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

Rajashri Chakrabarti[†]
Federal Reserve Bank of New York

Preliminary, please do not cite without permission.

Abstract

NCLB mandated the institution of 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. In this paper, I study the incentives and responses of schools that failed AYP once. Using data from Wisconsin and regression discontinuity designs, I find evidence of improvements in high stakes reading and spillover effects to low stakes language arts in these schools. The patterns are consistent with focus on marginal students around the high stakes cutoff, but this did not come at the expense of the ends. There is not much evidence in favor of improvement in high stakes math, or low stakes science and social studies. Performance in low stakes grades suffered, and so did performance in weaker subgroups inspite of their inclusion in AYP computations. While there is no robust evidence in favor of effects in test participation and graduation, attendance improved in threatened schools where they mattered for AYP. There is strong evidence in favor of response to incentives—schools that failed AYP by failing only in reading and/or math did substantially better in these subject areas. Credibility of threat mattered—AYP-failed schools that faced more competition responded both more strongly and more broadly, with robust evidence in favor of improvements in all AYP objectives.

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

[*]I thank Damon Clark, Randy Reback, Jonah Rockoff, Wilbert van der Klaauw, Matt Wiswall, Basit Zafar and seminar participants at Columbia University, FRBNY/NYU Education Seminar Series, 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 that failed AYP for two or more consecutive years faced ESEA sanctions that escalated with the number of years of failed AYP. Title 1 schools missing AYP 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 study the incentives and responses of schools that failed AYP once. 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 (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 (example, advantaged versus disadvantaged) were affected? I also look at the effect of threatened status 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 regression discontinuity designs to investigate these questions. Using Wisconsin data, I find evidence in favor of an improvement in reading in the threatened schools. These schools indeed tended to focus more on the high stakes students close to and around the cutoff. But this did not come at the expense of the ends—rather, there was a rightward shift of the whole distribution. I find that language arts patterns paralleled those in reading closely, possibly due to spillover effects. I do not find evidence in favor of improvement in either high stakes math or low stakes science and social studies.[1] In Wisconsin, around this time period, WKCE math was generally regarded as harder to improve in than reading. Consistent with this notion, these schools did worse in math relative to reading before the program also—this pattern seems to have continued after the program. The general notion of difficulty may have discouraged focus in math (or students may not have improved inspite of some focus). It is also worth noting here that responses of the threatened schools depended on performance in the AYP objectives in the previous year. Schools that missed AYP by missing only the reading and/or math objectives did substantially better in reading; their performance in math was also relatively better, but still the effects in math were small and not statistically different from zero.

The patterns in subgroup performance reveal that the weaker groups lost irrespective of the law's emphasis on these groups. Both special education and economically disadvantaged groups (and especially the latter) showed evidence in favor of deterioration in each subject area. 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 in threatened schools where they mattered for AYP (graduation, instead of attendance counted in AYP for the high schools).

Thus, while there is evidence that the threatened schools responded according to incentives in some areas—example, in high stakes reading and attendance rates—there is not much evidence of response

[1] This result is qualitatively similar to Reback et al. (2011). Studying responses in low stakes subjects, they find evidence of statistically significant improvements in reading in schools on the AYP margin. Effects in math for these schools were considerably smaller and not statistically different from zero.

in other high stakes indicators.[2] Was this because these schools may not have faced a credible threat? Did schools that faced more competition respond more strongly and broadly? 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.[3] This indeed seems to have been the case. Schools that faced more competition not only responded more strongly in reading and attendance rates, but also responded more broadly. They showed improvements in the other high stakes criteria as well—math, test participation, attendance and graduation. These results confirm that schools did indeed respond according to incentives—feasibility and credibility of threat mattered. Schools that faced meaningful threat of loss of students[4] responded strongly in each of the AYP indicators.

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 (Greene (2001), Greene and Winters (2003), Figlio and Rouse (2006), West and Peterson (2006), Chakrabarti (2008a), Rouse et al. (2007), Chiang (2009), Rockoff and Turner (2010)). But, there is also evidence in favor of reclassification of low performing students into disabled categories (Cullen and Reback (2006), Figlio and Getzler (2006) and Jacob (2005)), teacher cheating (Jacob and Levitt (2003)), strategic suspensions of students (Figlio (2006)), increased focus on high stakes marginal students (Reback (2008), Ladd and Lauen (2010), Chakrabarti (forthcoming)) 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 (2008b)) 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

[2] Also of note here is that most of the reading gains came from improvements in elementary grade rather than middle and high school grades which possibly indicates that the schools desisted from making substantial investments in response to AYP-failure.

[3] Under NCLB rules, students who became eligible for public school choice could only move to higher performing (AYP passed) schools. Also if just the AYP-fail status encourages students to move, it is likely that they will move to better performing (that is, AYP-passed) schools. So the extent of effective threat faced by an AYP-failed school should be judged by the density of AYP-passed schools (rather than any school) in its near vicinity.

[4] Note that both stigma and the threat of ESEA public school choice threatened schools with potential loss of students in the near future. Just AYP-failed status (both the stigma associated with it as well as the prospect of studying in an AYP-failed school) might induce students to move away to better performing (AYP-passed) schools. In this paper, I do not distinguish between stigma and the threat of public school choice. In a parallel paper (Chakrabarti 2011), I study and contrast the effects of NCLB threat of sanctions versus stigma on public school responses in both cognitive and non-cognitive outcomes.

(2010) study the effects of NCLB on 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 margin 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 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 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 that also entered AYP formation. 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.

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 Adequate Yearly Progress (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.

To understand the effect of NCLB on public school incentives in Wisconsin, we have to first understand the accountability and testing systems in Wisconsin during this period. 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[5], the tests have been given in grades 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 performance, 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 years under consideration in this paper (2003 and 2004). 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%) during the period under consideration. 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.

If the cell size of 40 was not met for the All Student group, then proficiency data from the previous year was required to be combined with the current year. Thus combined, if the All student group reached

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

40 or above, then it was held accountable for the reading and math objectives. If the cell size was still not met, then the school was not included in AYP calculations.

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.[6]

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 rate of 81.75% in high schools.

3 Data

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 and 2003-04.

The high stakes examination in Wisconsin is known as the Wisconsin Knowledge and Concepts Examination (WKCE). It is administered annually statewide in the subject areas of reading, language arts, math, science and social studies. 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 and 2004.

