Incentives and Responses under *No Child Left Behind*: Credible threats and the Role of Competition

Rajashri Chakrabarti[†]
Federal Reserve Bank of New York

Abstract

NCLB mandated the institution of Adequate Yearly Progress (AYP) objectives, and schools are assigned an AYP pass/fail based on performance in these objectives. AYP-fail status is associated with negative publicity and often sanctions. Using data from Wisconsin and alternate regression discontinuity designs, I study the incentives and responses of schools that failed AYP once. Math-induced AYP-failures showed strong improvements in math, while reading-induced AYP-failures showed marked improvements in reading. Consistent with incentives, these schools showed no positive effect in other high stakes objectives. In contrast, test-participation failures showed no effect in either high stakes reading or math, while they showed some evidence of positive (though, not statistically significant) effects in test participation. Improvements in reading are associated with parallel effects in low stakes language arts (possibly, due to spillover effects), while there is no evidence of effects in low stakes science or social studies. Nor is there evidence of effects on graduation rates. Performance in low stakes grades suffered, and so did performance in weaker subgroups in spite of their inclusion in AYP computations. There is evidence of focus on marginal students around high stakes cutoffs in subject areas AYP-failed schools improved in, but this did not come at the expense of ends. Credibility of threat mattered—AYP-failed schools that faced more competition responded considerably more strongly in the objectives they had incentives in.

Keywords: No Child Left Behind, Incentives, Public School Performance, Regression Discontinuity
JEL Classifications: H4, I21, I28

[*]I thank Damon Clark, Caroline Hoxby, Randy Reback, Jonah Rockoff, Wilbert van der Klaauw, Matt Wiswall, Basit Zafar, and seminar participants at Columbia University, FRBNY/NYU Education Seminar Series, NBER Economics of Education Meeting, University of Houston, American Economic Association Conference and Association for Education Finance and Policy conference for helpful discussions and the Wisconsin Department of Public Instruction for data used in this analysis. Sophia Gilbukh and Noah Schwartz provided excellent research assistance. The views expressed in this paper are those of the author and do not necessarily reflect the position of the Federal Reserve Bank of New York or the Federal Reserve System. All errors are my own.

[†]Federal Reserve Bank of New York, 33 Liberty Street, New York, NY 10045. Email: Rajashri.Chakrabarti@ny.frb.org

1 Introduction

Concern over public school performance since the mid-1980s led to strong demands for public school reform. To some extent as a response to these demands, the 1990s saw a surge of test-based state accountability systems. These culminated in the federal *No Child Left Behind* (NCLB) law that stipulated the implementation of statewide test-based accountability systems and assignment of pass/fail statuses to schools based on these tests.

Signed into law on January 8, 2002, NCLB mandated testing of all students in reading and math in grades 3-8. States were required to institute "Adequate Yearly Progress" objectives and schools were assigned an AYP-pass or AYP-fail based on performance in these objectives. These AYP statuses were made publicly available and were often associated with media publicity and visibility. Consequently, AYP-failure was associated with shame and stigma. In addition, Title 1 schools[1] that failed AYP *in the same objective* for two or more consecutive years faced federal sanctions that escalated with the number of years of failed AYP. Title 1 schools missing AYP (in the same objective) for two consecutive years were required to provide public school choice to their students, for three consecutive years were required to provide other supplemental services like tutoring. These services were associated with loss of public school funds as they were funded by public school money. These sanctions cumulated over the years until the school was restructured if it failed AYP for five consecutive years.

In this paper, I focus on the state of Wisconsin and study the incentives and responses of schools that failed AYP once ("threatened schools"). These schools were threatened in the sense that they faced stigma and often possibility of impending sanctions. Hence they had strong incentives to try to avoid AYP-failure in the next year. How might we expect these schools to respond? Reading and math were high stakes subject areas in the sense that scores from these tests entered AYP computations. Did this induce the threatened schools to focus more on high stakes subject areas and did this lead to a shift away from low stakes subject areas? In addition, under AYP rules, certain percentages of students had to score above some pre-designated cutoffs on the score scale to pass AYP. Did this induce the threatened schools to focus more on students expected to score just around the cutoffs and did this come at the expense of the ends? Moreover, while some grades were included in AYP computations, others were not. Did the threat of sanctions and stigma lead to a shift of focus away from low stakes grades and students to high stakes grades? In contrast to most other accountability systems, NCLB holds demographic and economic subgroups accountable. The purpose was to prevent weaker subgroups (for

[1] Title 1 schools are schools that receive federal funding under the federal Title 1 program.

example, economically disadvantaged, special education groups) from falling through the cracks. Did this provision lead the threatened schools to focus more on various subgroups, and were there differences in how different subgroups (for example, advantaged versus disadvantaged) were affected? I also look at the effect of "threatened status" (having failed AYP once) on performance in other high stakes indicators such as test participation, attendance, and graduation that also entered AYP formation.

Exploiting the institutional details of the system, I use alternate regression discontinuity (RD) designs to investigate these questions. As a first pass, I use a regression discontinuity strategy based on school's "minimum distance" of the various high stakes subject-subgroup criteria from the corresponding cutoffs. This RD design is based on the intuition that the lowest performing criterion determines the AYP status of schools—under NCLB, a school fails AYP even if it fails only one subgroup-criterion combination while passing all other criteria. While there is no threat to the validity of this design, the interpretations of the estimates from this design are less clear. By construction, schools that just barely missed the cutoff may have been placed there on the basis of different criteria and hence might have different incentives. Therefore, it is not apriori clear how this heterogenous group might respond in the various criteria/objectives[2]. Therefore, the bulk of the paper centers around alternative RD designs that focus more closely on incentives and aim to study the corresponding responses. I consider alternative RD strategies that probe into math-induced failure, reading induced failure, and test-participation induced failure. In each case, I consider schools that just barely missed/made the corresponding criterion (for example, say math)—here the discontinuity in AYP failure is generated solely by math-induced failure. These RD designs enable me to ask some targeted questions. Did math induced failure lead to more focus on math relative to the other criteria? Similarly, did reading induced failure lead to more focus on reading? Did test participation induced failure lead to more focus on test participation and less on other indicators?

In the "minimum distance" sample, I find that the threatened schools responded by improving in high stakes reading. While there is evidence of parallel patterns in language arts (possibly due to spillover effects), there is no evidence of any improvement in high stakes math or any of the other low stakes subject areas (science, social studies). In contrast, I find that math-induced failure led to strong improvements in math, while reading-induced failure led to marked improvements in reading. Interestingly, there is no evidence of improvements of these schools in other objectives. Notably, consistent with incentives, there is no evidence that test-participation induced failure led to any improvement in either reading or math, while there is some evidence (though not statistically significant) that it led to positive effects on

[2] I use "criteria" and "objectives" interchangeably in this paper.

test participation. These findings suggest that responses of the threatened schools depended crucially on performance in the AYP objectives in the previous year. AYP-failed schools responded in the criterion that induced failure, presumably in an effort to obviate a second consecutive failure in the same criterion which would expose them to further stigma and sanctions.

Results from the various RD designs show that the AYP-failed schools indeed tended to focus more on the high stakes students close to and around (or just below) the cutoff in the high stakes subject areas they improved in. But this did not come at the expense of the ends—rather, there was a rightward shift of the whole distribution. However, there is robust evidence that improvements in high stakes subject areas in high stakes grades did come at the expense of performance in low stakes grades.

The patterns in subgroup performance reveal that the weaker subgroups (economically disadvantaged, special education) lost irrespective of the law's emphasis on these groups. In contrast, white students showed positive effects in all subject areas, though the effects were not always statistically significant. As far as the other indicators go, I do not find evidence of robust effects either in test participation or in graduation. In contrast, I find some evidence that attendance improved (relatively) in threatened schools where they mattered for AYP.

I also investigate whether the response of the AYP-failed schools depended on the extent of competition they faced. To cast some light into this question, I investigate whether schools that faced more competition (had more AYP-passed schools in their near vicinity) responded more strongly. This indeed seems to have been the case. Using the "minimum distance" RD design I find that schools that faced more competition not only responded more strongly in reading, but also showed improvements in the other high stakes criteria as well—math, test participation, attendance and graduation. By construction, the group of schools that just barely missed the cutoff under this design did so on the basis of the different high stakes criteria (footnote 10), and hence had incentives in these various criteria. While these AYP-failed schools as a group did not show statistically significant evidence of improvement in high stakes areas other than reading, AYP-failed schools that faced viable competition responded much more strongly and broadly, and showed improvements in each of the high stakes criteria.

These competition effects become even more clear in the latter RD designs, where the incentives were more precise. There is robust evidence that math-induced failures (reading-induced failures) facing more competition responded much more strongly in math (reading), while there is no evidence that these schools responded in the other high stakes criteria. These competition effects are very robust, and survive a battery of sensitivity checks. These results imply that AYP-failed schools did indeed respond according to incentives—feasibility and credibility of threat mattered. AYP-failed schools that faced

meaningful competition responded strongly in the criteria they faced incentives in.

A rich literature on school accountability studies the effects of various accountability systems on public school performance and behavior. This literature generally finds positive effects on public school performance (Figlio and Rouse (2006), West and Peterson (2006), Rouse et al. (2013), Chiang (2009), Rockoff and Turner (2010), Chakrabarti (Forthcoming)). But, there is also evidence in favor of reclassification of low performing students into disabled or limited english proficient categories (Cullen and Reback (2006), Figlio and Getzler (2006) and Jacob (2005), Chakrabarti (2013b)), teacher cheating (Jacob and Levitt (2003)), strategic suspensions of students (Figlio (2006)), increased focus on high stakes marginal students (Reback (2008), Chakrabarti (2013a)) and even strategic boosting of caloric content of school lunches during the test taking period (Figlio and Winicki (2005)). This study is also related to the literature on voucher competition (Hoxby (2003a), Hoxby (2003b), Chakrabarti (2008)) where schools face loss of students to private schools (rather than public schools under NCLB). This literature finds positive effects of credible voucher competition.

However, this study is most closely related to a slowly-emerging but still relatively sparse literature on *No Child Left Behind*. Ballou and Springer (2008), Krieg (2008), Springer (2008), Neal and Schanzenbach (2010) study the effects of NCLB on the test score distribution—while students close to the center of the distribution are found to gain, there is no consensus relating to the effects on students at the ends of the distribution. Using NAEP data, Dee and Jacob (2011) find that students in states that had no previous accountability systems gained more. To identify the effects of NCLB, Reback et al. (2011) exploit the variation in state policies by which schools near the cutoff for meeting their own state's AYP requirements would have failed or passed if they were situated in other states. They find that NCLB lowered teachers' perceptions of job security, induced untenured teachers in high stakes grades to work longer hours, and had positive or no effects on low stakes tests in reading, math and science.

This paper has been greatly informed by this literature and builds on it. It differs from the existing literature in several important dimensions. First, exploiting the design of NCLB in general, and AYP in particular, it uses alternative regression discontinuity designs to identify the effects of AYP failure. Some key advantages of the RD strategies are that they serve to eliminate such factors as differences in pre-program trends, differences in observed and unobserved factors (often time-varying), and mean reversion that can potentially confound program effects. Second, in addition to looking at the distributional effects of AYP-failure, this study looks at the effects of AYP-failure on language arts, science and social studies (low stakes subject areas), students in low stakes grades, effects on score distributions of various student subgroups (both advantaged and disadvantaged), as well as test participation, attendance rates,

and graduation rates. Third, to have a closer look at the linkage between incentives and responses, I investigate whether schools that missed AYP by missing a certain criterion focused more on that criterion the following year and less on others. Finally, I also investigate whether competition mattered—that is, whether schools that faced more credible competition and hence a more effective threat of future loss of students responded more strongly in the objectives they had incentives in.

2 Program Details

The *No Child Left Behind* Act, a major reform of the Elementary and Secondary Education Act (ESEA), was signed into law on January 8, 2002. The law mandates implementation of a statewide accountability system and testing of all students in reading and math in grades 3-8. States are required to establish AYP targets, and schools are assigned an AYP-pass or AYP-fail status based on these criteria. The AYP-statuses are publicly available (online as well as sent home in the form of report cards) and AYP-failure is associated with negative publicity and visibility, and hence stigma and shame. In addition, Title 1 schools missing AYP targets for two or more consecutive years face ESEA sanctions. These sanctions start with two years of missed AYP and escalate with the number of years of missed AYP. A Title 1 school missing AYP for two consecutive years is required to provide public school choice to its students, and money follows students where they move. If the school misses AYP for three consecutive years, it is required to fund supplemental educational services in addition. If it misses AYP for four consecutive years, it is required to undertake corrective action in addition, and for five consecutive years restructuring in addition.

It is worth noting here that Wisconsin had an accountability system in the pre-NCLB era as well. Initiated in the 1997-98 school year, Wisconsin subjected its schools to an "annual review criteria". Schools were identified as "in need of improvement" in a subject area if a school failed to meet the annual review criteria for two consecutive years in one or more subject areas. However, this accountability system was much less salient that NCLB. It did not have any sanction associated with it (unlike NCLB). It also had less visibility in the sense that these ratings were not reported online or sent home in report cards, nor was there much media attention on them (unlike under NCLB).

The state tests in Wisconsin are known as the Wisconsin Knowledge and Concepts Examination (WKCE) and they have been given each year starting from 1997. They are given in five subject areas: reading, math, language arts, science and social studies. Till the 2004-05 school year, they have been given in grades 4, 8 and 10. Starting from the 2005-06 school year[3], the tests have been given in grades

[3] In the remainder of the paper, I refer to school years by the calendar year of the spring semester.

3-8 in reading and math, while the other three subject areas still continue to be tested in grades 4, 8 and 10. Based on scores, students are placed in four proficiency categories in each subject area—minimal, basic, proficient, advanced—minimal being the lowest category and advanced being the highest.

In accordance with the NCLB rules, there are four AYP objectives in Wisconsin. These objectives are known as the reading objective, the math objective, the test participation objective and the other indicator objective. The rules I outline here pertain to the school years 2003 and 2004 (which are exploited in this paper to construct the various alternative RD designs). The cutoffs changed later over the years since the schools were required to be 100% proficient by 2013-14.

According to the reading (math) objective, the percentage of students scoring at or above proficient in reading (math) was required to equal or exceed 61% (37%). The All Student group and each subgroup of sufficient cell size were required to meet these objectives. Eight subgroups (White, Black, Hispanic, Asian, American Indian, Limited English Proficient, Students with Disabilities, Economically Disadvantaged) in addition to the All Student group were held accountable. The cell size in Wisconsin was 40 students, except for students with disabilities where the cell size was 50.[4]

If the All Student group or a subgroup failed to meet the above reading or math cutoffs, the school could still pass the relevant AYP objective if it passed the "confidence interval" requirement which allowed it to score within a confidence interval of the cutoffs. If the reading and/or math objectives were still not met using the confidence interval rule, then "Safe Harbor" allowed the reading or math objectives to be met for the All Student group or a subgroup if the percentage of students not yet proficient was decreased by 10% from the prior year for that group/sub-group. In addition, for the Safe Harbor criterion to be applicable, a safe harbor step 2 or "Other Indicator" criterion for the All Students or that subgroup was required to be met.[5]

The test participation objective required at least 95% of the students in the All Student group and each subgroup of sufficient cell size to participate in the reading and math tests. The "other objective" criterion was attendance rate in elementary and middle schools, and graduation rate in high schools. It required schools of sufficient cell size to meet attendance rates of at least 84.9% in elementary and middle schools, and graduation rates of 81.75% in high schools.

[4] If the cell size of 40 was not met for the All Student group, then proficiency data from the previous year was combined with the current year. Thus combined, if the All Student group reached 40 or above, then it was held accountable for the reading and math objectives.

[5] While the law stipulated that this Safe Harbor step 2 "Other Indicator" should be school attendance or high school graduation, Wisconsin did not have disaggregated data for graduation and attendance for the period under consideration. So while these criteria were used for the All Student group, science proficiency rates were used for the subgroups' "Other Indicator".

