.0170) | (0.0137) | (0.0220) | (0.0216) | (0.0116) | (0.0110) | (0.0130) | (0.0126) |
| Observations | 213,279 | 213,279 | 213,279 | 213,279 | 213,279 | 213,279 | 213,279 | 213,279 |
| Year FE | YES | YES | YES | YES | YES | YES | YES | YES |
| Subgroup FE | YES | YES | YES | YES | YES | YES | YES | YES |
| Student controls | NO | YES | NO | YES | NO | YES | NO | YES |

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1
Student controls include lagged test scores, gender,
limited English proficiency, and grade FE.
The bandwidth is 5 for the subgroup RDD
and MSE optimal (around 0.08) for the school RDD.
Standard errors are clustered at the subgroup-by-count level
for the subgroup RDD and at the school level for the school RDD.

Table A19: Changes in Inputs around Subgroup Cutoff

*Subgroup RDD*

| | Teacher Math VA | | Teacher Reading VA | | Teacher Experience | | Transferred Teacher | | Class Size | |
|---|---|---|---|---|---|---|---|---|---|---|
| Treated | -0.00138 | -0.00198 | 0.000644 | 0.000147 | 0.0324 | 0.0478 | -0.00338 | -0.00343 | -0.0907 | -0.0804 |
| | (0.00665) | (0.00662) | (0.00566) | (0.00564) | (0.165) | (0.164) | (0.00757) | (0.00756) | (0.118) | (0.114) |
| Observations | 283,023 | 272,380 | 260,162 | 260,162 | 301,015 | 311,215 | 198,502 | 191,332 | 240,689 | 240,689 |

*School RDD*

| | Teacher Math VA | | Teacher Reading VA | | Teacher Experience | | Transferred Teacher | | Class Size | |
|---|---|---|---|---|---|---|---|---|---|---|
| Treated | 0.0469** | 0.0454** | 0.0651*** | 0.0643*** | 0.178 | 0.128 | -0.0270 | -0.0251 | -0.446 | -0.468 |
| | (0.0229) | (0.0224) | (0.0186) | (0.0184) | (0.755) | (0.758) | (0.0262) | (0.0261) | (0.558) | (0.550) |
| Observations | 166,430 | 171,470 | 133,904 | 133,786 | 215,890 | 214,115 | 208,550 | 209,177 | 169,122 | 171,724 |
| Year FE | YES | YES | YES | YES | YES | YES | YES | YES | YES | YES |
| Subgroup FE | YES | YES | YES | YES | YES | YES | YES | YES | YES | YES |
| Student controls | NO | YES | NO | YES | NO | YES | NO | YES | NO | YES |

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1
Student controls include lagged test scores, gender,
limited English proficiency, and grade FE.
The bandwidth is 5 for the subgroup RDD
and MSE optimal (around 0.08) for the school RDD.
Standard errors are clustered at the subgroup-by-count level
for the subgroup RDD and at the school level for the school RDD.