Apart from the WKCE, Wisconsin also administers a Reading test in grade 3. It is known as the

[6] 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 Safe Harbor step 2 criterion for each subgroup was the state's starting point for that subgroup or growth from the prior year's proficiency rate in Science.

Wisconsin Reading Comprehension test (WRCT). It is a low stakes test in the sense that scores from this test do not contribute to AYP formation. Nor was it given in a high stakes grade. (The high stakes grades during this period were 4, 8 and 10). 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 in 2004.

Data on percentage of students tested in each 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 The Questions Posed

In the analysis that follows, I investigate the effect of the threat of NCLB sanctions and stigma on schools that failed AYP once. Did the threatened schools focus more on reading and math, the high stakes subject areas, relative to the schools that did not face an immediate threat? Did the threat of sanctions and stigma lead to a shift of focus from low stakes to high stakes subject areas? Did they lead to a shift away from students in the low stakes grades to high stakes grades? Under NCLB guidelines, certain percentages of students in each group (or subgroup) had to score above some specified thresholds on the score scale to make AYP. Did this motivate the threatened schools to focus more on students expected to score close to and just around these high stakes cutoffs rather than equally on all students?

Unlike most state accountability regimes, NCLB holds schools accountable not only for test scores, but also for attendance, graduation and test participation, and they form an integral part of AYP. The objective was to have a more broad-based effect on schools rather than a narrow effect on test scores. Did this provision lead the threatened schools to improve their performance in these indicators?

Still another feature of NCLB is that it holds not only the whole school accountable, but also individual subgroups. The objective here was to guard against the possibility of some subgroups (especially the

weaker ones) from falling through the cracks. Did this rule motivate the threatened schools to improve the performance of the weaker subgroups, such as economically disadvantaged and special education students?

In addition, I also look at the effect of the threat of sanctions and stigma on performance of white students. This provides an interesting contrast to the above exercise and helps understand how one of the more advantaged subgroups was affected. It is also instructive from the point of view that the whites constituted, on average, the most numerous group. While the less numerous subgroups could potentially move in and out of the designated minimum cell size, the more numerous subgroups would likely always count towards AYP. This might induce threatened schools to focus more on numerous subgroups.

I also investigate whether there was heterogeneity of responses based on schools' incentives. AYP-failed schools are likely to be a heterogenous group in terms of their past year's individual criteria pass/fail histories. While they might be unified in terms of their aim to avoid failure in the following year, they might emphasize different indicators differently depending on their previous year's performance—I examine whether this has been the case. While the first part of the paper looks at the effect of general AYP-failure, in the latter part I take a closer look at these incentives—Did schools that failed AYP by only failing in test participation focus more on test participation and less on other indicators? Did schools that failed AYP by failing only in reading/math focus more on these subject areas and less on others?

I also analyze the role of credible threats and competition. Did schools' responses depend on whether the threats were credible? To study this, I investigate whether AYP-failed schools that had more high performing (AYP-passed) schools in their near vicinity showed larger responses. Under NCLB rules, students who became eligible for public school choice could only move to higher performing (AYP-passed) schools. Similarly, if stigma associated with attending AYP-failed schools induced students to move, they would likely move to better (AYP-passed) schools. So the extent of threat a AYP-failed school perceived (and hence the responses) likely depended on the density of AYP-passers in its near vicinity—I investigate whether this was indeed the case.

4.2 Methodology

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. I use a regression discontinuity strategy to study the causal effect of failing AYP. The analysis entails comparing the responses of schools that barely missed AYP with schools that barely made AYP.

The AYP formula is complex, contributed by pass rates in different subgroup–subject categories, subgroup–test-participation categories, attendance and graduation rates. However, the institutional details of the AYP formula enable me to reduce it to a single, continuous, one-dimensional measure 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 criterion. 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 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. 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.

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, that is $p_{jkst} \geq (10 + 0.9 p_{jks,t-1}).I_1$, where I_1 is an indicator variable denoting whether or not safe harbor 2 is satisfied.[7,8]

Now, consider the test-participation criterion. Each subgroup j passes the test-participation objective

[7] As described in section 4.2, safe harbor requires the percentage of students not yet proficient to be decreased by at least 10% from the prior year for that group/sub-group. Denoting the percentage of students scoring below proficient in subgroup j subject k school s and year t by q_{jkst}, this implies $\frac{q_{jkst} - q_{jks,t-1}}{q_{jks,t-1}} \leq -0.1 \Rightarrow \frac{q_{jks,t-1} - q_{jkst}}{q_{jks,t-1}} \geq 0.1 \Rightarrow 0.9 q_{jks,t-1} - q_{jkst} \geq 0 \Rightarrow 0.9(100 - p_{jks,t-1}) - (100 - p_{jks,t}) \geq 0 \Rightarrow p_{jkst} - 0.9 p_{jks,t-1} \geq 10 \Rightarrow p_{jkst} \geq 10 + 0.9 p_{jks,t-1}$.

[8] For the All Student group, $I_1 = 1[OI_{st} - \bar{OI} \geq 0]$, where OI denotes the other Indicator criteria (see discussion below). For all other subgroups, $I_1 = 1[max[p^s_{jkst} - c^s_j, p^s_{jkst} - p^s_{jks,t-1}] \geq 0]$ where p^s_{jkst} denotes percentage of students scoring at or above the Science cutoff in supgroup j, school s, year t, and c^s_j denotes the Science cutoff.

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 my 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 $TP_{jst} = Max_k\{TP_{jkst}\}$. I consider 2002-03 as the pre-program year and use data from this year to calculate the running variable. Recall that NCLB was signed into law in January 2002. Tests in Wisconsin were held in October-November 2002, so schools did not have much time to respond before the tests. But, perhaps more 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) or enough time to manipulate their position on the AYP scale before October 2002. Therefore, 2002-03 is treated as a pre-program year.

Figure 1 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. The figure suggests that the minimum distance predicts AYP status well and can potentially serve as the running variable in a regression discontinuity strategy, provided other validity assumptions (tested below) are satisfied.

An advantage of a regression discontinuity analysis is that identification relies on a discontinuous jump in the probability of treatment at the cutoff. Consequently, a potential confounding factor such as mean reversion that is important in a difference-in-differences setting is not likely to be important here, as it likely varies continuously with the running variable at the cutoff. Also, regression discontinuity analysis essentially entails comparison of schools that are very similar to each other (a testable assumption that I test later) except that the schools to the left faced a discrete increase in the probability of treatment. As a result, another potential confounding factor in a difference-in-differences setting, existence of differential pre-program trends, is not likely to be important here.