The timing of the WKCE tests relative to the time when the schools first came to know of their scores and AYP ratings needs some discussion. In the 2002-03 school year, schools notified parents of the test results and individual performance reports on May 1, 2003, and the schools came to know of these results prior to that in April. Schools received notification of their AYP status in May, and the AYP statuses were publicly announced in June. On the other hand, the WKCE tests for the 2003-04 school year were held in November 2003. In 2004, schools notified parents in April and they came to know of the test results before that, but the timing of the AYP status notification, public announcement, and the testing window remained the same. It follows that there was not much time gap between the point when the schools first came to know of their performance (March/April) and the point when the tests were held (November). Correspondingly, the opportunities for schools to implement changes and those changes to take effect before the tests were limited. This should be kept in mind while interpreting the results. It should however be noted that benchmark exams and district assessments were held throughout the year which to some extent conveyed information to the schools relating to their ongoing performance in these related tests.

3 Data

In this study, I focus on the state of Wisconsin, as it has some unique advantages. Wisconsin is one of a very few states that has continued to test in low stakes subject areas (language arts, science, social studies) in addition to high stakes subject areas (reading and math) even after NCLB. Moreover, Wisconsin continued to administer its Wisconsin Reading Comprehension Test (WRCT) in its third grade, a low stakes test in a low stakes grade in the period under consideration. (The high stakes grades during this period were 4, 8, and 10.) These features enable me to study whether AYP failure under NCLB led to a shift away from low stakes subject areas to high stakes subject areas, and from low stakes grades to high stakes grades even in a high stakes subject area (reading).

The data for this paper consist of disaggregated school, grade, and subgroup level data, and are mostly obtained from the Wisconsin Department of Public Instruction. They include data on test scores, attendance rate, graduation rate, AYP status, socio-economic characteristics, per pupil expenditure, and school addresses for the school years 2002-03, 2003-04, and 2004-05.

Test score data include data on WKCE (grades 4, 8, 10) and WRCT (grade 3). WKCE data include data on percentage of students scoring in each proficiency category (minimal, basic, proficient and advanced), in each subgroup, in each subject area, and in each tested grade in 2003, 2004, and 2005.

Student scores in the WRCT were also classified into four categories (Minimal, Basic, Proficient and Advanced). Data were obtained on percentage of students scoring in each of the four categories and percentage of students tested in WRCT. Data on percentage of students tested in each WKCE subject area in each subgroup are also available for the tested grades. Attendance rate data are available for all schools, while graduation rate data are available for high schools.

AYP data include overall AYP statuses of schools as well as data on AYP statuses in each of the four component AYP objectives of schools. Data on socioeconomic characteristics include data on race and gender distribution of students, percentage of students eligible for free or reduced-price lunches, and real per pupil expenditure. Street addresses of schools are obtained from the Common Core of Data of the National Center for Education Statistics.

4 Empirical Strategy

4.1 Alternative Regression Discontinuity Designs

Simple comparison of schools that missed AYP with schools that made AYP will yield biased estimates of AYP failure. This is because AYP status is not randomly distributed among schools, and schools that missed AYP are likely to differ substantially in terms of both observed and unobserved characteristics from schools that made AYP. As described below, I exploit the institutional details of NCLB and the structure of the AYP formula to construct alternative regression discontinuity designs to study the causal effect of failing AYP.

In this study, 2002-03 is treated as the pre-program year, and I use data from this year to calculate the running variable in each of the RD designs below (unless otherwise stated). Recall that NCLB was signed into law in January 2002. Tests in Wisconsin were held in November 2002. But, importantly, details of the AYP formula were not yet worked out then. The AYP formula was debated and developed later in the school year. Consequently, the schools did not have knowledge of the AYP formula (or cutoffs) to manipulate their position on the AYP scale before November 2002.

4.1.1 'Minimum Distance' Regression Discontinuity Design

As a first pass, I use a strategy based on the minimum distance of a school from the relevant cutoffs. The intuition that guides this construct is that a school fails AYP even if it fails in one subgroup–criterion combination, while passing every other criteria. So what matters to a school (and the determining factor as far as AYP status is concerned) is the distance of its lowest performing subgroup–criteria from the cutoff. In other words, minimum of the distances of the various subgroup–criteria from the

relevant cutoffs determines how far the school is from making or missing AYP. Based on this argument, I characterize each school by the minimum of its distances of the various subgroup–criteria combinations from the relevant cutoffs.[6]

To construct the one-dimensional "minimum distance" measure, I use the following steps. First consider the reading and math objectives. Let p_{jkst} denote percentage of students scoring at or above proficient in subgroup j, subject k ($k \in \{reading, math\}$), school s and year t. Let C_k denote the cutoff in subject k. Subgroup j passes subject k if $p_{jkst} \geq C_k$. But even if this is not satisfied, the subgroup can pass if p_{jkst} exceeds the confidence interval adjusted cutoff (c_{jkst}), that is, if $p_{jkst} \geq C_k - \gamma_{jkst} = c_{jkst}$ where γ_{jkst} denotes the confidence interval cutoff. Note that even if the confidence interval adjusted cutoff is not met, the subgroup can make the subject AYP if it passes the safe-harbor condition *and* satisfies the corresponding qualifying safe harbor 2 ("Other Indicator") condition (as described in section 2), that is $p_{jkst} \geq (10 + 0.9p_{jks,t-1}) * I_1$, where I_1 is an indicator variable denoting whether or not safe harbor 2 is satisfied.

Now, consider the test-participation criterion. Each subgroup j passes the test-participation objective in year t if $Max_k\{TP_{jkst}\} \geq \bar{TP}$, where $\bar{TP}$ denotes the test-participation cutoff of 95% and TP_{jkst} denotes percentage of students tested in subgroup j, subject k, school s and year t. Finally, a school passes the "Other Indicator" objective if $OI_{st} \geq \bar{OI}$. For schools with a twelfth grade, OI_{st} denotes its graduation rate in year t. For schools without a twelfth grade, OI_{st} denotes the attendance rate of school s in year t. $\bar{OI}$ denotes the corresponding graduation rate or attendance rate cutoff. Taking all the indicators into account, r_{st} denotes the grand minimum of all distances from the respective cutoffs and constitutes the running or assignment variable.

$$r_{st} = min[P_{jkst} - min\{c_k, (10 + 0.9P_{jks,t-1}) * I_1\}, TP_{jst} - \bar{TP}, OI_{st} - \bar{OI}] \quad (1)$$

where $TP_{jst} = Max_k\{TP_{jkst}\}$. Figure 1A illustrates the relationship between assignment to treatment (that is, failing AYP) and schools' minimum distances from the cutoff (normalized to zero). As can be seen, there is a discontinuous change in the probability of treatment at the cutoff. Schools that lie to the left of the cutoff have a considerably higher probability of failing AYP than schools to the right of the cutoff.[7]

[6] For a similar characterization of the running variable, see Bacolod et al. (2009). Using a regression discontinuity analysis, they investigate the effect of California's accountability based financial awards program on resource allocation and academic achievement.

[7] The safe harbor rule was still under development during the period under consideration, especially because disaggregated data by subgroup on graduation and attendance were not yet available. Moreover, while AYP related to percentage of full academic year students above the various cutoffs, the available data relate to all students. These factors led to fuzzy discontinuities in the various designs.

Identification requires that the conditional expectations of various pre-program characteristics are smooth through the cutoff. Using a local linear regression technique with a triangular kernel and the Silverman rule of thumb bandwidth, the discontinuity estimates for various pre-program characteristics are presented in Appendix Table A1. Panel A reports discontinuity estimates for percentage of students at or above proficient in the five subject areas; panel B presents estimates for percentages of students tested in these subject areas; Panels C and D present results for socio-economic characteristics, attendance rate as well as number of subgroups that counted towards AYP; Panels E and F present estimates for indicator variables that indicate whether each of the eight subgroups mattered for AYP purposes in 2002-03. The discontinuity estimates are never statistically distinguishable from zero, except in two cases (% Hispanic and whether Hispanics counted) out of twenty eight cases. Note that with a large number of comparisons, one might expect a few to be statistically different from zero just by sheer random variation.

Following McCrary (2008), I also test whether there is unusual bunching at the cutoff. Using density of the running variable and the strategy above, I find no evidence of a discontinuity in the density function at the cutoff in 1999 (the discontinuity estimate is 0.03 and not statistically significant). The histogram in Figure 1B shows the distribution of the running variable. While there is no evidence of discontinuity in the density function at the cutoff, there is evidence of spikes in density at 3 and 5, especially at 5. A valid question here is whether these spikes pose a threat to the validity of the regression discontinuity design? These spikes are generated by the construction of the running variable, which in turn follows from the design of the AYP formula. Recall that the test participation cutoff was 95%. This implies that the minimum of the distances from the cutoffs of the schools to the right (who passed all criteria) can never exceed 5. There were a number of schools with test participation 100% and 98% which generated the spikes at 5 and 3 respectively.[8] So the heapings at 5 and 3 are artefacts of the AYP rule and not caused by manipulation at these points/spikes.[9]

[8] Note that, as can be seen from Figure 3, there were some schools to the right of 5 as well. These were the small schools (schools with less than 40 students in 2002-03, but at least 40 when 2001-02 and 2002-03 are pooled together) that were accountable only for the reading and math objectives, so their minimum distances were not constrained to be at 5 or below.

[9] I also look at the distribution of test participation the year before (i.e., in 2002) to investigate whether the spiky pattern at the upper end of the distribution is specific to 2003 only. As Figure 1C shows, the distribution of test participation in 2002 looks very similar with a spike at the 100% mark, confirming that the spiky pattern in 2003 is not specific to that year. But, as Barreca et al. (2011) point out, RD estimates can be biased if attributes relating to the outcomes predict heaping in the running variable, even if the heaping is away from the cutoff. In other words, these spikes might indicate composition bias that can serve as a potential threat to identification of the treatment effect at the cutoff. To investigate further, following Barreca et al. (2011), Appendix Figure A1 presents means plots of various pre-program characteristics against the running variable. The points 3 and 5 are highlighted with bigger markers and different colors. As can be seen, in each of the figures, the patterns through 3 and upto 5 are continuous. So the spikes do not appear to be problematic. However, note that some of the graphs show a jump right after 5. This is because, as mentioned in footnote 8 the points to the right of 5 pertain to the small schools that probably were different. Note that the jump seen in the case of the variable "Number of Subgroups" is an artefact of the AYP rule. For the regular schools, various subgroups in addition to the "All

4.1.2 A Closer Look at Incentives: Investigating the Impact of Math Induced Failure

While there is no threat to the validity of the RD estimates in the above "Minimum Distance" design, the interpretation of the estimates is less clear. By construction of the above running variable, AYP-failed schools were a heterogenous group in terms of their past year's individual criteria pass/fail histories. Schools to the left of the AYP-fail cutoff missed AYP in different criteria. For example, some schools missed AYP in reading, others missed in math or test participation or "other indicator".[10] These schools likely faced different incentives and responded in different ways. Therefore, it is not apriori clear how one might expect this heterogenous group to respond in the various objectives.

In the remainder of the paper, I look at incentives and responses more closely. For example, does math-induced failure lead to stronger response in math than in the other objectives? What about reading-induced and test-participation induced failures? These questions are all the more relevant as under NCLB a school must fail the same objective in consecutive years to receive sanctions. Consequently, a school that failed in the math (reading) objective would be expected to focus more on math (reading) the following year to escape consecutive failures in the same objective. The RD designs in this section and the ones below are geared to investigate whether this has indeed been the case.

To study the impact of math-induced failure, I compare the response of schools that just barely missed the math cutoff with those that just barely made it. In other words, my running variable here is the distance of the school's math score (percent of students at or above proficient in math) from the math cutoff in 2003. Note that this analysis does not constrain the AYP statuses in the other objectives, that is, these schools may have passed or failed in the other objectives. For validity of this design, the probabilities of AYP failures in the other objectives should be continuous through the math cutoff (a restriction I test below).

Figure 2A illustrates the relationship between the assignment to treatment (that is, failing AYP) and schools' distances from the math cutoff. As can be seen, there is a sharp discontinuity in the probability of AYP failure at the cutoff, and this discontinuity is solely due to math induced failure (as the running variable is based on math, and probabilities of AYP failures induced by the other objectives are smooth through the math cutoff as shown below). Figure 2B finds no evidence of any discontinuity in the density of the running variable at the cutoff. Online appendix Table B1 reveals that the pre-program characteristics are continuous through the cutoff. In fact, none of the thirty one discontinuity estimates

Student" group counted towards AYP. But for the small schools, only the "All Student" subgroup counted.

[10] In the above sample of schools, out of the schools in the bandwidth that fell to the left of the cutoff, 9% missed AYP in "other indicator", 32% in math, 38% in reading and 53% in test participation. 18% of these schools missed only the reading cutoff, 15% only math and 41% only in test participation, no school missed only "other indicator".

are statistically different from zero. Moreover, Panel G of Table B1 reveals that the probabilities of failures in the other objectives are indeed continuous through the math cutoff.

4.1.3 Investigating the Impact of Reading Induced Failure

To study the impact of reading-induced failure, I compare the response of schools that just barely failed the reading cutoff with those that just barely made it. The running variable here is the distance of the schools' reading score (percent of students at or above proficient in reading) from the reading cutoff in 2003. Figure 3A finds that there is a sharp discontinuity in the probability of AYP failure at the cutoff—the schools to the immediate left of the cutoff have a higher probability of failing AYP than the schools to the immediate right. Validity tests reveal that a rich set of pre-program characteristics (the same as those considered in online Appendix Table B1) including probabilities of AYP failures in the other objectives are continuous through the reading cutoff. (The estimates not reported for lack of space, but available on request.) Figure 3B does not find any evidence of discontinuity in the density of the reading running variable at the reading cutoff. Note that the discontinuity in Figure 3A is solely due to reading-induced failure, as (i) the running variable and the corresponding cutoff are based only on reading and (ii) the probabilities of failures in the other objectives are smooth through the reading cutoff.

4.1.4 Investigating the Impact of Test Participation Induced Failure

I also consider an alternative regression discontinuity strategy where I investigate whether schools that missed AYP by missing the test participation criterion focused more on test participation in the high stakes subject areas in the next year. The strategy is as follows. Consider schools that passed the reading, math and other indicator criteria. In this sample, I compare schools that just barely missed the test participation cutoff with schools that just barely made it.[11]

Figure 4A looks at the relationship between AYP status and the running variable. There is a sharp discontinuity at the cutoff. Schools to the left of the cutoff have a higher probability of failing

[11] It is worth noting here that a large mass of schools lie close to the test participation cutoff in terms of the distances of their test participation (percent tested) from this cutoff. On the one hand, this is an artifact of the AYP formula which stipulates the test participation cutoff to be 95% thus forcing the schools on the right hand side to lie within a distance of 5. On the other hand, schools falling to the left of the cutoff typically did not have very low test participation. The availability of a large mass around the cutoff enables me to condition on schools that passed in all other objectives. Note, though that the results remain qualitatively similar if I do not restrict the sample to schools that pass the other criteria. It is also worth noting here that the number of schools close to the test participation cutoff in this RD design is considerably larger than the number of schools close to the math (reading) cutoffs in the above two subsections where I consider the impacts of math and reading induced failure. This is also an artifact of the design of the AYP formula and the test participation cutoff as discussed above, and the fact that many more schools in reading and math fell closer to the ends (that is, were low scoring or high scoring).

AYP compared to schools to the right of the cutoff. I also test for the continuity assumptions—I find that both pre-program characteristics and the density of the running variable (Figure 4B) are indeed continuous through the cutoff (former not reported for lack of space, but available on request).