I first examine whether the use of a regression discontinuity strategy is valid here. Identification of β_1 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, I test if this was indeed the case. The discontinuity estimates are presented in Table 1 and graphs corresponding to a subset of these (to save space) are presented in Figure 2. Panel A reports discontinuity estimates for pre-program (2001-02) percentage of students at or above proficient in the five subject areas; panel B presents estimates for pre-program percentages of students tested in these subject areas; Panel C and Panel D present results for pre-existing socio-economic characteristics, attendance rate as well as number of subgroups that counted towards AYP. Panels E and F present discontinuity 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. So, from the above discussion, it seems reasonable to say that this case passes the test of smoothness of predetermined characteristics through the cutoff.

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 3 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.[9] So the heapings at 5 and 3 are artefacts of the AYP rule and not caused by manipulation at these points/spikes.

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 4 shows, the

[9] 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 who were accountable only for the reading and math objectives, so their minimum distance was not constrained to be at 5 or below.

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. (2010) point out, regression discontinuity 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.

Following Barreca et al. (2010), I first try to investigate whether these data heaps at 3 and 5 are problematic. Figure 5 represents 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 9 the points to the right of 5 pertain to the small schools that probably were different.[10] To take account of this, I use two strategies. First, as suggested by Barreca et al. (2010), I drop the sample to the right of 5 and repeat the analysis. Second, I construct an alternate running variable that is the same as the running variable in (1) except that it excludes the test-participation criterion. The results in each of these cases remain qualitatively similar. (They are not reported to save space, but are available on request). So, these heapings at 3 and 5 do not seem to pose a threat to the identification of LATE.

Having established that the use of regression discontinuity strategy in this setting is valid, I next proceed to look at the effect of AYP failure on the behavior of threatened schools. As Figure 1 shows, I have a fuzzy discontinuity. 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)). 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, $F_{st} = 1(r_{st} < 0)$, $y_{jks,t+1}$ denotes outcome of subgroup (or group) j of school s in criterion k in year $t+1$. If $g(r_{st})$ and

[10] 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 Student" group counted towards AYP. But for 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), only the "All Student" subgroup counted. So the plot jumps down to one just to the right of 5 since we only have small schools in that range.

$h(r_{st})$ 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.[11] The results remain qualitatively similar and are available on request.

5 Results

5.1 Effect on Performance in High Stakes and Low Stakes Subject Areas in High Stakes Grades

Using the regression discontinuity strategy described above, Table 2 columns (1)-(5) look at the effect of "threatened status" on percentage of students scoring in minimal, basic, advanced, proficient and "at or above proficient" respectively in the five subject areas. The first panel reports second stage results for reading, the second panel reports language arts results, the third reports math, the fourth science, and the fifth social studies. All regressions reported in this paper control for 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 (Table 1), the purpose of inclusion of covariates here is variance reduction. Indeed, as expected, the results do not depend on inclusion/exclusion of controls. The corresponding first stage (Appendix table 1, column 1) is strong as reflected in statistical and economic significance of F_{st} and the F-statistic for excluded instruments.

The reading results show a rightward shift of the distribution with 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.[12] 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. Interestingly, language arts patterns mirror those in reading. There

[11] 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.

[12] 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.

is a rightward shift in the language arts distribution with a statistically significant decline in percentage of students just below proficient and a corresponding increase in percentage of students scoring at or above proficient. There is no evidence in favor of improvements in either Math, Science or Social Studies.

A question that naturally arises in this context is whether the improvements of the marginal students in reading and language arts came at the expense of the non-marginal (lower and higher performing) ones. There is no evidence of such a pattern except for a small increase in percent of students in the lowest performing category (minimal) in reading, but this effect is also statistically insignificant.

It should be noted here that the changes in Table 2 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. It is also possible that some students moved from the advanced to the proficient category. But this does not seem to have been a major factor because the net increase in the advanced category is positive. Similarly, to the extent that there may have been moves from the upper or lower proficiency categories to basic, the actual decline in basic is even larger than that seen in the estimate. Again, these patterns do not seem to have been prominent because cumulative percentage change in categories above basic is large and positive, and percent change in minimal is also positive, though small. While there seems to have been a small net flow into the lowest category (minimal), the basic thrust is that there have been net declines around the cutoff (basic and proficient) and this was associated with net flows into the highest category, advanced. It follows that there has been a net rightward shift of the distribution and the patterns are consistent with the hypothesis of increased emphasis on the students expected to score around the cutoff (the marginal students). In reading, except for a small increase in minimal, there is no evidence that this increased focus came at the expense of the ends.

One potential concern here is that the schools close to the cutoff may be failing or passing in different objectives. However, the finding in Table 1 that the pre-existing characteristics are balanced on either side of the cutoff allays this concern, at least to some extent. But to explicitly control for the differences in pass/fail statuses in different objectives, I 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.

$I_{k,s} = 1[min_j\{P_{jkst} - min\{c_k, (10 + 0.9P_{jks,t-1}) * I_1\}\} < 0]$ *where* $k = \{Reading, Math\}$

$I_{TP,s} = 1[min\{TP_{jst} - \bar{TP}\} < 0]$

$I_{OI,s} = 1[min\{OI_{st} - \bar{OI}\} < 0]$

The results from re-estimating the above regressions (including these indicator variables in addition) are qualitatively similar. These are not reported to save space, but are available on request. In partial response to this issue of different AYP-failed schools failing in different criteria, I later use alternative regression discontinuity designs to study the responses of schools that failed AYP by failing only in test participation or only in reading and/or math criteria. The analysis (section 6) yields some interesting patterns.

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.[13] An advantage of this strategy is that it permits me to conduct a test of the over-identifying restrictions. The test statistic is never statistically different from zero, thus confirming the validity of the instruments.[14] Using these instruments, Table 2 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 are qualitatively similar to above, where F_{st} was used to instrument for AYP failure. Once again, there is evidence of fall in percentage of students around the cutoff (basic and proficient categories) in reading, and a statistically significant increase in percentage of students scoring in advanced category. As earlier, Language Arts patterns are similar to Reading, suggesting spillover effects. There is not much evidence of effects in Math, Science and Social Studies.

To summarize, there is some evidence in favor of improvement in reading (one of the high stakes subject areas) and the patterns are consistent with the hypothesis that the threatened schools tended to focus on students expected to score around and close to the proficiency cutoff. Interestingly, the increased focus on the high stakes students around the proficiency cutoff did not come at the expense of the ends. Although Language Arts was low stakes, there is no evidence of deterioration in this subject area. In fact, there seems to have been spillover effects from Reading to Language Arts. The skills required in language arts are likely similar to those required in reading, which might have led to such spillover effects. There is no evidence of improvement in the low stakes subject areas of Science and Social Studies.[15]

[13] The first stage results (Table A1 column 2) show that the other instruments also have predictive power, especially test-participation, but F_s always has the most predictive power. Probability of treatment is always discontinuously higher (and also statistically so for F_s and $I_{TP,s}$) if a school lies to the left of the cutoff of the corresponding running variable.

[14] In this paper, I use Wooldridge's (1995) robust score test of over-identifying restrictions because it is robust to heteroscedasticity. Sargan's (1958) and Basmann's (1960) tests of over-identifying restrictions give similar results.

[15] Note that Science was not completely low stakes for schools that missed the AYP cutoff in Reading or Math both in the regular and "Confidence Interval" way. For those schools "safe harbor" could make them pass these criteria. But for "safe harbor" to be applicable, percentage of students in the failing subgroup had to score above a certain cutoff in Science. Thus while Science did matter for some schools and subgroups, it did not matter for all schools and probably was also not among the most salient components of AYP criteria.

Although Math was also a high stakes subject area, there is no evidence of improvement in Math. It might be worth thinking a little bit here why that might have been the case. In Wisconsin (WKCE), math was widely regarded as harder to improve in, which might to some extent explain the lack of improvement in Math. In fact, because Math was regarded as harder to improve in, Wisconsin started with a considerably lower target proficiency cutoff (37%) in Math in contrast to Reading (61%). Also of note here is that Wisconsin schools performed considerably worse in math compared to reading in the pre-program years. For example 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. Thus historically these schools did worse in math (relative to reading) and this trend seems to have continued later as well, despite NCLB. It is worth mentioning here that these reading and math patterns are consistent with those obtained in Reback et al. (2011). While they find a statistically and economically significant effect in reading for schools at the AYP margin, the effect in math is considerably smaller and statistically not different from zero.

5.1.1 Effect on Subgroup Performance in High Stakes and Low Stakes Subject Areas

Tables 3 and 4 look at the effect of AYP failure on the distribution of scores in three subgroups—whites, economically disadvantaged and special education. The results reported in this and the next subsection are obtained from regressions that include F_s as well as indicators of criteria missed as instruments. The results from models that use F_s as the only instrument are very similar and not reported for lack of space, but are available on request. Interestingly, the patterns for whites (Table 3) suggest rightward shifts in the distributions of both reading and math, although the effects are often not statistically significant. In both subject areas, there is evidence of net movements of students from lower performance categories to the advanced category. Language Arts patterns once again mirror reading patterns. Patterns in science and social studies also show some evidence in favor of improvements. The findings in this table suggest improvements in the high stakes subject areas, but also spillover effects to other low stakes subject areas. While Reading and Language Arts on the one hand, and Math and Science on the other, arguably use similar skills and have synergies that might make spillover effects more natural, spillover effects to social studies is not immediately obvious. However, if AYP failure brings in a new sense of urgency and drive

within this subgroup and/or the threatened school tends to, in general, focus more on this subgroup (either because it is a majority group or because improving scores in this group is less costly), then this might lead to overall improvements that might get reflected in high stakes as well as low stakes subject areas including social studies.

Table 4 looks at the performance effects in two relatively weaker groups, Economically Disadvantaged and Special Education. There is evidence in favor of deterioration in each of the subject areas for both subgroups. In stark contrast to the patterns for whites, in each of the subject areas, there seems to have been a leftward shift of the score distributions with net movements of students from the higher performance to lower performance categories, and many of these effects are highly significant (especially for the economically disadvantaged group). The negative effects are most pronounced in Social Studies, as might be expected.

5.2 Test-Participation of "All Students" and Other Subgroups

Table 5 looks at the effect of AYP failure on test participation of students. Panel A looks at the effect on test participation in the "All Student" category, Panel B looks at the test-participation effect for White students, Panel C for Economically Disadvantaged students, and Panel D for Special Education students. The first five columns in each panel look at test participation in the five subject areas, while the last column looks at the effect on the AYP test participation criterion. In neither the "All Student" group nor the "Whites" subgroup, is there any evidence of improvements in test participation. This is true in all the five subject areas as well as for the AYP test participation criterion. The effects are always negative, but they are always small and never statistically significant.

The effects on Economically Disadvantaged test participation (Panel C) are also negative for each of the subject areas, but they are almost always larger in magnitude than the above effects and also statistically different from zero for social studies. The picture for Special Education is somewhat different. Here, there is clear evidence of a decline in test participation in each of the subject areas as well as the AYP test participation criterion—the effects are always economically larger than above and always statistically different from zero. Also, interestingly, the negative effect is economically the largest for social studies. In spite of the heterogeneities, a common thread permeates test participation patterns in each of these groups/subgroups. While the effects are negative across the board (often small and not statistically significant), these negative effects are almost always economically stronger in the low stakes (rather than the high stakes) subject areas, with this negativity being the most prominent in social

studies. To summarize, while there is no evidence of any improvement in test participation in any group or subject area, the weaker subgroups and the low stakes subject areas seem to have suffered/lost the most.

5.3 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, that is, students in that grade did not face the high stakes tests. It would be interesting to see whether the threatened schools tended to focus less on low stakes grades and students.

Table 6 looks at the effect of AYP failure on third grade WRCT scores and test-participation. The top panel uses F_s as the instrument, while the lower panel uses F_s as well as indicators for cutoffs missed as additional instruments. While the effects in the upper panel are not statistically significant, the patterns in both panels suggest a leftward shift in the WRCT distribution. There seems to have been a net move from the higher proficiency categories (Proficient and Advanced) towards lower proficiency categories (Minimal and Basic). The findings suggest that the threatened schools tended to shift their focus away from the low stakes grade 3 reading. While the threatened schools may have focused more on WKCE Reading in the high stakes grades, this seems to have come at the expense of performance in the low stakes grade 3. Columns (6) and (12) look at the effect on test participation in WRCT—there is not much evidence in favor of any effect on test participation in third grade WRCT.