4.2 Estimation

Having established the validity of the above regression discontinuity designs, I next proceed to study the effect of AYP failure on the behavior of threatened schools using each of the above RD designs. As the identification graphs above show, I have fuzzy discontinuities in each case. I use a two stage least squares estimator to identify the local average treatment effect (LATE) at the cutoff (Hahn, Todd and van der Klaauw (2001))[12]. Consider the following model, where specification (2) denotes the first stage, and specification (3) the second stage.

$$AYPfail_{st} = \alpha_0 + \alpha_1 F_{st} + g(r_{st}) + \epsilon_{it} \tag{2}$$

$$y_{jks,t+1} = \beta_0 + \beta_1 AYPfail_{st} + h(r_{st}) + \xi_{it} \tag{3}$$

where $AYPfail_{st}$ takes a value of 1 if the school s failed AYP in year t and 0 otherwise, $F_{st} = 1(r_{st} < 0)$, r_{st} denotes the running variable for the corresponding design, $y_{jks,t+1}$ denotes outcome of group (or subgroup) j of school s in criterion k in year $t+1$. If $g(r_{st})$ and $h(r_{st})$ are continuous at the cutoff and the probability of treatment is discontinuous at the cutoff, then β_1 identifies the LATE at the cutoff and is given by the ratio of the discontinuity in the outcome variable to the discontinuity in treatment. As shown by Hahn, Todd and van der Klaauw (2001), this estimator is identical to a two stage least squares estimator of β_1 with $F_{st} = 1(r_{st} < 0)$ as the excluded instrument. I use a rule of thumb bandwidth as suggested by Silverman (1986) and a linear spline functional form for $g(r_{st})$ and $h(r_{st})$.

To test robustness of the results, I also experiment with alternative bandwidths, and alternative functional forms that include third order and fifth order polynomials as well as third order and fifth order splines.[13] The results remain qualitatively similar. I estimate alternate specifications that do not include controls as well as those that use controls (racial composition of schools, gender composition of schools, percentage of students eligible for free or reduced price lunches and real per pupil expenditure). Since the covariates are balanced on both sides of the cutoff (as established above), the purpose of

[12] It should be noted here that, as is common in the accountability literature that exploits RD designs (Chiang (2009), Rockoff and Turner (2010), Chakrabarti (2013a, 2013b), Rouse et al. (2013)), the impact estimates in this study are likely to be underestimates. This is because while the schools to the immediate right of the cutoff face a considerably lower probability of facing stigma/sanctions in the near future, they still face the threat to some extent as they are very close to the cutoff.

[13] I use odd order polynomials because they have better efficiency (Fan and Gijbels (1996)) and are not subject to boundary bias problems unlike even order polynomials.

including covariates is variance reduction. The estimates reported in this paper are obtained from specifications that include these controls, though the results do not depend on the inclusion/exclusion of controls.

5 Results: Analyzing the Impact of 'Threatened Status'

5.1 Using the Minimum Distance 'Regression Discontinuity' Design

High Stakes and Low Stakes Subject Areas in High Stakes Grades: Using the regression discontinuity strategy described above in section 4.1.1, Table 1 columns (1)-(5) look at the effect of "threatened status" on percentage of students scoring in minimal, basic, proficient, advanced and "at or above proficient" respectively in the five subject areas. The reading results show a rightward shift of the distribution with a fall in the percentage of students in basic and proficient categories (that is, just around the cutoff) in threatened schools, and a corresponding statistically significant increase in the percentage of students in the advanced category. These patterns are consistent with the hypothesis that the threatened schools chose to focus on students expected to score close to and around the cutoff.[14] Interestingly, language arts patterns mirror those in reading, even though language arts was low stakes. The skills required in language arts are likely similar to those required in reading, which might have led to such spillover effects. Note that there is no evidence that the improvements of the marginal students in reading and language arts came at the expense of the non-marginal ones (except for a small statistically insignificant increase in percent of students in the lowest performing category in reading).[15] There is no evidence in favor of improvements in either math, science or social studies.[16]

Since the schools close to the cutoff may have passed/failed in different objectives, I also estimate alternate specifications that control for four indicator variables that respectively capture whether or not the school fell to the left or right of the cutoff in the corresponding criteria to explicitly control for the differences in pass/fail statuses in different objectives: $I_{k,s} = 1[min_j\{P_{jkst} - min\{c_k, (10 + 0.9P_{jks,t-1}) * I_1\}\} < 0]$ $where\ k = \{reading, math\}$; $I_{TP,s} = 1[min\{TP_{jst} - \bar{TP}\} < 0]$; $I_{OI,s} = 1[min\{OI_{st} - \bar{OI}\} < 0]$.

[14] Since it might have been difficult to precisely target students who would score just below the proficiency cutoff, they might have increased their attention towards students expected to score around the proficiency cutoff thus leading to decreases in percentages of students in the basic and proficient categories and an increase in the advanced category.

[15] Of note here is that, as can be seen in online Appendix Table B4, the improvement of AYP-failed schools in reading was driven essentially by improvement in reading in elementary schools (grade 4) and not in middle and high schools (grades 8 and 10). Reading in elementary grades is less costly to improve in and does not require a lot of additional investments, either in terms of money or personnel (unlike middle and high school), which may have led to this pattern.

[16] It should be noted here that the changes in Table 1 (as well as in the tables that will follow) are net changes. For example, consider the "proficient" category in reading. It is possible that some students moved from the lower categories to "proficient" category. If this did happen, then the actual fall in proficient category is even larger than that suggested by the estimates.

The results from this re-estimation are qualitatively similar (and available on request). Next, I use an alternative strategy where in addition to F_{st}, I use these indicators for cutoffs missed as additional instruments for AYP failure. Using these instruments, Table 1 columns (6)-(10) investigate the effect of AYP failure on percentage of students scoring in the various proficiency levels in the five subject areas. The results remain qualitatively similar to above.[17]

While the above analysis looks at the effects of AYP failure on school scores, I also investigate the effect of AYP failure on the distribution of scores in three subgroups—Whites, economically disadvantaged and special education (online appendix Tables B2 and B3). Online Appendix Table B2 finds evidence in favor of deterioration in each of the subject areas for both subgroups—Economically Disadvantaged and Special Education—especially the former. This is in spite of the inclusion of these subgroups in AYP formula with a stated objective of improving their performance. In stark contrast to the patterns for these two subgroups, the patterns for Whites (online Appendix Table B3 columns 1-5) suggest rightward shifts in the distributions of not only the high stakes subject areas but also the low stakes ones, although the effects are not always statistically significant.[18]

Effect on Low Stakes Grades: The above analysis looks at the effects on high stakes WKCE reading and math tests, as well as tests given in low stakes subject areas (language arts, science, social studies) to the *same cohort of students*. In contrast, WRCT was a reading test given in a low stakes grade (grade 3), that is, students in that grade did not face the high stakes tests. Interestingly, Table 2 Panel A

[17] The above results reveal that although math was a high stakes subject area, there is no evidence of improvement in math. It might be worth thinking a little bit here why this might have been the case, especially because the literature has often found larger effects in math (than reading) in other settings. First, note that in the sample of schools that fell in the bandwidth, the schools close to and on the left of the cutoff were actually more likely to be reading failures than math failures (footnote 10), which might have contributed to this pattern. But, at the risk of foreseeing upcoming results, the reading improvement caused by reading-induced failure later in this paper is quantitatively somewhat larger than the math improvement caused by math-induced failure. Note that this pattern to some extent may have been contributed by the setting (Wisconsin) and the WKCE test. Of note here is that Wisconsin schools performed considerably worse in math compared to reading in the pre-program years. (In the immediate pre-program year (2002-03), in the sample of schools that fell in the bandwidth, 5.8% of students scored in minimal in reading compared to 12.3% in minimal in math; 46% of students scored in advanced in reading in the same year in contrast to 28% in math. In 2001-02 (in the same sample), 4.5% of students scored in minimal in reading while 12% did so in math; 45.6% scored in advanced in reading in the same school year while 29% did so in math.) It is also relevant to note here that other studies in the context of other educational interventions in Wisconsin found larger impacts in WKCE reading than in WKCE math (Witte et al. (2004) in the context of Wisconsin charter schools; Witte et al. (2012) studying the impact of the Milwaukee voucher program on choice students in 2010; Chakrabarti (2008) studying the impact of the Milwaukee voucher program on public school responses).

[18] Note that the differences in results between the Whites on the one hand and the economically disadvantaged and special education groups on the other are not driven by differences in samples. Online Appendix Table B3 columns 6-10 re-estimate the impact on Whites constraining the sample only to schools where the economically disadvantaged subgroup satisfied minimum cell size. This essentially is the sample used in online Appendix Table B2 columns 1-4, yet the results for Whites remain qualitatively similar to those obtained from the unconstrained sample in online Appendix Table B3 columns 1-5 (except for Language Arts). (The differences in the number of observations (between online Appendix Tables B2 columns (1)-(4) and B3 columns (6)-(10)) are because of two reasons (i) Whites did not make cell size in some of these schools and (ii) their scores were missing in a few cases.)

finds strong evidence in favor of a leftward shift in the WRCT distribution. This suggests that while the threatened schools tended to focus more on WKCE reading in the high stakes grades, this seems to have come at the expense of reading performance in the low stakes grade 3. Column (6) finds that there is no evidence of any positive effect on test participation in third grade WRCT.

Test-Participation: Table 2 Panel B finds that there is no evidence of improvements in test participation in any of the subject areas (high stakes or low stakes). Nor is there any evidence of improvement in the AYP test participation criterion. Of note here is that while the effects are negative across the board (often small and never statistically significant), these negative effects are always economically stronger in the low stakes subject areas, with this negativity being the most prominent in social studies.

Attendance and Graduation Rates: Table 2 Panel C looks at the effect of AYP failure on attendance and graduation rates. Column (1) finds that there is no evidence of any positive effect on attendance. While this column includes all schools, column (2) looks at the effect on attendance constraining the sample only to schools where attendance counted in AYP. Interestingly, the effect on attendance is more positive in this column (though still small and not statistically significant). These findings suggest that the threatened schools for whom attendance mattered may have focused more on attendance relative to schools for whom attendance was not high stakes. There is also some evidence in favor of positive effect on graduation, although the effect is not statistically significant.

5.2 Impact of Math Induced Failure

High and Low Stakes Subject Areas in High Stakes Grades, and Test Participation: In this section, I analyze the impact of math-induced failure using the strategy described in section 4.1.2. For context and for easier interpretation of the estimates that follow, Appendix Table A3 Panel A presents summary characteristics (socioeconomic characteristics, % scoring at or above proficient in the various subject areas, real per pupil expenditure, number of accountable subgroups, attendance rate, % of schools scoring below the various cutoffs) of this sample at baseline (2003). Approximately 41% of the schools in this sample scored below the math cutoff, 43% below the reading cutoff, 14% below the test-participation cutoff, 14% below the attendance cutoff, and 32% of the high schools scored below the graduation cutoff.

Table 3 Panel A presents the impact on the distribution of WKCE test scores and test participation in the various high stakes and low stakes subject areas in high stakes grades. It shows a sharp rightward shift in the distribution in math, suggesting that math-induced failure indeed led to an increase in emphasis in math. This is in contrast to the results in Table 1 where schools close to the cutoff in terms of multiple objectives did not show any evidence in favor of improvement in math. Table 3 Panel A

also reveals that while math-induced failure led to improvements in math performance, there is no clear evidence of effects in any of the other subject areas, including high stakes reading. In fact, there is some evidence of deterioration in reading with a leftward shift in the distribution, but the effects are not statistically significant. Nor is there any evidence of positive effects on test participation in any of the subject areas. Figure 5 plots the post-treatment outcomes in math (Panel A) and reading (Panel B) as a function of the running variable (distance from the math cutoff)[19]. Note that these are intent-to-treat estimates, while the table reports treatment-on-treated estimates. As can be expected, these estimates are consistent with those reported in the table.

Low stakes grades, Attendance and Graduation: Table 3 Panel B columns (1)-(5) look at the effect on the distribution of third grade WRCT scores, while column (6) looks at WRCT test participation. There seems to have been a deterioration in both WRCT performance and test participation, but none of the effects are statistically significant, likely because of small sample size.[20] Also, column (7) finds no evidence of any effect of math-induced failure on the "other indicator" (attendance/graduation).

5.3 Impact of Reading Induced Failure

High and Low Stakes Subject Areas in High Stakes Grades, and Test Participation: This section examines the impact of reading-induced failure using the RD design described in section 4.1.3. Appendix Table A3 Panel B presents baseline (2003) characteristics of this sample. Approximately 49% of the schools scored below the reading cutoff, 27% scored below the math cutoff, 13% scored below the test-participation cutoff, 9% scored below the attendance cutoff, and 26% of high schools scored below the graduation cutoff.

Table 4 presents the impact estimates for reading-induced failure. Interestingly, there is strong evidence of improvement in reading with a sharp rightward shift in the distribution. Also, notably (and interestingly), the effect in reading is much stronger than that in Table 1. Specifically, the percent of students at or above proficient in reading, the measure that mattered for AYP purposes, shows a considerably larger increase here. There is a sharp decline in the percentage of students scoring in "basic" in the threatened schools suggesting that the schools tended to focus more on students expected to score just below the high stakes cutoff "proficient". There is no evidence that the improvements of the marginal students in reading came at the expense of the ends (except for a small statistically insignificant

[19] Plots for the other subject areas and for the other RD designs are not reported here for lack of space.

[20] Note that the WKCE analysis in Panel A includes schools that have either elementary, middle or high school grades. In contrast, Panel B includes schools that have elementary grades (and drops schools that have only middle and/or high grades) and also have WRCT scores, which explains the drop in sample size. More specifically, 57 of the 130 schools in the bandwidth in Panel A are elementary schools, and 47 of them have WRCT scores (and are included in Panel B).

increase in percent of students in the lowest performing category in reading).[21] In language arts, there seems to have been some movement from lower performance categories to "proficient"—this may have been contributed by spillover effects from reading. Note, though, that there is also some evidence of decline in the percent of students in the highest performance category "advanced".

Table 4 also shows that reading-induced failure led to a sharp decline in the performance in math, with a left-ward shift in the distribution. There is also strong evidence of sharp declines in both science and social studies.[22] Apart from some evidence of an increase in test participation in language arts, there is no evidence of effects on test participation in any of the other subject areas.

Low stakes grades, Attendance and Graduation: Panel B columns (1)-(6) investigate the effect of threatened status on performance and test participation in the low stakes reading test WRCT given in the low stakes grade 3. The estimates show that AYP failure led to a negative effect on performance of third graders in WRCT—there was a clear leftward shift of the WRCT distribution. Most of the effects are highly statistically significant and precisely estimated, in spite of the small sample size.[23] Interestingly, this negative effect is despite the fact that the threatened schools showed improvements in reading in the high stakes grades. The results suggest that the increased focus on high stakes reading in high stakes grades may have led to a shift of emphasis away from reading in low stakes grades. Panel B column (7) shows that there is no evidence that reading induced failure led to positive effects on the "other indicator" objective (attendance/graduation).

It might be worth summarizing the patterns obtained in the last two sections. There is strong evidence that math-induced failure led to marked improvements in math, while reading-induced failure led to strong positive effects in reading. In contrast, there is no clear evidence of improvement in any of the other subject areas (high stakes or low stakes), nor is there any evidence of improvement in the other objectives, such as test participation, attendance or graduation. Also, interestingly, while reading induced failure led to improvements in high stakes reading in high stakes grades, it seems to have come at the cost of reading in low stakes grades. These results suggest that the schools responded strategically according to incentives focusing on the subject area that caused them to fail AYP in an effort to escape

[21] This pattern of rightward shift of the whole distribution (in contrast to gain of marginal students at the expense of the ends) is similar to findings in the context of some other accountability programs (Ballou (2008), Chakrabarti (2013a), Springer (2008)). This pattern is suggestive of an increase in efficiency after the program, implying that these schools were possibly less than fully efficient in the pre-program period.

[22] Online Appendix Table B5 shows that the impacts of math-induced and reading-induced failure on reading and math are robust to alternative bandwidths and to inclusion/exclusion of covariates. The other estimates are also robust to such sensitivity analyses, but are not reported here for lack of space.

[23] Refer to footnote 20 for reasons behind the drop in sample size in Panel B (relative to Panel A). Out of 138 schools included in Panel A, 59 are elementary. Of these, 53 have WRCT scores and are included in Panel B.

a second AYP failure in the same objective and correspondingly more dire consequences.

5.4 Impact of Test Participation Induced Failure

Based on the RD design outlined in section 4.1.4, Table 5 investigates the effect of threatened status on score distribution in the five subject areas as well as test participation. Interestingly, now the evidence in favor of improvements in reading and math found earlier disappear—in fact, there is a small negative (statistically insignificant) effect on percentage of students scoring at or above proficient in reading (the statistic that matters for AYP computation purposes for reading). There is not much evidence of effects in any of the other subject areas and the effects are never statistically significant. Table 5 column (6) investigates the effects on test participations in the various subject areas. There are small positive effects on test participations in reading, language arts, and math in the threatened schools, but they are not statistically significant. The effects on test participation in the other subject areas are smaller (though still positive) and never statistically significant.

5.5 Are Differences in Subgroup Accountability Driving Results?

Recall that under NCLB subgroups are accountable only if they pass the minimum cell size requirement. The composition of accountable subgroups in schools is likely to affect results. For example, schools that have many weaker subgroups accountable are likely to have a harder time improving relative to schools that don't, other things equal. So, an important concern is whether the composition of accountable subgroups is similar for schools just below and above the cutoff, and if the results above are biased by differences in subgroup composition.

First, I find no evidence of discontinuities in the number of accountable subgroups at the cutoffs for any of the RD designs (Table A1 column 18 and Table B1 column 18 for the first two designs, others available on request). To explore further, I construct indicator variables corresponding to each subgroup that represent whether the corresponding subgroup met the designated cell size. None of these "whether subgroup counted" variables show evidence of any discontinuity at the cutoffs (Table A1 columns (21)-(28) and Table B1 columns (21)-(28) for the first two designs, others available on request), except Hispanics for the minimum distance criterion design. Finally, I also re-run the above regressions, now explicitly controlling for each of the "whether subgroup counted" variables. The results remain qualitatively similar. The results for the minimum distance design are presented in online Appendix Table B6, the others are available on request. So, it does not seem that differences in accountable subgroups are driving the above results.

5.6 Are Compositional Changes or Sorting Driving Results?

The effects obtained above might be biased if AYP failure led to differential changes in composition or sorting in these schools. Note that public school choice came into effect only if a school missed AYP two years in a row, so public school choice changing student composition is not a concern here. However, a failing grade might induce some students to move away from these schools. The existence of such phenomenon can confound the results obtained above.

To investigate whether sorting might have driven the results above, I first examine whether the demographic composition of the treated schools saw a relative shift in 2004. The results are reported in Table 6. Panel A reports results for the sample of schools under the "minimum distance" design and Panel B for the sample of schools under the "missed/made math cutoff" design. (The results corresponding to the other designs are similar and are available on request.) As columns (1)-(7) in both panels show, there is no evidence of any effect on any of the demographic variables except on percentage Hispanic in Panel A. Note that this pattern for percentage Hispanic for the "minimum distance design" is very similar to the corresponding patterns in the pre-program demographics (Table A1) where percentage Hispanic was the only statistically significant variable. In fact, the discontinuity in the percentage Hispanic variable here is economically and statistically similar to that in the pre-program year.

Further, Table 6 also looks at the effect of AYP failure on real per pupil expenditure, number of accountable subgroups (that is, number of subgroups that made minimum cell size), and accountability of individual subgroups (that is, whether the program led to increase or decrease in the likelihood of some subgroups getting counted in the threatened schools). The intuition here is that any perceptible sorting or change in demographic composition will likely get reflected in these variables. Once again, there is no evidence of any effect on these variables, except on the accountability of Hispanics (whether Hispanics counted) in the "minimum distance" design. Again, this pattern for Hispanics in the "minimum distance design" is similar to that in the pre-program scenario (Table A1). Also, as noted above, with a large number of comparisons, one would expect a few to deviate statistically from zero just by random variation. So AYP failure does not seem to have led to shifts in these variables in the treated schools in 2004—it follows that it is unlikely that the results above are driven by sorting.

5.7 Are the Results being Driven by SIFI schools?

Recall that while adequate yearly progress ratings under NCLB were given for the first time in the 2002-03 school year, Wisconsin had an accountability system (though much less salient) that preceded NCLB. Under this prior accountability system, schools had to meet an annual review criteria, as outlined

in section 2. Schools that failed to meet the annual review criteria for two consecutive years in one or more assessed subject areas were designated as "in need of improvement" (SIFI or "schools identified for improvement"). While deciding on the SIFI ratings for 2002-03, Wisconsin took into account the ratings in the previous years under the prior accountability system—schools that failed to meet the annual review criteria the year before and failed AYP in 2002-03 were taken to have missed AYP for two consecutive years, and designated SIFI level 1. Similarly the SIFI ratings of the previous years were taken into account to determine the rating of 2002-03 school year. Schools that fell into SIFI status in 2002-03 based on current year and previous years' ratings faced NCLB sanctions (if they were Title 1 schools) and/or an elevated stigma. Since the incentives of schools that were identified to be SIFI based on previous years' ratings might have been different from those that missed AYP for the first time in 2002-03, in this section I drop these SIFI schools and re-estimate the above effects.

The results from this analysis after dropping the SIFI schools are presented in Appendix Table A2. Panel A uses the sample of schools that just missed or made the math cutoff based on the design outlined in section 4.1.2, while Panel B uses the sample of schools that just missed or made the reading cutoff based on the design outlined in section 4.1.3. The results remain qualitatively similar to those obtained above (Tables 3 and 4 respectively). Once again math induced failure led to improvements in math with a clear rightward shift in its distribution. There is no evidence that math-induced failure led to improvements in reading or test participation in any of these subject areas. Similarly, as earlier, reading induced failure led to improvements in reading, but deterioration in math. There is no evidence of effects of reading-induced failure on test participation in these subject areas.[24] These results suggest that the SIFI schools did not respond in a very different way from the schools that missed AYP for the first time in 2002-03. This is likely because for both types of schools future sanctions or worse designations kicked in only if they failed the same objective in the next year. Therefore, both groups of schools had strong incentives to improve in the criterion they failed in.

5.8 Testing Robustness of the Results to additional years

In this section, I use an additional year of post-NCLB data and investigate whether the results are robust to this inclusion. An advantage of this analysis is that pooling of multiple post-program years lead to a larger sample size. It is important to note here that there were important changes to the WKCE in the 2005-06 school year, and it was also calibrated to a different scale in this year. As a result, I limit myself

[24] The results for the other subject areas are not reported to save space, but they remain qualitatively similar as earlier and are available on request.

to data through the 2004-05 school year. The purpose of this section is to analyze the impact of AYP failure in year T (where $T = \{2003, 2004\}$) on outcomes in the following year, $T + 1$. I use the various RD strategies outlined in section 4.1 for this analysis, but to save space I only report the impacts of math-induced failure and reading induced failure.

Using data from 2003 and 2004 and exploiting the AYP criteria in these years, Figure 6A illustrates the relationship between AYP failure and distance from the math cutoff, while Figure 7A illustrates the relationship between AYP failure and distance from the reading cutoff. Both figures show sharp discontinuities at the respective cutoffs. These discontinuities are induced by math-induced failure in Figure 6A and reading induced failure in Figure 7A. Figures 6B and 7B respectively show that there is no evidence of any discontinuity in the density functions of the running variables at the corresponding cutoffs. Further there is no evidence of discontinuities in pre-existing observable characteristics at the respective cutoffs; nor is there any evidence of discontinuities in the probabilities of AYP failures induced by the other objectives at these cutoffs (online Appendix Table B7 for "missed/made math cutoff" sample; estimates for "missed/made reading cutoff" sample available on request). An important additional consideration here is whether the 2003 AYP pass/fail history for the 2004 schools is smooth through the 2004 cutoff. Online Appendix Table B7 Panel H shows that this is indeed the case for the "missed/made math cutoff" sample—there is no evidence of any discontinuity in the various measures of their 2003 AYP status (overall AYP, reading AYP, test participation AYP, other indicator AYP) at the 2004 cutoff. (The results for the "missed/made reading cutoff" sample are similar and available on request.)

Having established the validity of the regression discontinuity design, I next look at the impact of math-induced failure and reading-induced failure (Table 11 Panels A and B respectively). The results remain qualitatively similar to those obtained above (Tables 3 and 4). Once again, math induced failure led to improvements in math with a rightward shift of the math distribution, while reading induced failure led to improvements in reading. Also, similar to the results above, reading induced failure led to a deterioration in math. There is some evidence that math induced failure led to deterioration in reading, but the effects are not statistically significant.[25] These findings increase confidence on the results obtained above.

[25] The impacts on the other subject areas are also similar to above, and are available on request.

6 Incentives and Responses: Credibility of Threat and the Role of Competition

In this section, I analyze whether the responses of the AYP-failed schools depended on the extent of competition they faced. The idea is to investigate whether AYP-failed schools that had more AYP-passed schools in their near vicinity exhibited larger responses. Using street addresses of schools, I geocode public schools and find the number of AYP-passed schools within a certain radius of each public school.[26] I consider the number of AYP-passed schools as a measure of the extent of competition because, first, the public school choice provision under ESEA makes students eligible to transfer to AYP-passed schools only. In addition, if stigma is the motivating factor for student moves, students will likely choose to move to AYP-passed schools rather than to another AYP-failed school. So the density of AYP-passed schools can be taken as a measure of competition.

6.1 Competition Response in the "Minimum Distance" Sample

This section uses the minimum distance sample, and investigates whether the threatened schools in this sample responded differently to competition. The results from this analysis are presented in Table 7. Panel A shows that AYP-failed schools that had more AYP-passers in their near vicinity experienced a larger shift of their reading distribution to the right. Specifically, they exhibited larger net moves from lower performance categories (minimal and basic) to proficient, and also exhibited economically and statistically larger net moves into the key "at or above proficient" category. These patterns were mirrored in language arts as well, possibly due to spillover effects.

What is perhaps more interesting is that not only in reading, but the competition effects appear in math as well—AYP-failed schools that had more schools in their near vicinity show a relative shift of their distributions to the right in math. They exhibit larger net falls in percentage of students around the cutoff and larger net increases in the advanced category. While competition effects are not very clear in science, they again show up in social studies.

Panel A column (6) looks at the competition effects in test participation, while Panel B looks at the effects on attendance and graduation rates. AYP-failed schools that faced more competition exhibited economically and statistically larger improvements in test participation, attendance and graduation rates.

The above analysis suggests that competition did indeed matter in this sample. While in general there was not a strong response from the AYP-failed schools and the response was mostly limited to reading,

[26] The results presented in this study pertain to a one mile radius. But I have also experimented with 2 mile, 3 mile and 5 mile radii—the results remain qualitatively similar and are available on request.

AYP-failed schools that faced more competition responded more broadly and strongly. Their increased response was not only limited to reading, but was also seen in math, test participation, attendance, and graduation rate. The weak response of the threatened schools in general was most likely due to lack of adequate competition and inadequate perception of threat.[27] The strong response *in all criteria* exhibited by the AYP-failed schools that faced more competition is striking. In the "minimum distance" RD design, schools to the left of the cutoff were close to the cutoff in a variety of objectives (by definition) and may have chosen to respond in the corresponding objectives facing competition, thus yielding strong responses in these different objectives. In the two sections that follow, I look at more targeted failures (where failure was math-induced and reading-induced respectively)—it would be interesting to see whether the competition responses (if any) were focused on the corresponding subjects.

6.2 Competition Response of Math-induced failures

Table 8 Panel A uses the sample of schools that just missed or made the math cutoff. Interestingly, math-induced failures that had more AYP-passers in their near vicinity exhibited larger shifts of their math distributions to the right. Specifically, these schools exhibited larger and statistically significant increase in the percent of students scoring at or above proficient, the metric that mattered the most for AYP calculation purposes and governed their sanction/stigma status the next year. In contrast, there is no evidence of any differential response of these schools in any of the other subject areas.

6.3 Competition Response of Reading-induced failures

Table 8 Panel B explores how competition affected the responses of Reading-induced failures. Reading induced failures that had more AYP-passers in their near vicinity responded more strongly in reading with a larger rightward shift in the reading distribution. This was associated with larger net moves from the minimal category to basic and especially proficient category, and a correspondingly large net increase in the key "at or above proficient" category. Interestingly, the patterns in language arts are once again similar to reading, likely due to spillover effects. There is no evidence of any differential response in math. There is some evidence of stronger rightward shifts in science and social studies distributions for reading induced AYP-failures that faced more competition. A possible reason is spillovers from reading with better reading skills helping science and social studies performance, but most of these effects are not statistically significant.

[27] This is supported by the fact that, on average, the distribution of AYP-passed schools was pretty sparse around the AYP-failed schools. For example, in 2003-04, 85% (65%) of the schools had no higher performing school within its one (two) mile radius, 10% (12%) had one, 3% (7%) had 2 and 1% (4%) had three.

The findings in the last two sections can be summarized as follows. Math-induced AYP-failures and reading-induced AYP-failures that faced more competition responded more strongly in the subject area that generated this failure (math and reading respectively). Apart from evidence of some spillover effects from reading to language arts (and to some extent to science and social studies, though often not statistically significant), there is no evidence of effects in the other areas. This suggests that AYP-failed schools that faced more competition strove to improve performance in the criterion that caused the failure, presumably in an effort to escape a second AYP-failure in the same criterion. Recall that a second AYP-failure in the same criterion would trigger sanctions and further stigma.

6.4 Testing Robustness of the Competition Responses

6.4.1 Is the Balance in Socio-Economic Characteristics Satisfied Regardless of the Extent of Competition?

In this section, I investigate whether the balance in pre-program socio-economic characteristics at the cutoff is maintained for different values of competition. Note that for the validity of the above RD designs, it is essential to have balance in pre-program characteristics irrespective of the extent of competition.

Table 9A investigates whether this was indeed the case. Panel A uses the sample of schools that just barely missed/made the math cutoff, while Panel B uses the sample of schools that just barely missed/made the reading cutoff.[28] In each panel, the first row of results uses pre-program socio-economic characteristics as dependent variables. The coefficients of the interaction terms between AYPfail and competition ("count") are always small and never statistically different from zero. Thus, there is no evidence in favor of any imbalance in pre-program socio-economic characteristics irrespective of the extent of competition. This testifies to the validity of the RD design for all values of "count".

6.4.2 Did Schools Facing More Competition Have Different Sorting?

While there is no evidence of different pre-program demographics in AYP-failed schools that faced more competition, AYP-failed schools that faced more competition might have faced differential sorting in the post-program period. If so, this potentially could have driven the effects seen in Table 8, in spite of the fact that there was no evidence of differential sorting in the total group of AYP-failed schools (Table 6). The second row of results in Table 9A Panel A and Panel B respectively investigates this issue for math-induced and reading-induced failures. Specifically, I investigate whether there was differential sorting in the AYP-failed schools that faced more competition. For this purpose, I use post-program

[28] In this section and the ones that follow, I focus on the impacts of competition on math-induced and reading-induced failures. The results for the other designs are similar and are available on request.

(2004) demographic variables as dependent variables in the second row of each panel, and investigate whether the AYP-failed schools that faced more competition showed a relative shift in their demographic composition in 2004. I find no evidence in favor of such differential effects.

6.4.3 Does Controlling for Interactions of Demographics and Competition Affect Results?

In this section, I include interactions of competition ("count" or number of AYP-passers in near vicinity) with each of the demographic variables and percentage of students eligible for free/reduced price lunches to investigate whether such inclusion affects results.