5.4 Attendance and Graduation Rates

Table 7 looks at the effect of AYP failure on attendance and graduation rates. The upper panel uses F_s as the instrumental variable, while the lower panel uses indicators of criteria failed as additional instruments. Columns (1) and (5) find that there was not much effect on attendance rate. While these columns include all schools, columns (2) and (6) look at the effect on attendance constraining the sample only to schools where attendance counted in AYP. (Recall that attendance rate contributed to AYP formation only if schools did not have a twelfth grade.) Interestingly, the effect on attendance is considerably more positive in these columns, and also statistically significant in the top panel. These findings suggest that the threatened schools for whom attendance mattered tended to focus more on attendance relative to schools for whom attendance was not high stakes. There is also some evidence in

favor of positive effects on graduation, although the effects are not always significant.

Columns (4) and (8) look at the effect on "Other Indicator" as defined in the AYP formula. Once again, the results suggest that AYP failure led to improvement in attendance and graduation of schools where they mattered, as reflected in positive effects on the "Other Indicator" criterion, which is also statistically significant in the top panel. The results in this table suggest that the threatened schools tended to focus on parts of the "Other Indicator" criterion that mattered for their AYP formation.

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.

As discussed above, table 1 columns (21)-(28) and (18) find no evidence of any discontinuity in the distribution of accountable subgroups at the cutoff in the pre-program year. First, there is no evidence of discontinuity in the number of subgroups that counted at the cutoff. In addition, I construct indicator variables corresponding to each subgroup that represent whether the corresponding subgroup met the designated cell size. Except Hispanics, none of these "whether subgroup counted" variables show evidence of any discontinuity at the cutoff. I also re-run the above regressions, now explicitly controlling for each of the "whether subgroup counted" variables. The results are presented in appendix table A2 and are similar to above.[16] 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.

[16] Also, note that as table 8 shows later, there is no evidence of differences in the distribution of accountable subgroups at the cutoff after the program.

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. I use the same regression discontinuity strategy as above except now the above outcome variables are replaced by the various variables reported in Table 8. As columns (1)-(7) show, there is no evidence of any effect on any of the demographic variables except in percentage Hispanic. Note that this pattern is very similar to the patterns in the pre-program demographics (table 1) 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, I also look 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 the accountability of Hispanics (whether Hispanics counted).[17] Again, this pattern is similar to the pre-program scenario. 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.

6 A Closer Look at Incentives: Exploiting Alternate Regression Discontinuity Designs

The above sensitivity tests suggest that the effects obtained in tables 2-7 indeed capture the effect of AYP failure at the cutoff (LATE). But, it might be worthwhile here to look more closely at the incentives faced by these AYP-failed schools. By construction of the running variable, schools to the left of the AYP-fail cutoff may have missed AYP in different criteria. For example, some schools may have missed AYP in reading, others may have missed in math or test participation or "other indicator".[18] These schools would likely face different incentives and might respond in different ways. For example, one

[17] Also note that, as mentioned earlier, controlling for these variables do not affect results (table A2) which further indicates that sorting was not a major factor.

[18] In fact, 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. (These numbers add to more than 100 because many schools missed AYP in multiple criteria.) 18% of these schools missed only the reading cutoff, 15% only math and 41% only in test participation, no school missed only other indicator.

might argue that reading AYP-failures would respond in reading, math AYP-failures would respond in math, and so on and so forth. Therefore, since schools to the left of the cutoff above failed in a variety of criteria, it is not apriori clear that this heterogenous group would show improvements in each of the high stakes criteria. In this section, I look at incentives and responses more closely using two strategies, as discussed below.

6.1 Focusing on Reading and Math Objectives

First, consider the sample of schools that passed the test participation and "other indicator" criteria. In this sample of schools, I compare schools that just barely missed the reading and/or math cutoffs with schools that just made the cutoffs. According to the AYP rules, schools to the right of the cutoff should make AYP while schools to the left should not. Note that schools to the left of the cutoff here will not have strong incentives in test participation and other indicator,—in contrast, they will likely have strong incentives in reading and math.[19]

Here the running variable is represented by the minimum of the distances of the reading and math criteria from the relevant cutoffs in the sample of schools that passed test participation and other indicator criteria:
$r_{1,st} = min_{jk}[P_{jkst} - min\{c_k, (10+0.9P_{jks,t-1}) * I_1\}]$, $\forall s$ that satisfies $min_j[max_k\{TP_{jkst}\} - \bar{TP}] > 0$ and $OI_{st} - \bar{OI} > 0$. Figure 6 illustrates the relationship between AYP status and schools' minimum distances from the cutoff in this sample of schools. Indeed, there is a discrete change in the probability of treatment at the cutoff. Schools that fell to the left of the cutoff had a discontinuously higher probability of failing than schools that fell to the right of the cutoff. To check the validity of this regression discontinuity strategy, I also investigate the continuity of the pre-program characteristics as well as the density of the running variable at the cutoff. I use the same pre-program characteristics as reported in table 1—I find no evidence that the continuity assumption was violated. I also do not find any evidence of discontinuity in the density of the running variable at the cutoff (Figure 7). It is worth noting here that while an important advantage of this strategy is that it affords a closer look at the incentive effects of AYP-failed schools, it leads to a steep fall in the sample size due to the nature of the design.[20]

[19] Note that while a regression discontinuity strategy that compares schools that missed AYP by just missing only the reading (math) cutoff with schools that just made the cutoff would enable a more direct investigation of incentives in reading (math), very small sample sizes to the left of the cutoff precludes me from using these strategies. However, in the above sample, out of schools that fell to the left of the cutoff within the bandwidth,—42% missed AYP by missing only the reading cutoff, 35% by missing only the math cutoff and 23% by missing both cutoffs—in other words, 58% missed the math and 65% the reading cutoffs. So, on average, schools that missed AYP here had incentives in both reading and math.

[20] Recall that this strategy only keeps schools that pass the minimum criteria in test participation and other indicator

Exploiting this regression discontinuity strategy, table 9 looks at the effect of AYP failure on performance in various subject areas, high stakes and low stakes. Columns (1)-(5) report results where the instrument is given by $F_{1,st} = 1(r_{1,st} < 0)$, columns (6)-(10) represent results where I use indicators of missing reading cutoff and math cutoff as additional instruments.

Consistent with earlier patterns, the estimates show evidence in favor of a rightward shift in the distribution of threatened schools in reading, with more focus on the marginal students close to the cutoff. But, notably (and interestingly), the effects in reading are stronger than earlier—there is a considerably larger percentage of students scoring at or above proficient (the high stakes cutoff). This implies that schools indeed responded according to their incentives,—schools that missed AYP by missing only the reading and/or math criteria indeed focused more on reading.

Once again, the language arts patterns mirror the patterns in reading. But, as in reading, the language arts effects here are stronger than that obtained earlier. Interestingly, the patterns in math are more positive than above (section 5.1) with evidence consistent with a rightward shift of the distribution, but the effects are often not statistically significant. Once again, there is no clear evidence of improvements in either science or social studies.

Table 10 panels A and B look at the effect of AYP failure on test participation, attendance and graduation—Panel A uses $F_{1,st}$ as the only instrument; panel B uses indicators of whether reading criteria was missed and whether math criteria was missed as additional instruments. There is no evidence of any effect on test participation, but there is some evidence of positive effects on attendance in schools where attendance mattered (elementary/middle schools).

Panels C and D investigate the effect of threatened status on performance in the low stakes test WRCT given in the low stakes grade 3. Panel C uses $F_{1,st}$ as the sole instrument, while Panel D uses all three instruments. The estimates in these panels 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. Interestingly, this negative effect is despite the fact that the threatened schools showed improvements in reading in the high stakes grade. The results suggest that the increased focus on high stakes reading in high stakes grades might have led to a shift of emphasis away from the low stakes grades.

(which leads to the fall), and compares those that barely missed versus those that barely passed the reading/math cutoff. Because of this fall in sample size, I consider a constant bandwidth of 10 rather than the Silverman rule of thumb bandwidth. (Note though that the results obtained are not a function of bandwidth and results remain qualitatively similar with alternate bandwidths.) But, as this subsection shows, the results for this strategy are broadly similar qualitatively, with results in reading here stronger than earlier, as might be expected.

6.2 Focusing on Test Participation Incentive

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. Here, the running variable is given by: $r_{2,st} = min_j[max_k\{TP_{jkst}\} - \bar{TP}]$, $\forall s$ that satisfies $min_{jk}[P_{jkst} - min\{c_k, (10 + 0.9P_{jks,t-1}) * I_1\}] > 0$ and $OI_{st} - \bar{OI} > 0$.

Figure 8 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 considerably higher probability of failing 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 9) are indeed continuous through the cutoff (former not reported for lack of space, but available on request).

Table 11 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 found earlier for the threatened schools become considerably muted—in fact, there is a small negative (insignificant) effect on percentage of students scoring at or above proficient. There is not much evidence of effects in any of the other subject areas and the effects are never statistically significant. Table 11 column (6) investigates the effect on test participation. While there is a small positive effect on test participation in reading and language arts in the threatened schools, they are not statistically significant. There is no evidence of effects on test participation in any of the other subject areas. There is also no evidence of any effect on "other indicator" criteria and hence the results are skipped to save space.

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

The above analysis suggests that the effects of threatened status were not broad based, being mostly limited to reading, spillover effects to Language Arts, and attendance rates. Is this because the NCLB sanctions did not have teeth or were not regarded as credible? In general, compliance with NCLB sanctions and stigma was typically low in Wisconsin. For example, in the 2003-04 school year in Wisconsin, while 37,651 students became eligible to transfer to another higher performing (AYP-passed) public

school due to their school missing AYP two years in a row, only 758 students (2%) did actually transfer.[21] The threatened schools might have anticipated this low compliance and felt the lack of credibility of threats. If, indeed this was the case, then incentives to improve or respond to the program would be low, generating limited effects as seen above.

7.1 Probing the Extent of Responses: Were there heterogeneities of Responses by Grades?

To cast more light into the extent of responses of schools, I investigate whether there were heterogeneities of responses across grades. The intuition is as follows. There are likely heterogeneities of school efforts required to make improvements in the elementary, middle and high school levels. While relatively lower resource investments can presumably produce improvements in the elementary level, improvements in the higher levels likely require more substantial changes and more investment in terms of resources, personnel and time. Based on this argument, I investigate whether there were heterogeneities in responses of AYP-failed schools across the three different levels.

Table 12 presents the results of this analysis. Interestingly, there is no evidence of any effect in either reading or math in any of the middle or high school grades 8 and 10. In the elementary grade, once again there is evidence of improvement in reading, unlike in math. So, it seems that the improvement of AYP-failed schools in reading seen above was driven by improvement in reading in grade 4. Reading in elementary grades is probably less costly to improve in and possibly does not require a lot of additional investments, either in terms of money or personnel. The above analysis supports the hypothesis that the response of AYP-failed schools was generally weak and was restricted to more easily malleable and less costly grades and subject areas.

Is the weak response a result of limited perception of threat? Did schools doubt the credibility of sanctions/stigma and the feasibility of their implementation? Did schools that faced a more credible threat (example, more competition) respond more strongly?

7.2 Did Competition Matter?

In this section, I look more closely at the effect of competition on the responses of AYP-failed schools. The idea is to investigate whether schools that had more AYP-passed schools in their near vicinity exhibited larger responses.

[21] Take-up/participation was similarly low in latter years as well. In 2004-05, 16430 students became eligible for transfer, but only 197 (1.2%) did; in 2005-06, 17010 became eligible, but only 197 (1.2%) participated.

First, using street addresses of schools, I geocode public schools and find the number of AYP-passed schools within a certain radius of each AYP-failed school.[22] 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.

The results from this analysis are presented in tables 13 and 14. Table 13 looks at the effect on the distribution of scores. The first panel 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. Note that there is some evidence that in reading, improvement in these schools came at the expense of the highest performing students. But this effect is small and not statistically significant.

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. It should be noted though that there is some evidence that the relative improvement in math (in the AYP-failed schools that faced more competition) came at the expense of the lowest performing students, but the effect is not statistically significant.

Table 14 looks at the competition effects in test participation, 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. While in general there was not a strong response from the AYP-failed schools and the response was mostly limited to reading (specifically, elementary school reading), schools that faced more competition and threat of future loss of students 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 threat-

[22] The results presented here 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.

ened schools in general was most likely due to lack of adequate competition and inadequate perception of threat. 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.