Table 9B presents results from this analysis—Panel A focuses on math-induced failures while Panel B focuses on reading induced failures.[29] Once again, the table finds that math-induced AYP-failures that faced more competition responded more strongly in math. In contrast, there is no evidence that competition led these schools to respond more strongly in reading. Similarly, reading induced AYP-failures that faced more competition responded more strongly in reading (with some evidence of spillover effects to language arts), while there is no such evidence in math. In fact, the effects remain very similar to those obtained in Table 8, thus giving confidence to the results above.

It is worth mentioning here that this analysis further confirms the validity of the above RD design. If geographic location and demographics are both continuous through the cutoff, as they should be in a valid RD design (and are here), then the interaction of demographics and competition should also be balanced at the cutoff, and inclusion of such interactions should not affect results. This indeed is the case here.

6.4.4 Did Competition Affect Resource Allocation?

In this section, I investigate whether AYP-failed schools that faced more competition availed of more resources. For example, one can think of a scenario where districts allocated more resources to AYP-failed schools that faced more competition in an effort to help them improve because it is these schools that faced a credible threat of losing students rather than their counterparts in less competitive areas.

Using real per pupil expenditure as the measure of school resources, Table 10 investigates whether this was indeed the case. The first column uses the sample of schools that just missed/made the math cutoff, while the second column uses the sample of schools that just missed/made the reading cutoff. There is no evidence that the AYP-failed schools that faced more competition had higher real per pupil

[29] In this table, I focus on impacts on reading and math for math-induced failure, and reading, language arts, and math for math induced failure. The results for the other subject areas are also similar to those obtained in Table 8 and are available on request.

expenditures than their counterparts in less competitive areas. This is true for both math-induced failures and reading-induced failures. It follows that there is no evidence that differential resource allocation in AYP-failed schools that faced more competition may have caused the results obtained in Table 8.

These sensitivity analyses show that the competition impacts obtained above are reasonably robust. They increase confidence in the finding that AYP-failures facing more competition indeed responded more strongly, and perhaps more interestingly, focused more on the subject area that induced failure and where a further failure would trigger more serious consequences (sanctions and further stigma).

7 Conclusions

In this paper, I study the incentives and responses of schools that failed AYP once. Exploiting the institutional details of the program, I use alternate regression discontinuity methods to analyze these effects. I find strong evidence that the threatened schools responded according to incentives. Math-induced failures showed marked improvements in math, while reading-induced failures showed strong improvements in reading. Consistent with incentives, there is not much evidence that these schools responded in the other high stakes criteria. The various RD designs show that, in general, the AYP-failed schools tended to focus more on students expected to score just below or around the high stakes proficiency cutoffs in the subject areas they showed improvements in. But, interestingly, this gain of the marginal students did not come at the expense of the ends, rather there was a rightward shift in the corresponding distributions. Results for Language Arts reveal patterns similar to reading, suggesting spillover effects from reading. In contrast, there is no evidence of improvement in the low stakes subject areas, science and social studies.

Unlike state accountability systems that preceded NCLB, NCLB included subgroup scores in AYP formation with a declared objective to prevent weaker subgroups from falling through the cracks. However, unfortunately, in spite of this, there is evidence in favor of deterioration in performance for weaker subgroups (Economically Disadvantaged, Special Education) in all subject areas. In contrast, performance effects for Whites not only show improvements in the high stakes subjects (reading and math), but also in the low stakes ones (language arts, science, social studies). Interestingly, while there is no effect on attendance in the group of threatened schools where it did not count, the effects are more positive for threatened schools where it did count.

While there were notable improvements in high stakes reading in the schools that faced incentives in reading, the picture was quite different for reading in low stakes grade 3. In fact, threatened status (even

when it was caused by reading-induced failure) led to a sharp deterioration of reading performance in grade 3. This suggests that the increased focus on high stakes students may have come at the expense of low stakes ones.

I also analyze the role of competition in the incentives and responses of the AYP-failed schools. The results are revealing. Recall that in the "minimum distance" design, while the AYP-failed schools responded in reading, there was no statistically significant evidence of improvements in the other criteria. In contrast, I find that the AYP-failed schools in this sample that faced higher competition responded both more strongly and broadly, with statistically significant improvements noted in all high stakes criteria. This is likely because in this design, by definition schools that barely missed the cutoff faced incentives in a variety of high stakes criteria. The results are even more telling for the latter RD designs where the incentives were more precise. I find that the math-induced (reading-induced) AYP-failures that faced more competition responded much more strongly in math (reading). In contrast, there is no evidence that these schools responded any more strongly in the other high stakes subject areas. These results strongly suggest that competition and the credibility of consequences mattered—AYP-failed schools that perceived credible threats responded considerably strongly in the objectives they had incentives in.

8 References

Bacolod, Marigee, John Dinardo, and Mireille Jacobson (2009),"Beyond Incentives: Do Schools use accountability Rewards Productively?," NBER Working Paper Number 14775.

Ballou D., and M. Springer (2008),"Achievement Trade-Offs and NCLB," Urban Institute.

Barreca, Alan, Jason Lindo, and Glen Waddell (2011),"Heaping-Induced Bias in Regression-Discontinuity Designs," NBER Working Paper Number 17408.

Chakrabarti, Rajashri (2008),"Can Increasing Private School Participation and Monetary Loss in a Voucher Program Affect Public School Performance? Evidence from Milwaukee," Journal of Public Economics, 92 (5-6), 1371-1393.

Chakrabarti, Rajashri (2013a),"Vouchers, Public School Response and the Role of Incentives: Evidence from Florida," Economic Inquiry volume, 51(1), 500-526.

Chakrabarti, Rajashri (2013b),"Accountability with Voucher Threats, Responses, and the Test-Taking Population: Regression Discontinuity Evidence from Florida," Education Finance and Policy, 82(2), 121-167.

Chakrabarti, Rajashri (Forthcoming),"Impact of Voucher Design on Public School Performance: Evidence from Florida and Milwaukee Voucher Programs," B.E. Journal of Economic Analysis and Policy: Contributions.

Chiang, Hanley (2009),"How Accountability Pressures on Failing Schools Affects Student Achievement," Journal of Public Economics, 93, 1045-1057.

Cullen, Julie and Randall Reback (2006), "Tinkering towards Accolades: School Gaming under a Performance Accountability System," in T. Gronberg and D. Jansen, eds., Improving School Accountability: Check-Ups or Choice, Advances in Applied Microeconomics, 14, Amsterdam: Elsevier Science.

Dee, T., and B. Jacob (2011), "The impact of *No Child Left Behind* on student achievement," Journal of Policy Analysis and Management, 30(3), 418-446.

Figlio, David (2006), "Testing, Crime and Punishment", Journal of Public Economics, 90, 837-851.

Fan, Jianqing and Irene Gijbels (1996), "Local Polynomial Modeling and Its Applications", Chapman and Hall, London.

Figlio, David and Lawrence Getzler (2006), "Accountability, Ability and Disability: Gaming the System?", in T. Gronberg ed., Advances in Microeconomics, Elsevier.

Figlio, David and Cassandra Hart (2010),"Competitive Effects of Means-Tested Vouchers," National Bureau of Economic Research Working Paper Number 16056.

Figlio, David and Cecilia Rouse (2006), "Do Accountability and Voucher Threats Improve Low-Performing Schools?", Journal of Public Economics, 90 (1-2), 239-255.

Figlio, David and Joshua Winicki (2005), "Food for Thought? The Effects of School Accountability Plans on School Nutrition", Journal of Public Economics, 89, 381-394.

Hahn, Jinyong, Petra Todd, and Wilbert van der Klaauw (2001), "Identification and Estimation of Treatment Effects with a Regression Discontinuity Design," Econometrica 69 (1): 201-209.

Hoxby, Caroline (2003a), "School Choice and School Productivity (Or, Could School Choice be the tide that lifts all boats?)", in C. Hoxby (ed.) Economics of School Choice, University of Chicago Press.

Hoxby, Caroline (2003b), "School Choice and School Competition: Evidence from the United States", Swedish Economic Policy Review 10, 11-67.

Jacob, Brian (2005), "Accountability, Incentives and Behavior: The Impacts of High-Stakes Testing in the Chicago Public Schools", Journal of Public Economics, 89, 761-796.

Jacob, Brian and Steven Levitt (2003), "Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating", Quarterly Journal of Economics, 118(3).

Krieg, J. (2008), "Are students left behind? The distributional effects of *No Child Left Behind*", Edu-

cation Finance and Policy, 3(2), 250-281.

McCrary, Justin (2008),"Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test," Journal of Econometrics, 142(2): 698-714.

Neal, Derek and Diane W. Schanzenbach (2010),"Left Behind By Design: Proficiency Counts and Test-Based Accountability," The Review of Economics and Statistics, 92(2): 263-283.

Reback, Randall (2008), "Teaching to the Rating: School Accountability and Distribution of Student Achievement," Journal of Public Economics 92, June 2008, 1394-1415.

Reback, Randall, Jonah Rockoff and Heather Schwartz (2011), "Under Pressure: Job Security, Resource Allocation, and Productivity in Schools Under NCLB," NBER Working Paper Number 16745.

Rockoff, Jonah E. and Lesley J. Turner (2010), "Short Run Impacts of Accountability on School Quality," American Economic Journal: Economic Policy, 2(4): 119-147.

Rouse, Cecilia E., Jane Hannaway, David Figlio and Dan Goldhaber (2012), "Feeling the Florida Heat: How Low Performing Schools Respond to Voucher and Accountability Pressure," American Economic Journal: Economic Policy, 5(2), 251-281.

Silverman, Bernard W. (1986), "Density Estimation for Statistics and Data Analysis," New York: Chapman and Hall, 1986.

Springer, Matthew (2008), "The influence of an NCLB accountability plan on the distribution of student test score gains," Economics of Education Review 27(5), 556-563.

West, Martin and Paul Peterson (2006), "The Efficacy of Choice Threats within School Accountability Systems: Results from Legislatively Induced Experiments", The Economic Journal 116, C46-C62.

Witte, J., D. Carlson, J. Cowen, D. Fleming and P. Wolf (2004), "The Performance of Charter Schools in Wisconsin", La Follette School of Public Affairs, University of Wisconsin-Madison.

Witte, J., D. Weimer, P. Schlomer and A. Shober (2012), "MPCP Longitudinal Educational Growth Study: Fifth Year Report", SCDP Milwaukee Evaluation Report # 29.

Table 1: Effect of "Threatened Status" on Percent of Students Scoring in Various Proficiency Categories

						Using Indicators of Cutoffs Missed as Additional Instruments				
	% Min	% Basic	% Prf	% Adv	% At/Abv Prf	% Min	% Basic	% Prf	% Adv	% At/Abv Prf
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Reading	1.02	-5.59	-5.62	10.19*	4.57	0.55	-4.81	-6.58	10.84**	4.26
	(4.90)	(4.32)	(5.40)	(5.34)	(5.95)	(3.95)	(3.55)	(5.23)	(5.24)	(5.67)
Observations	1328	1328	1328	1328	1328	1328	1328	1328	1328	1328
R^2	0.25	0.28	0.15	0.43	0.34	0.24	0.29	0.15	0.42	0.34
Bandwidth	6.69	6.69	6.69	6.69	6.69	6.69	6.69	6.69	6.69	6.69
p-value[1]						0.18	0.94	0.27	0.74	0.33
Language Arts	-0.53	-7.55**	3.83	4.26	8.09	-1.49	-5.80*	5.51	1.78	7.29
	(3.67)	(3.54)	(6.06)	(7.80)	(6.63)	(3.61)	(3.04)	(6.24)	(7.37)	(6.24)
Observations	1328	1328	1328	1328	1328	1328	1328	1328	1328	1328
R^2	0.26	0.20	0.06	0.19	0.29	0.25	0.21	0.05	0.20	0.29
Bandwidth	6.69	6.69	6.69	6.69	6.69	6.69	6.69	6.69	6.69	6.69
p-value[1]						0.31	0.43	0.46	0.29	0.22
Math	8.88	-5.10	-2.68	-1.10	-3.78	2.25	-4.56	1.84	0.46	2.30
	(8.07)	(3.94)	(5.51)	(3.61)	(6.61)	(5.21)	(2.90)	(4.03)	(3.22)	(4.87)
Observations	1328	1328	1328	1328	1328	1328	1328	1328	1328	1328
R^2	0.50	0.13	0.16	0.29	0.45	0.50	0.13	0.16	0.29	0.45
Bandwidth	6.69	6.69	6.69	6.69	6.69	6.69	6.69	6.69	6.69	6.69
p-value[1]						0.08	0.52	0.05	0.35	0.08
Science	-2.74	4.02	3.55	-4.83	-1.28	-0.24	-2.27	3.11	-0.60	2.51
	(6.25)	(5.85)	(8.35)	(3.63)	(6.93)	(5.50)	(3.23)	(7.31)	(3.73)	(5.80)
Observations	1328	1328	1328	1328	1328	1328	1328	1328	1328	1328
R^2	0.28	0.48	0.14	0.39	0.49	0.29	0.48	0.14	0.39	0.49
Bandwidth	6.69	6.69	6.69	6.69	6.69	6.69	6.69	6.69	6.69	6.69
p-value[1]						0.13	0.10	0.15	0.04	0.09
Social Studies	9.92	-5.91	-9.61	5.60	-4.00	4.83	-4.66**	-4.39	4.22	-0.16
	(9.24)	(3.70)	(6.25)	(7.51)	(7.81)	(5.47)	(2.18)	(3.82)	(6.88)	(5.53)
Observations	1328	1328	1328	1328	1328	1328	1328	1328	1328	1328
R^2	0.18	0.36	0.15	0.29	0.34	0.20	0.37	0.15	0.29	0.34
Bandwidth	6.69	6.69	6.69	6.69	6.69	6.69	6.69	6.69	6.69	6.69
p-value[1]						0.10	0.69	0.19	0.26	0.09

*, **, ***: significant at the 10, 5, and 1 percent level, respectively. Robust standard errors adjusted for clustering by school district are in parentheses. All regressions include controls for racial composition, gender composition, percentage of students eligible for free/reduced price lunches, and real per pupil expenditures.

[1] Reports p-value for Wooldridge's (1995) robust score test of over-identifying restrictions. Sargan's (1958) and Basmann's (1960) tests of over-identifying restrictions give similar results; Wooldridge's test is reported here as it is robust to heteroskedasticity.

Table 2: Effect of "Threatened Status" on WRCT, Test Participation, Attendance, and Graduation

	WRCT Scores and Participation					
Panel A	% Minimal	% Basic	% Proficient	% Advanced	% Prof/Adv	% Tested
	(1)	(2)	(3)	(4)	(5)	(6)
Failed AYP	0.37	7.22**	-7.16***	-0.43	-7.59*	-1.69
	(0.70)	(3.43)	(2.53)	(2.77)	(3.91)	(1.59)
Observations	683	683	683	683	683	683
R^2	0.27	0.57	0.21	0.59	0.59	0.48
Bandwidth	7.67	7.67	7.67	7.67	7.67	7.67
Over-id. test p-value[1]	0.45	0.40	0.42	0.17	0.14	0.19

	WKCE Test Participation					
Panel B:	Reading	Lang. Arts	Math	Science	Soc. Studies	AYP Test Part.
All Students	(1)	(2)	(3)	(4)	(5)	(6)
Failed AYP	-0.79	-0.88	-0.05	-1.57	-2.18	0.00
	(1.24)	(1.25)	(0.68)	(1.53)	(1.88)	(0.64)
Observations	1329	1329	1329	1329	1329	1329
Bandwidth	6.69	6.69	6.69	6.69	6.69	6.69
Over-id. test p-value[1]	0.76	0.75	0.47	0.30	0.23	0.67

	Attendance and Graduation			
Panel C	Attendance	Attendance (No Grad)[2]	Graduation	AYP Other Indicator
	(1)	(2)	(3)	(4)
Failed AYP	-2.15	0.53	3.39	1.06
	(1.84)	(0.82)	(4.52)	(5.47)
Observations	1329	984	352	1329
R^2	0.31	0.55	0.21	0.05
Bandwidth	6.69	7.08	8.86	6.69
Over-id. test p-value[1]	0.18	0.66	0.43	0.36

*, **, ***: significant at the 10, 5, and 1 percent level, respectively. Robust standard errors adjusted for clustering by school district are in parentheses. All regressions include controls for racial composition, gender composition, percentage of students eligible for free/reduced price lunches, and real per pupil expenditures. Regressions in this table include indicators of criteria missed as instruments in addition to F.

[1] Reports p-value for Wooldridge's (1995) robust score test of over-identifying restrictions. Sargan's (1958) and Basmann's (1960) tests of over-identifying restrictions give similar results; Wooldridge's test is reported here as it is robust to heteroskedasticity. [2] Uses sample of schools where attendance matters (elementary and middle).

Table 3: Investigating the Effect of Math-Induced Failure on Percent of Students Scoring in Various Proficiency Categories and Test Participation
(Using Sample of Schools that Just Missed/Made the Math Cutoff)

Panel A: WKCE	% Minimal (1)	% Basic (2)	% Prof. (3)	% Adv. (4)	% at/above Prof. (5)	% Tested (6)
Reading	6.81	0.57	-8.58	1.20	-7.38	-2.40
	(5.63)	(3.73)	(8.20)	(6.70)	(8.68)	(3.09)
R^2	0.46	0.58	0.49	0.79	0.61	0.23
Lang. Arts	5.77	-0.26	-10.13***	4.63	-5.50	-2.45
	(3.97)	(4.21)	(3.54)	(5.34)	(6.89)	(3.08)
R^2	0.42	0.45	0.31	0.32	0.54	0.23
Math	-3.08	-3.11	0.52	5.67*	6.19**	-0.43
	(4.68)	(4.78)	(6.97)	(3.04)	(3.14)	(1.24)
R^2	0.66	0.19	0.47	0.69	0.67	0.31
Science	-7.43	9.00	-5.03	3.47	-1.56	-5.47
	(7.51)	(8.17)	(7.80)	(6.02)	(6.96)	(3.66)
R^2	0.37	0.54	0.39	0.82	0.72	0.06
Social Studies	1.15	-9.32	0.71	7.46	8.17	-6.01
	(5.50)	(6.42)	(7.09)	(9.06)	(8.06)	(4.58)
R^2	0.40	0.28	0.20	0.49	0.39	0.11
Observations	130	130	130	130	130	130
Bandwidth	10.00	10.00	10.00	10.00	10.00	10.00

Panel B: WRCT/ AYP OI	% Minimal (1)	% Basic (2)	% Prof. (3)	% Adv. (4)	% at/above Prof. (5)	% Tested (6)	AYP OI (7)
AYPfail	18.75	13.34	-14.62	-17.46	-32.08	-5.67	-3.70
	(17.38)	(15.70)	(15.21)	(14.01)	(24.52)	(10.08)	(5.41)
Observations	47	47	47	47	47	47	106
Bandwidth	10.00	10.00	10.00	10.00	10.00	10.00	10.00

*, **, ***: significant at the 10, 5, and 1 percent level, respectively. Robust standard errors adjusted for clustering by school district are in parentheses. All regressions include controls for racial composition, gender composition, percentage of students eligible for free/reduced price lunches, and real per pupil expenditures. This RD design uses the sample of schools that just missed or made the Math cutoff. It does not constrain AYP statuses in the other criteria, that is, these schools may have passed or failed the other criteria. Validity tests confirm that the probabilities of AYP failures in the other criteria are continuous through the Math Cutoff (Online Appendix Table B1 Panel G).

Table 4: Investigating the Effect of Reading-Induced Failure on Percent of Students Scoring in Various Proficiency Categories and Test Participation (Using Sample of Schools that Just Missed/Made the Reading Cutoff)

Panel A:	% Minimal	% Basic	% Prof.	% Adv.	% at/above Prof.	% Tested
WKCE	(1)	(2)	(3)	(4)	(5)	(6)
Reading	0.39	-8.62***	4.01	4.23	8.24**	2.08
	(7.83)	(2.71)	(7.98)	(6.22)	(4.13)	(1.29)
R^2	0.54	0.49	0.37	0.71	0.64	0.06
Lang. Arts	-3.85	0.75	8.13*	-5.03*	3.10	2.13*
	(5.81)	(2.19)	(4.93)	(4.04)	(6.56)	(1.28)
R^2	0.42	0.46	0.28	0.35	0.56	0.06
Math	9.99	8.54**	-4.13	-14.40*	-18.53**	2.51
	(8.06)	(4.24)	(4.23)	(7.71)	(8.41)	(1.78)
R^2	0.63	0.40	0.56	0.30	0.55	0.06
Science	23.58***	-9.61**	-12.62**	-1.35	-13.97	1.27
	(6.95)	(4.80)	(6.15)	(5.78)	(9.80)	(2.18)
R^2	0.33	0.58	0.39	0.73	0.63	0.07
Social Studies	21.90**	-2.84	0.17	-19.23**	-19.06*	1.38
	(10.29)	(2.27)	(3.59)	(8.01)	(10.15)	(2.77)
R^2	0.33	0.49	0.29	0.41	0.52	0.08
Observations	138	138	138	138	138	139
Bandwidth	10.00	10.00	10.00	10.00	10.00	10.00

Panel B:	% Minimal	% Basic	% Prof.	% Adv.	% at/above Prof.	% Tested	AYP OI
WRCT/ AYP OI	(1)	(2)	(3)	(4)	(5)	(6)	(7)
AYPfail	-0.72	15.28***	-13.41***	-1.15	-14.56***	-4.11	-7.80
	(2.46)	(1.95)	(3.78)	(2.80)	(3.38)	(3.42)	(8.93)
Observations	53	53	53	53	53	53	139
Bandwidth	10.00	10.00	10.00	10.00	10.00	10.00	10.00

*, **, ***: significant at the 10, 5, and 1 percent level, respectively. Robust standard errors adjusted for clustering by school district are in parentheses. All regressions include controls for racial composition, gender composition, percentage of students eligible for free/reduced price lunches, and real per pupil expenditures. This RD design uses the sample of schools that just missed or made the Reading cutoff. It does not constrain AYP statuses in the other criteria, that is, these schools may have passed or failed the other criteria. Validity tests confirm that the probabilities of AYP failures in the other criteria are continuous through the Reading Cutoff.

Table 5: Investigating the Response of Schools that Missed AYP by Missing Test Participation only: Effect on Percent of Students Scoring in Various Proficiency Categories and Test Participation

	% Minimal	% Basic	% Prof.	% Adv.	% at/above Prof.	% Tested
	(1)	(2)	(3)	(4)	(5)	(6)
Reading	0.49	1.40	-11.18	9.28	-1.89	1.61
	(4.30)	(4.54)	(8.56)	(6.81)	(7.79)	(1.11)
Observations	1296	1296	1296	1296	1296	1297
R^2	0.15	0.22	0.15	0.37	0.24	0.04
Bandwidth	10.00	10.00	10.00	10.00	10.00	10.00
Lang. Arts	1.42	-3.44	1.36	0.65	2.02	1.14
	(4.55)	(4.35)	(6.91)	(9.27)	(8.46)	(1.10)
Observations	1296	1296	1296	1296	1296	1297
R^2	0.19	0.17	0.05	0.16	0.23	0.05
Bandwidth	10.00	10.00	10.00	10.00	10.00	10.00
Math	-1.22	-1.83	3.69	-0.65	3.04	0.84
	(4.16)	(2.30)	(3.80)	(3.10)	(5.27)	(1.17)
Observations	1296	1296	1296	1296	1296	1297
R^2	0.42	0.11	0.12	0.26	0.36	0.04
Bandwidth	10.00	10.00	10.00	10.00	10.00	10.00
Science	2.19	-0.26	2.06	-3.99	-1.93	0.63
	(6.49)	(2.73)	(9.69)	(4.90)	(7.56)	(1.14)
Observations	1296	1296	1296	1296	1296	1297
R^2	0.20	0.42	0.09	0.35	0.38	0.04
Bandwidth	10.00	10.00	10.00	10.00	10.00	10.00
Social Studies	3.06	-1.38	-4.04	2.36	-1.68	0.45
	(5.55)	(2.63)	(6.00)	(9.52)	(7.15)	(1.17)
Observations	1296	1296	1296	1296	1296	1297
R^2	0.11	0.32	0.14	0.22	0.23	0.04
Bandwidth	10.00	10.00	10.00	10.00	10.00	10.00

*, **, ***: significant at the 10, 5, and 1 percent level, respectively. Robust standard errors adjusted for clustering by school district are in parentheses. All regressions include controls for racial composition, gender composition, percentage of students eligible for free/reduced price lunches, and real per pupil expenditures. This RD design uses the sample of schools that passed reading, math, and "other indicator". Schools to the left missed AYP by just missing the test participation cutoff. Schools to the right just made the test participation cutoff.

Table 6: Are Compositional Change or Sorting Driving Results? Investigating Demographic Shifts Using Regression Discontinuity Designs

Panel A	Using Sample of Schools Under Minimum Distance Criteria				
	% White	% Black	% Hispanic	% Asian	% Am. Indian
	(1)	(2)	(3)	(4)	(5)
	1.56	15.10	-8.46***	-3.51	-0.44
	(23.41)	(23.87)	(3.18)	(2.11)	(2.15)
Observations	1343	1343	1343	1343	1343
	% Male	% Free/Reduced Price Lunch	No. of Subgroups Counted	Real PPE	
	(6)	(7)	(8)	(9)	
	0.00	-3.72	-0.53	-1.55	
	(0.03)	(22.69)	(0.65)	(1.73)	
Observations	1343	1329	1263	1343	
	Whites Counted	Blacks Counted	Hispanics Counted	Asians Counted	
	(21)	(22)	(23)	(24)	
	0.16	0.01	-0.15**	-0.02	
	(0.27)	(0.17)	(0.06)	(0.02)	
Observations	1343	1343	1343	1343	
	Am. Indians Counted	Limited English Prof. Counted	Special Ed. Counted	Econ. Disadv. Counted	
	(25)	(26)	(27)	(28)	
	-0.01	-0.06	-0.06	-0.31	
	(0.05)	(0.04)	(0.20)	(0.20)	
Observations	1343	1343	1343	1343	
Panel B	Using Sample of Schools that just Missed/Made Math Cutoff				
	% White	% Black	% Hispanic	% Asian	% American Indian
	(1)	(2)	(3)	(4)	(5)
	-3.48	11.38	-5.46	-1.93	-0.50
	(11.87)	(12.26)	(5.49)	(2.22)	(0.65)
Observations	139	139	139	139	139
	% Male	% Free/Reduced Price Lunch	No. of Subgroups Counted	Real PPE	
	(6)	(7)	(8)	(9)	
	0.00	3.13	-0.67	-1.22	
	(0.01)	(11.24)	(0.62)	(1.68)	
Observations	139	136	129	136	
	Whites Counted	Blacks Counted	Hispanics Counted	Asians Counted	
	(21)	(22)	(23)	(24)	
	-0.27	0.08	0.02	-0.00	
	(0.16)	(0.17)	(0.13)	(0.00)	
Observations	139	139	139	139	
	Am. Indians Counted	Limited English Prof. Counted	Special Ed. Counted	Econ. Disadv. Counted	
	(25)	(26)	(27)	(28)	
	-0.04	-0.11	-0.25	-0.05	
	(0.10)	(0.08)	(0.19)	(0.19)	
Observations	139	139	139	139	

*, **, ***: significant at the 10, 5, and 1 percent level, respectively. Robust standard errors adjusted for clustering by school district are in parentheses.

Table 7: Did Competition Matter? Examining the Effect on Percent of Students Scoring in Various Proficiency Categories, Test Participation, Attendance, and Graduation (Using Sample of Schools Under Minimum Distance Criteria)

Panel A	% Minimal	% Basic	% Proficient	% Advanced	% at/above Prof	% Tested
	(1)	(2)	(3)	(4)	(5)	(6)
	Reading					
AYPfail	3.47	-4.99	-7.81	9.33*	1.52	-0.78
	(6.12)	(3.64)	(6.51)	(5.18)	(6.70)	(1.18)
AYPfail * count	-5.51**	-3.76	12.52***	-3.24	9.28**	4.44***
	(2.60)	(2.32)	(3.54)	(2.66)	(4.67)	(1.09)
	Language Arts					
AYPfail	0.77	-6.64*	2.62	3.25	5.87	-0.77
	(3.92)	(3.49)	(5.29)	(8.07)	(6.96)	(1.21)
AYPfail * count	-4.36	-6.41***	5.18	5.60**	10.77**	4.81***
	(3.37)	(2.06)	(5.01)	(2.34)	(5.25)	(1.06)
	Math					
AYPfail	7.86	-4.24	-1.61	-2.01	-3.62	-0.72
	(7.18)	(3.44)	(4.53)	(3.62)	(6.46)	(1.35)
AYPfail * count	3.61	-3.76***	-7.28**	7.43**	0.15	5.00***
	(2.99)	(1.07)	(3.06)	(2.09)	(2.89)	(1.32)
	Science					
AYPfail	-0.03	2.48	1.47	-3.92	-2.45	-2.25
	(5.65)	(4.51)	(9.08)	(3.65)	(7.54)	(2.04)
AYPfail * count	-11.88***	8.19***	9.64**	-5.96***	3.68	4.97***
	(4.53)	(1.62)	(4.72)	(2.08)	(5.08)	(1.39)
	Social Studies					
AYPfail	11.70	-5.09*	-8.45	1.84	-6.61	-2.93
	(11.07)	(2.90)	(5.33)	(9.25)	(10.09)	(2.38)
AYPfail * count	-8.31***	-3.57*	-7.32***	19.20***	11.88***	5.56***
	(3.16)	(2.04)	(2.49)	(3.89)	(4.47)	(1.66)
Observations	1294	1294	1294	1294	1294	1295
Bandwidth	6.69	6.69	6.69	6.69	6.69	6.69

Panel B	Attendance and Graduation			
	Attend.	Attend. (No G)[1]	Grad.	AYP OI
	(1)	(2)	(3)	(4)
AYPfail	-2.05	0.45	0.45	4.07
	(1.84)	(1.44)	(1.44)	(3.72)
AYPfail * count	4.61***	4.74***	4.74***	9.70***
	(1.68)	(1.42)	(1.42)	(2.31)
Observations	1295	965	965	1295
Bandwidth	6.69	7.08	7.08	6.69

*, **, ***: significant at the 10, 5, and 1 percent level, respectively. Robust standard errors adjusted for clustering by school district are in parentheses. The number of AYP-passed schools in the near vicinity of a public school is denoted "count". All regressions include "count", racial composition, gender composition, percentage of students eligible for free/reduced price lunches, and real per pupil expenditures. [1] Uses sample of schools where attendance matters (elementary and middle). AYP OI indicates AYP "other indicator" criterion.

Table 8: Do the Effects of Math-Induced Failure & Reading-Induced Failure Vary By Competition?