8 Conclusions

In this paper, I study the effect of AYP failure on high stakes and low stakes outcomes of threatened schools. Exploiting the institutional details of the program, I use regression discontinuity methods to analyze these effects. I find evidence that the threatened schools tended to improve in reading, and consistent with incentives, they tended to focus more on students expected to score around the high stakes proficiency cutoff. But, interestingly, this improvement of the marginal students did not seem to have come at the expense of the ends. Patterns in Language Arts reveal patterns similar to Reading, suggesting spillover effects from Reading. Language Arts requires skills that are similar to those developed in Reading, so such spillover effects are perhaps not surprising. In contrast, there is no evidence of improvement in the low stakes subject areas, science and social studies, or in the high stakes subject area, math. It is worth noting that there is evidence in favor of heterogeneity of responses based on prior year's performance in AYP objectives, and this behavior conforms well with incentives. Schools that missed AYP by only missing AYP in reading and/or math, responded more strongly in reading. While there is still no statistically significant evidence of their improvement in math, at least economically their response in math is more positive. As discussed above, there was a general sentiment in Wisconsin that math was considerably harder to improve in, that may have to some extent discouraged efforts in math (or in spite of focus, students failed to improve). This general sentiment is also consistent with the fact that these schools had historically performed worse in math relative to reading.

Performance effect patterns for Whites show a slightly different picture. While once again there are improvements in Reading and parallel patterns in Language Arts (suggesting spillover effects), there is now evidence in favor of improvement in Math, the other high stakes subject area. There are similar patterns in Science and Social Studies as well, suggesting spillover effects. Reasoning and other skills developed in Math can arguably be useful in Science making spillover effects possible. While it is not clear how much Reading or Math skills can translate to Social Studies, general build-up of skills or a

sense of urgency and motivation generated by AYP failure may have helped.

A typical shortcoming of many accountability policies is that weaker student groups are often neglected,—more so because they either often do not count towards ratings or grades, or it is possible to bypass them and still make the grade. NCLB was designed with an objective to overcome these shortcomings by including the scores of these groups in AYP formation. However, unfortunately, inspite of this measure, there is no evidence of improvement of these groups in any of the subject areas. In fact, there is evidence in favor of deterioration in performance in all subject areas, especially in the Economically Disadvantaged subgroup. Similar patterns are reflected in test participation in the various subject areas. There is evidence of deterioration in test participation in these groups, especially in the Special Education group.

The patterns in attendance are interesting. Attendance entered AYP formation only for schools that did not have a twelfth grade, while schools that had a twelfth grade were accountable for graduation rate. While there is no effect on attendance in the group of threatened schools where it did not count, the effects are much more positive in threatened schools where it did count.

WKCE was the high stakes test in Wisconsin given only in the high stakes grades. Increased focus on high stakes grades or high stakes students might yield positive externalities to other related subject areas for the same students, as seen above to some extent (language arts). It can be due to skills earned due to increased "teaching to the test" that are readily transferable to other subject areas or due to better or more productive school/class environment caused by the urgency generated by AYP failure. But a valid question is did this increased focus on the high stakes students come at the expense of low stakes students or grades? To shed some light into this question, I look at the effect of threatened status on third grade WRCT. Interestingly, while Reading was the subject area that showed the most consistent positive effect in high stakes grade 4, the picture is quite different in Reading in low stakes grade 3. While not all the effects are statistically significant, the patterns suggest that threatened status led to a deterioration of reading performance in grade 3. The entire distribution seems to have shifted to the left with moves from Advanced and Proficient categories to Basic and Minimal. This suggests that the increased focus on high stakes students may have come at the expense of low stakes ones.

Of note is that while there is evidence that schools responded according to incentives in WKCE reading (and attendance), there is not much evidence of improvement in other high stakes margins (example, math). Moreover, the response in reading seems to have largely come from improvements in elementary reading (rather than middle and high school reading), which arguably is less costly to

improve. A question that naturally arises here is why was there a lack of strong and broad responses. Wasn't there a credible perception of threat on the part of the threatened schools? Take-up of public school choice provision was pretty low (2 %). Consequently, it is possible that the threatened schools doubted the feasibility of the consequences which in turn got reflected in the responses. To cast more light into this issue, I investigate whether schools that faced more competition and hence a more credible threat responded more strongly. Indeed threatened schools that had more high performing (AYP-passed) schools in their vicinity responded more strongly. And this stronger response was seen not only in reading, but also in other indicators—math, test participation, attendance and graduation. Therefore, it seems that competition and the credibility of consequences mattered—schools that perceived credible threats responded considerably strongly than those that did not.

9 References

Ballou Dale, and Matthew Springer (2008),"Achievement Trade-Offs and No Child Left Behind," Working Paper, Urban Institute.

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

Chakrabarti, Rajashri (2008a),"Impact of Voucher Design on Public School Performance: Evidence from Florida and Milwaukee Voucher Programs," Federal Reserve Bank of New York Staff Paper Number 315.

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

Chakrabarti, Rajashri (Forthcoming),"Vouchers, Public School Response and the Role of Incentives: Evidence from Florida," Economic Inquiry.

Chiang, Hanley (2009),"How Accountability Pressures on Failing Schools Affects Student Achievement," Journal of Public Economics volume 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 Maurice Lucas (2004), "What's in a Grade? School Report Cards and the Housing Market", American Economic Review, 94(3), 591-604.

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.

Greene, Jay and Marcus Winters (2003), "When Schools Compete: The Effects of Vouchers on Florida Public School Achievement," Education Working Paper 2.

Greene, Jay (2001), "An Evaluation of the Florida A-Plus Accountability and School Choice Program," New York: Manhattan Institute for Policy Research.

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.

Holmstrom, B., and P. Milgrom (1991), "Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design," Journal of Law, Economics, and Organization, 7, 24-52.

Hoxby, Caroline (2003a), "School Choice and School Productivity (Or, Could School Choice be the tide that lifts all boats?)", in Caroline Hoxby (ed.) The 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.

Imbens, Guido W., and Thomas Lemieux (2008), "Regression Discontinuity Designs: A guide to practice", Journal of Econometrics, 142 (2), 615-635.

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

Ladd, Helen and Douglas Lauen (2010), "Status Versus Growth: The distributional Effects of School Accountability Policies", Journal of Policy Analysis and Management 29(3), 426-450.

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 (2007), "Feeling the Florida Heat: How Low Performing Schools Respond to Voucher and Accountability Pressure," CALDER (National Center for Analysis of Longitudinal Data in Education Research) Working Paper 13.