Panel A	Using Sample of Schools that Just Missed/Made Math Cutoff				
	% Minimal	% Basic	% Proficient	% Advanced	% at / above Prof
	(1)	(2)	(3)	(4)	(5)
	Reading				
AYPfail * count	-1.82	-0.59	3.33	-0.93	2.41
	(4.11)	(1.81)	(5.16)	(2.00)	(5.80)
	Language Arts				
AYPfail * count	1.09	-3.95	2.19	0.67	2.86
	(2.92)	(2.72)	(3.83)	(2.16)	(5.41)
	Math				
AYPfail * count	-4.79	-2.05	3.52	3.33	6.85**
	(5.27)	(4.07)	(5.69)	(3.65)	(3.49)
	Science				
AYPfail * count	-5.02	2.46	2.30	0.27	2.56
	(10.84)	(4.43)	(6.85)	(1.59)	(6.90)
	Social Studies				
AYPfail * count	-4.24	0.03	-0.14	4.36	4.21
	(8.10)	(2.84)	(2.75)	(8.27)	(10.77)
Observations	131	131	131	131	131
Panel B	Using Sample of Schools that Just Missed/Made Reading Cutoff				
	% Minimal	% Basic	% Proficient	% Advanced	% at / above Prof
	(1)	(2)	(3)	(4)	(5)
	Reading				
AYPfail * count	-12.76**	3.72**	9.32***	-0.28	9.04**
	(5.12)	(1.68)	(3.24)	(2.37)	(4.59)
	Language Arts				
AYPfail * count	-8.79***	2.66	4.59***	1.54	6.13
	(3.07)	(2.01)	(1.76)	(3.98)	(4.98)
	Math				
AYPfail * count	2.65	-1.51	-1.32	0.17	-1.15
	(4.72)	(2.24)	(3.19)	(3.85)	(6.78)
	Science				
AYPfail * count	-11.28	5.69***	3.08	2.50	5.59
	(7.37)	(1.37)	(4.28)	(2.51)	(6.41)
	Social Studies				
AYPfail * count	-17.02	7.16***	5.08	4.78	9.86
	(10.59)	(0.82)	(3.85)	(6.87)	(10.44)
Observations	155	155	155	155	155

*, **, ***: significant at the 10, 5, and 1 percent level, respectively. Robust standard errors adjusted for clustering by school district are in parentheses. All regressions include "count", AYPfail, racial composition, gender composition, percentage of students eligible for free/reduced price lunches, and real per pupil expenditures, and use a bandwidth of 10.

Table 9A: Do Schools Facing Larger Competition have Different Socio-Economic Characteristics?

Panel A	Using Sample of Schools that Just Missed/Made Math Cutoff					
	% White 03	% Black 03	% Asian 03	% AmIndian 03	% Male 03	% Free/Reduced Price Lunch 03
	(1)	(2)	(3)	(4)	(5)	(6)
AYPfail*count	-3.47	4.28	0.24	0.09	-0.00	3.95
	(3.61)	(6.00)	(0.49)	(0.20)	(0.00)	(3.64)
	% White	% Black	% Asian	% AmIndian	% Male	% Free/Reduced Price Lunch
	(7)	(8)	(9)	(10)	(11)	(12)
AYPfail*count	-3.45	4.04	0.48	0.15	-0.00	2.68
	(3.62)	(6.15)	(0.52)	(0.20)	(0.00)	(3.55)
Observations	112	112	112	112	112	110
Panel B	Using Sample of Schools that Just Missed/Made Reading Cutoff					
	% White 03	% Black 03	% Asian 03	% AmIndian 03	% Male 03	% Free/Reduced Price Lunch 03
	(1)	(2)	(3)	(4)	(5)	(6)
AYPfail*count	-2.07	1.12	-0.91	-0.18	0.01	5.05
	(5.09)	(5.33)	(0.87)	(0.29)	(0.01)	(4.25)
	% White	% Black	% Asian	% AmIndian	% Male	% Free/Reduced Price Lunch
	(7)	(8)	(9)	(10)	(11)	(12)
AYPfail*count	-1.99	1.34	-1.33	-0.15	0.01	4.96
	(5.04)	(5.49)	(0.82)	(0.32)	(0.01)	(4.12)
Observations	147	147	147	147	147	144

Table 9B: Further Probing the Role of Competition: Can the Heterogeneity in Responses be Explained by Demographics in More Competitive Areas?

Panel A	Using Sample of Schools that Just Missed/Made Math Cutoff				
	% Minimal	% Basic	% Proficient	% Advanced	% at / above Prof
	(1)	(2)	(3)	(4)	(5)
	Reading				
AYPfail * count	-2.67	1.31	-1.05	2.41	1.36
	(5.13)	(3.81)	(5.92)	(2.26)	(5.02)
	Math				
AYPfail * count	-3.85	-2.60	3.21	3.24	6.45**
	(5.78)	(6.77)	(5.15)	(3.23)	(3.29)
Panel B	Using Sample of Schools that Just Missed/Made Reading Cutoff				
	% Minimal	% Basic	% Proficient	% Advanced	% at / above Prof
	(1)	(2)	(3)	(4)	(5)
	Reading				
AYPfail * count	-15.53***	6.44***	11.53***	-2.45	9.09**
	(5.28)	(1.82)	(3.34)	(3.48)	(4.60)
	Language Arts				
AYPfail * count	-9.75***	3.40	3.94	2.42	6.35
	(3.42)	(2.39)	(2.45)	(3.96)	(5.37)
	Math				
AYPfail * count	3.08	-2.30	-0.97	0.19	-0.78
	(4.23)	(2.53)	(2.96)	(3.86)	(6.44)

Notes for Table 9A & 9B: *, **, ***: significant at the 10, 5, and 1 percent level, respectively. Robust standard errors adjusted for clustering by school district are in parentheses. Table 9A regressions include "count" and AYPfail, and use a bandwidth of 10. Table 9B regressions include racial composition, gender composition, percentage of students eligible for free/reduced price lunches, and real per pupil expenditures, interactions of each of these variables with "count", AYPfail, "count", and use a bandwidth of 10.

Table 10: Did Competition Affect Resource Allocation?
(Using Sample that Just Missed/Made Math Cutoff & Sample that Just Missed/Made Reading Cutoff)

	Dependent Var : Real Per Pupil Expenditure	
	Sample: Just Passed/Failed in Math	Sample: Just Passed/Failed in Reading
	(1)	(2)
AYPfail * count	-0.82	-0.83
	(9.59)	(2.55)
Observations	128	137
Bandwidth	10.00	10.00

Table 11: Examining the Effect of Reading-Induced Failure & Math-Induced Failure
(Using Schools that Just Missed/Made the Math Cutoff and Reading Cutoff in 2002-03 and 2003-04)

Panel A	Impact of Math-Induced Failure					
	% Minimal	% Basic	% Prof.	% Adv.	% at/above Prof.	% Tested
	(1)	(2)	(3)	(4)	(5)	(6)
Reading	9.54	-1.90	-5.33	-2.32	-7.64	3.14
	(6.51)	(4.38)	(5.84)	(4.98)	(7.27)	(3.04)
R^2	0.42	0.53	0.46	0.81	0.62	0.50
Math	0.16	-2.48	-4.01	6.32**	2.31	1.35
	(4.30)	(3.51)	(3.38)	(3.19)	(2.75)	(1.58)
R^2	0.72	0.26	0.53	0.72	0.75	0.06
Observations	287	287	287	287	287	287
Bandwidth	10.00	10.00	10.00	10.00	10.00	10.00
Panel B	Impact of Reading-Induced Failure					
	% Minimal	% Basic	% Prof.	% Adv.	% at/above Prof.	% Tested
	(1)	(2)	(3)	(4)	(5)	(6)
Reading	3.31	-10.37**	1.48	5.58	7.06**	-1.64
	(3.41)	(4.29)	(4.10)	(3.96)	(3.55)	(1.75)
R^2	0.41	0.36	0.37	0.76	0.58	0.01
Math	7.03**	1.49	-6.97*	-1.55	-8.52	-2.17
	(3.38)	(2.97)	(4.02)	(3.42)	(5.30)	(1.45)
R^2	0.68	0.15	0.51	0.55	0.67	0.57
Observations	251	251	251	251	251	252
Bandwidth	10.00	10.00	10.00	10.00	10.00	10.00

*, **, ***: significant at the 10, 5, and 1 percent level, respectively. Robust standard errors adjusted for clustering by school district are in parentheses. Table 10 regressions include "count" and AYPfail. Table 11 regressions include racial composition, gender composition, percentage of students eligible for free/reduced price lunches, and real per pupil expenditures.

Table A1: Testing Validity of Regression Discontinuity Analysis: Looking for Discontinuities in Pre-Program Characteristics at the Cutoff

Panel A	% at or above Proficient				
	Reading	Language Arts	Math	Science	Social Studies
	(1)	(2)	(3)	(4)	(5)
	-3.66	1.06	-3.13	-3.92	0.15
	(10.73)	(12.96)	(15.58)	(14.89)	(9.49)
Panel B	% Tested				
	Reading	Language Arts	Math	Science	Social Studies
	(6)	(7)	(8)	(9)	(10)
	0.11	0.21	-0.82	-0.57	-0.72
	(1.52)	(1.51)	(1.42)	(1.45)	(1.38)
Panel C	% White	% Black	% Hispanic	% Asian	% American Indian
	(11)	(12)	(13)	(14)	(15)
	1.24	14.74	-7.77***	-3.91	-0.35
	(23.65)	(23.79)	(2.97)	(2.36)	(2.09)
Panel D	% Male	% Free/Reduced Price Lunch	No. of Subgroups Counted	Real PPE	Attendance Rate
	(16)	(17)	(18)	(19)	(20)
	0.01	-7.71	-0.55	-2.22	0.65
	(0.02)	(21.74)	(0.50)	(2.20)	(1.48)
Panel E	Whites Counted	Blacks Counted	Hispanics Counted	Asians Counted	
	(21)	(22)	(23)	(24)	
	0.07	0.04	-0.11**	0.03	
	(0.27)	(0.23)	(0.05)	(0.05)	
Panel F	Am. Indians Counted	Limited English Prof. Counted	Special Ed. Counted	Econ. Disadv. Counted	
	(25)	(26)	(27)	(28)	
	-0.06	-0.05	-0.21	-0.27	
	(0.05)	(0.03)	(0.14)	(0.18)	

*, **, ***: significant at the 10, 5, and 1 percent level, respectively. Robust standard errors adjusted for clustering by school district are in parentheses.

Table A2: Effect of Math-Induced and Reading-Induced Failure After Dropping 2002-03 SIFI Schools

Panel A	Using Sample of Schools that Just Missed/Made the Math Cutoff					
	% Minimal	% Basic	% Prof.	% Adv.	% at/above Prof.	% Tested
	(1)	(2)	(3)	(4)	(5)	(6)
Reading	4.05	-3.93	-2.65	2.52	-0.13	-2.18
	(3.99)	(4.33)	(7.37)	(8.15)	(6.85)	(3.81)
Math	-5.60	-0.73	-1.83	8.16*	6.33**	1.38
	(4.94)	(2.85)	(4.59)	(4.82)	(3.04)	(1.43)
Observations	97	97	97	97	97	97
Bandwidth	10	10	10	10	10	10

Panel B	Using Sample of Schools that Just Missed/Made the Reading Cutoff					
	% Minimal	% Basic	% Prof.	% Adv.	% at/above Prof.	% Tested
	(1)	(2)	(3)	(4)	(5)	(6)
Reading	3.56	-11.49**	-2.01	5.93	7.94	-1.30
	(4.11)	(5.17)	(10.95)	(5.37)	(5.33)	(1.60)
Math	10.49**	-1.38	-14.05***	4.93	-9.11	-1.55
	(5.27)	(3.12)	(4.20)	(5.23)	(7.28)	(1.54)
Observations	115	115	115	115	115	116
Bandwidth	10	10	10	10	10	10

*, **, ***: significant at the 10, 5, and 1 percent level, respectively. Robust standard errors adjusted for clustering by school district are in parentheses. All regressions include controls for racial composition, gender composition, percentage of students eligible for free/reduced price lunches, and real per pupil expenditures.

Table A3: Summary Statistics
(Using Schools that Missed/Made Math Cutoff (Panel A) and Reading Cutoff (Panel B))

Panel A	Sample of Schools that Just Missed/Made Math Cutoff				
	% at or above Proficient				
	Reading	Language Arts	Math	Science	Social Studies
	(1)	(2)	(3)	(4)	(5)
	61.71	52.49	47.16	47.36	64.57
	(15.76)	(17.52)	(21.34)	(22.20)	(18.72)
	% White	% Black	% Hispanic	% Asian	% American Indian
	(6)	(7)	(8)	(9)	(10)
	40.67	42.24	10.13	4.90	2.06
	(35.39)	(38.81)	(16.89)	(5.53)	(9.16)
	% Male	% Free/Reduced Lunch	Real PPE	No. of Subgroups Counted	Attendance Rate
	(11)	(12)	(13)	(14)	(15)
	0.52	57.63	59.90	3.45	91.30
	(0.03)	(31.58)	(4.85)	(1.60)	(5.48)
	% of schools below cutoff				
	Reading	Math	Attendance	Graduation Rate	Test Participation
	(16)	(17)	(18)	(19)	(20)
	42.86	41.35	13.53	32.00	13.53
	(49.67)	(49.43)	(34.34)	(47.12)	(34.34)
Panel B	Sample of Schools that Just Missed/Made Reading Cutoff				
	% at or above Proficient				
	Reading	Language Arts	Math	Science	Social Studies
	(1)	(2)	(3)	(4)	(5)
	61.86	53.93	52.51	52.12	67.64
	(14.65)	(15.47)	(19.89)	(21.29)	(16.07)
	% White	% Black	% Hispanic	% Asian	% American Indian
	(6)	(7)	(8)	(9)	(10)
	47.90	33.05	11.84	6.08	1.12
	(35.31)	(36.95)	(18.86)	(8.04)	(2.34)
	% Male	% Free/Reduced Lunch	Real PPE	No. of Subgroups Counted	Attendance Rate
	(11)	(12)	(13)	(14)	(15)
	0.52	54.29	58.76	3.13	92.41
	(0.03)	(31.58)	(4.34)	(1.59)	(4.50)
	% of schools below cutoff				
	Reading	Math	Attendance	Graduation Rate	Test Participation
	(16)	(17)	(18)	(19)	(20)
	48.95	27.27	9.09	26.00	13.29
	(50.16)	(44.69)	(28.85)	(44.31)	(34.06)

Figure 1A: Relationship Between Treatment Status and the Running Variable (Distance From the AYP Cutoff)

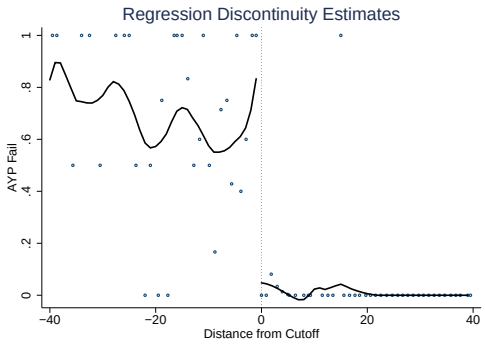


Figure 1B: Is there a Discontinuity in the Density of the Running Variable at the Cutoff?

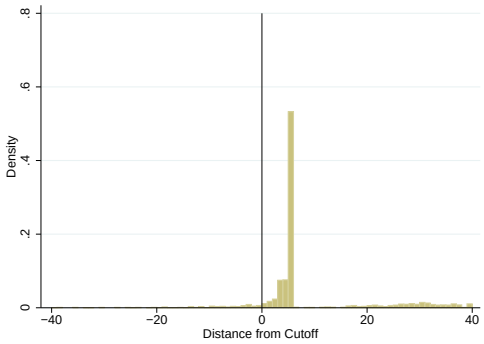
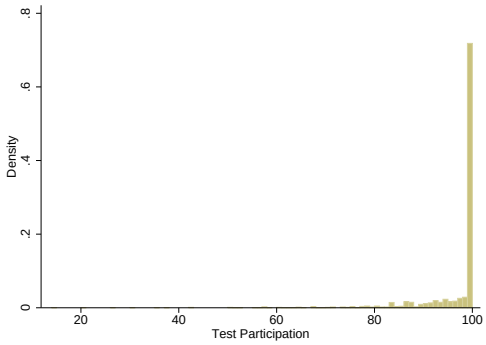


Figure 1C: Distribution of Test Participation in 2001-02



The running variable in Figures 1A-1B is the distance of the lowest performing subgroup/subject criterion from the corresponding cutoff.

Figure 2A: Relationship Between Treatment Status and distance from Math Cutoff

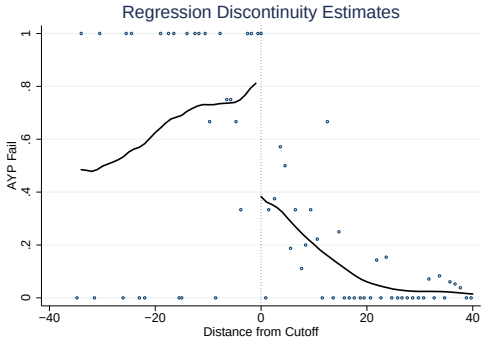
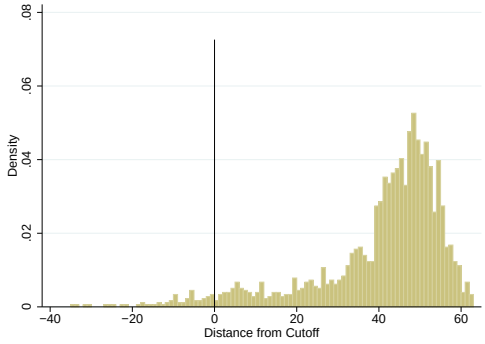


Figure 2B: Assessing Continuity in the Density of the Running Variable



The running variable is distance from the math cutoff.

Figure 3A: Relationship Between Treatment Status and distance from Reading Cutoff

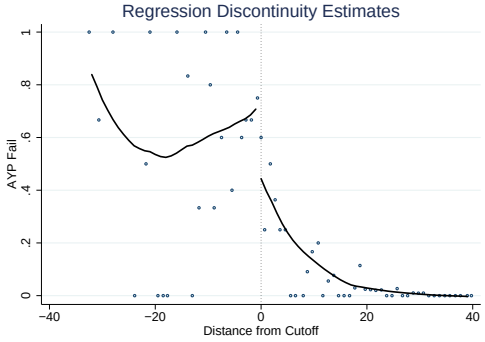
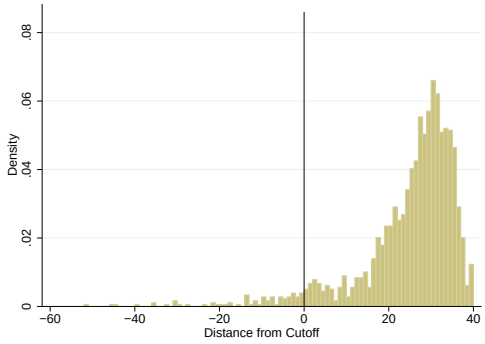


Figure 3B: Assessing Continuity in the Density of the Running Variable



The running variable is distance from the reading cutoff.

Figure 4A: Relationship Between Treatment Status and distance from Test Participation Cutoff

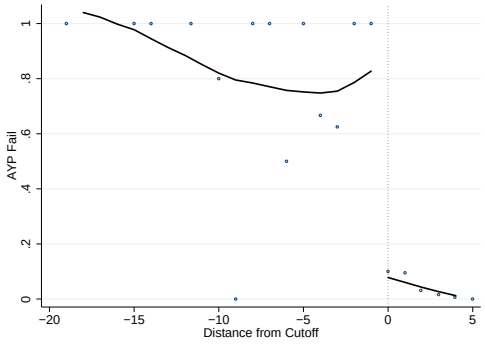
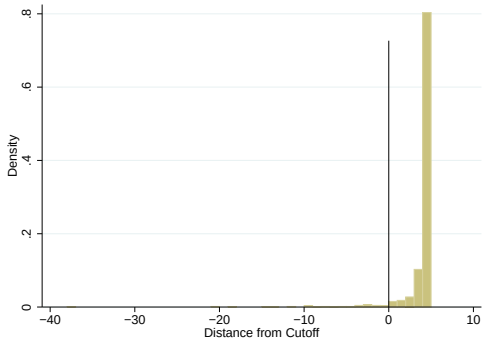


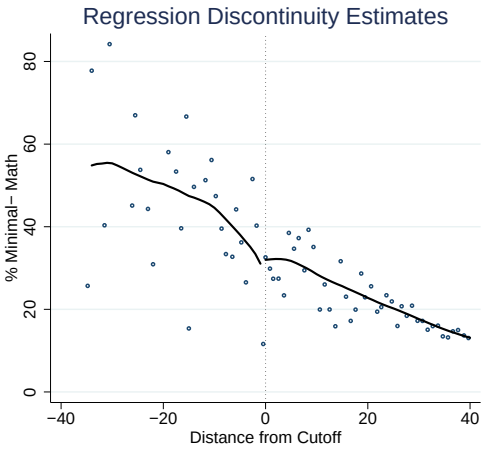
Figure 4B: Assessing Continuity in the Density of the Running Variable

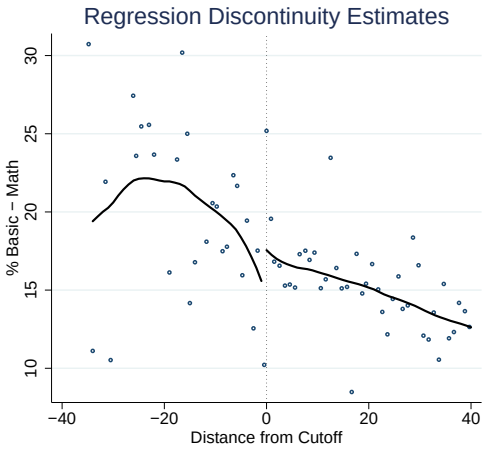


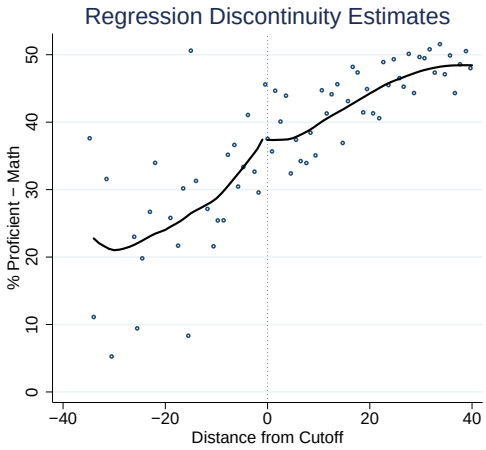
The running variable is distance from the test participation cutoff in the sample of schools that made reading, math, and "other indicator".

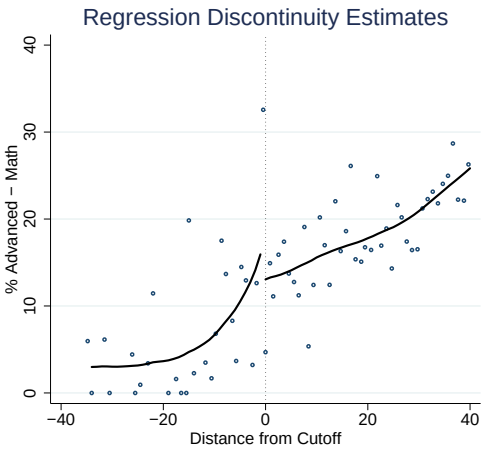
Figure 5: Investigating the Response of Schools that Missed AYP by Missing Math: Effect on Percent of Students Scoring in Various Proficiency Categories in Math and Reading

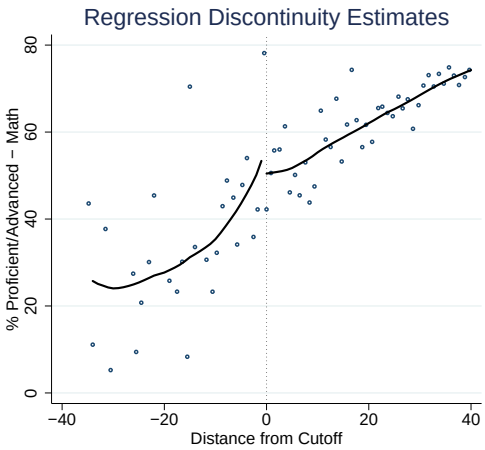
Panel A: Impacts on Math



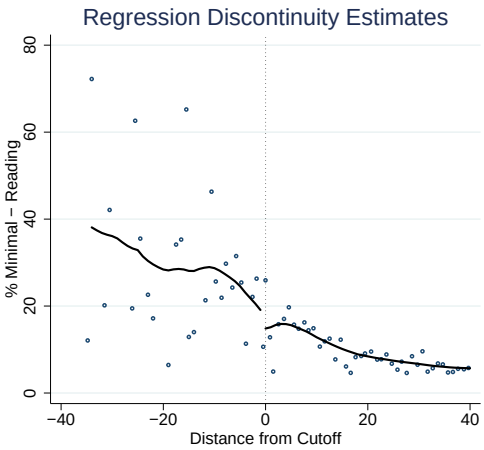


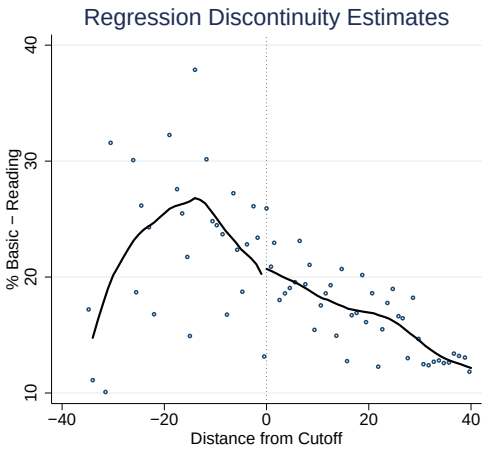


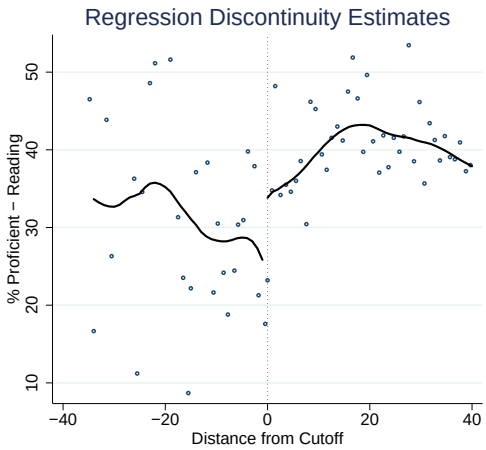


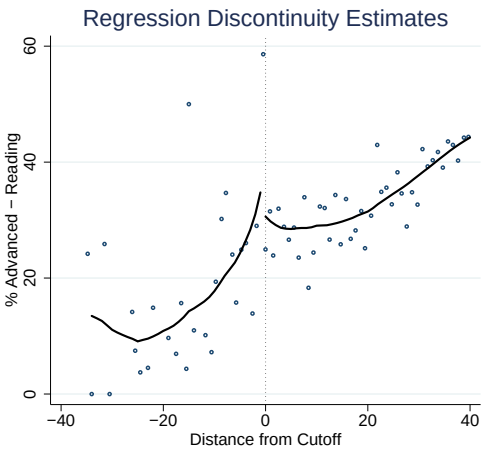


Panel B: Impacts on Reading









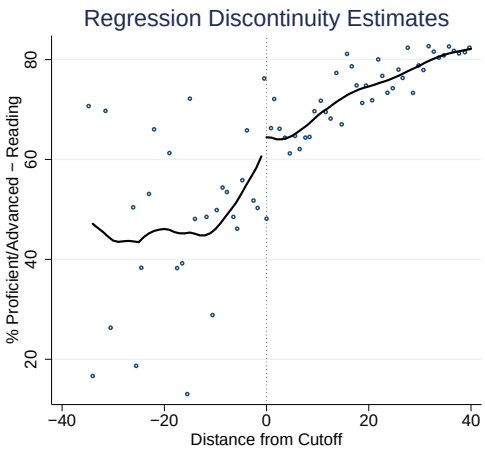
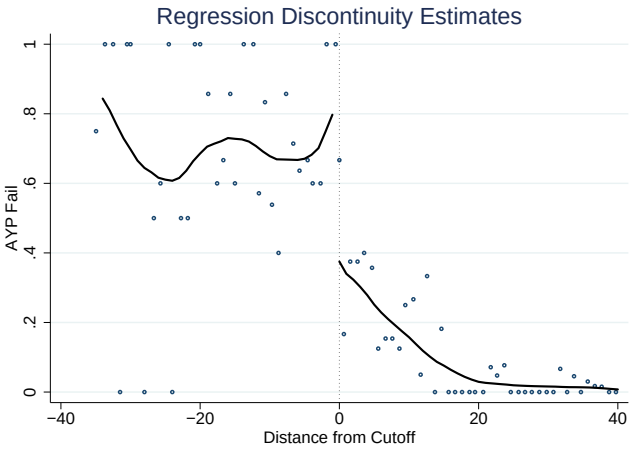
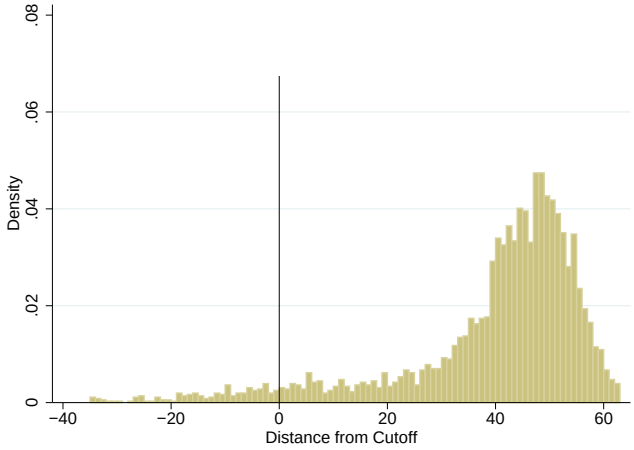


Figure 6A: Relationship Between Treatment Status and Math Running Variable (Using Multiple Years)



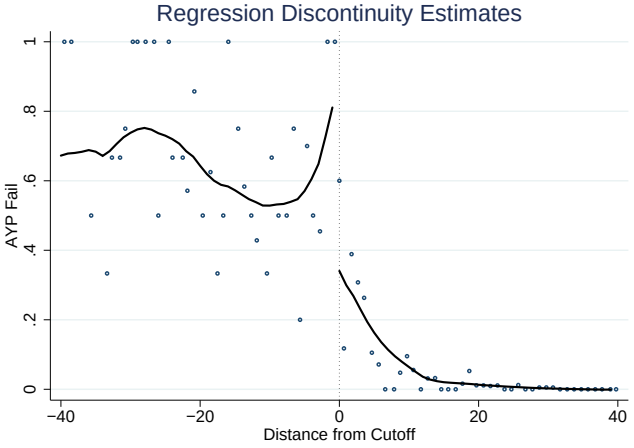
The running variable is distance from the math cutoff.

Figure 6B: Is there a Discontinuity in the Density of the Math Running Variable at the Cutoff? (Using Multiple Years)



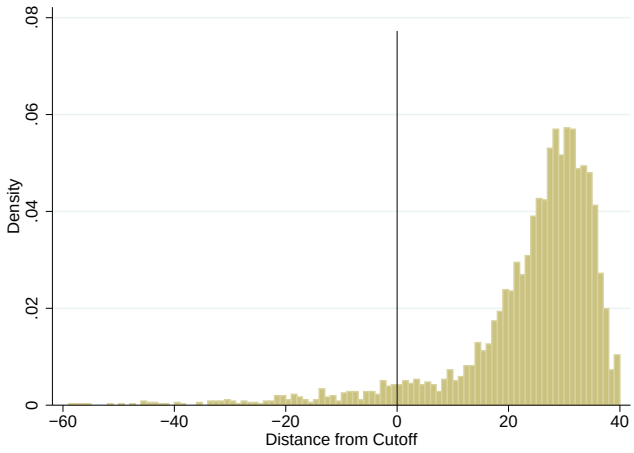
The running variable is distance from the math cutoff.

Figure 7A: Relationship Between Treatment Status and Reading Running Variable (Using Multiple Years)



The running variable is distance from the reading cutoff.

Figure 7B: Is there a Discontinuity in the Density of the Reading Running Variable at the Cutoff? (Using Multiple Years)



The running variable is distance from the reading cutoff.

Figure A1: Plotting Means of Pre-Program Characteristics