Van der Klaauw, Wilbert (2002), "Estimating the Effect of Financial Aid Offers on College Enrollment: A Regression-Discontinuity Approach," International Economic Review,Vol 43(4), November 2002.

Silverman, Bernard W. (1998), "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 (510), C46-C62.

Table 1: 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 2: 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 3: Effect of "Threatened Status" on Percent of White Students Scoring in Various Proficiency Categories

	% Min	% Basic	% Prf	% Adv	% At/Abv Prf
	(1)	(2)	(3)	(4)	(5)
Reading	-1.86	-0.58	-6.77	8.68*	1.28
	(5.09)	(5.51)	(13.58)	(5.20)	(10.08)
Observations	595	595	595	595	595
R^2	0.13	0.20	0.17	0.31	0.22
Bandwidth	5.71	5.71	5.71	5.71	5.71
Over-id. test p-value[1]	0.13	0.10	0.38	0.64	0.11
Language Arts	-3.57	-5.07	-2.01	10.66	8.64
	(3.62)	(6.56)	(4.32)	(11.25)	(9.73)
Observations	595	595	595	595	595
R^2	0.20	0.19	0.12	0.24	0.25
Bandwidth	5.71	5.71	5.71	5.71	5.71
Over-id. test p-value[1]	0.24	0.41	0.45	0.49	0.26
Math	-1.54	-9.45***	-7.40	18.39	10.99
	(6.38)	(3.61)	(5.63)	(11.32)	(8.62)
Observations	595	595	595	595	595
R^2	0.26	0.16	0.06	0.29	0.30
Bandwidth	5.71	5.71	5.71	5.71	5.71
Over-id. test p-value[1]	0.33	0.86	0.57	0.50	0.55
Science	-1.91	-0.60	-1.70	4.20	2.51
	(7.24)	(4.78)	(5.97)	(8.57)	(10.53)
Observations	595	595	595	595	595
R^2	0.13	0.22	0.13	0.29	0.21
Bandwidth	5.71	5.71	5.71	5.71	5.71
Over-id. test p-value[1]	0.12	0.64	0.50	0.43	0.20
Social Studies	-1.66	-6.29**	-6.35	14.31	7.96
	(10.27)	(3.18)	(5.66)	(11.99)	(9.39)
Observations	595	595	595	595	595
R^2	0.11	0.15	0.21	0.21	0.13
Bandwidth	5.71	5.71	5.71	5.71	5.71
Over-id. test p-value[1]	0.43	0.72	0.39	0.50	0.52

*, **, ***: 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. The regressions in this table use indicators of criteria failed 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.

Table 4: Effect of "Threatened Status" on Percent of Economically Disadvantaged and Special Education Students Scoring in Various Proficiency Categories

	Economically Disadvantaged Students					Special Education Students				
	% Min	% Basic	% Prf	% Adv	% At/Abv Prf	% Min	% Basic	% Prf	% Adv	% At/Abv I
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Reading	11.54***	2.96	-9.94	-4.56	-14.50***	5.50	0.28	-4.92	-0.86	-5.78
	(4.46)	(2.89)	(7.49)	(6.30)	(4.91)	(6.26)	(3.16)	(5.04)	(3.69)	(6.72)
Observations	125	125	125	125	125	65	65	65	65	65
R^2	0.33	0.21	0.28	0.32	0.37	0.27	0.11	0.24	0.42	0.23
Bandwidth	7.48	7.48	7.48	7.48	7.48	8.41	8.41	8.41	8.41	8.41
Over-id. test p-value[1]	0.38	0.34	0.63	0.63	0.21	0.23	0.18	0.58	0.34	0.62
Language Arts	14.30**	6.90***	-15.79***	-5.40***	-21.20***	5.49	0.49	-4.68	-1.30	-5.97
	(5.97)	(2.20)	(5.00)	(1.89)	(5.29)	(7.58)	(5.52)	(4.81)	(1.78)	(6.10)
Observations	125	125	125	125	125	65	65	65	65	65
R^2	0.26	0.12	0.18	0.27	0.24	0.51	0.36	0.39	0.14	0.35
Bandwidth	7.48	7.48	7.48	7.48	7.48	8.41	8.41	8.41	8.41	8.41
Over-id. test p-value[1]	0.71	0.97	0.86	0.92	0.95	0.56	0.25	0.35	0.12	0.20
Math	9.69	-1.59	-9.70***	1.60	-8.10	8.03**	-4.09*	-3.25	-0.68	-3.94
	(6.32)	(2.76)	(3.69)	(3.75)	(6.66)	(3.62)	(2.42)	(2.80)	(1.99)	(3.95)
Observations	125	125	125	125	125	65	65	65	65	65
R^2	0.41	0.12	0.43	0.10	0.34	0.58	0.35	0.45	0.26	0.45
Bandwidth	7.48	7.48	7.48	7.48	7.48	8.41	8.41	8.41	8.41	8.41
Over-id. test p-value[1]	0.31	0.96	0.33	0.34	0.17	0.24	0.18	0.09	0.45	0.13
Science	18.90***	-5.12	-8.23***	-5.55	-13.78**	11.19**	-4.47	-5.30	-1.42	-6.72*
	(4.84)	(3.93)	(3.04)	(4.13)	(6.48)	(5.11)	(3.81)	(3.31)	(2.73)	(3.79)
Observations	125	125	125	125	125	65	65	65	65	65
R^2	0.43	0.27	0.41	0.27	0.42	0.45	0.35	0.45	0.37	0.51
Bandwidth	7.48	7.48	7.48	7.48	7.48	8.41	8.41	8.41	8.41	8.41
Over-id. test p-value[1]	0.40	0.55	0.74	0.84	0.54	0.20	0.41	0.35	0.46	0.64
Social Studies	16.01***	1.40	-10.84**	-6.58	-17.42***	15.22**	-7.37**	-3.35	-4.51	-7.85
	(5.44)	(4.39)	(4.69)	(4.18)	(4.37)	(7.66)	(3.51)	(3.64)	(3.86)	(6.01)
Observations	125	125	125	125	125	65	65	65	65	65
R^2	0.37	0.23	0.17	0.27	0.37	0.32	0.51	0.32	0.22	0.26
Bandwidth	7.48	7.48	7.48	7.48	7.48	8.41	8.41	8.41	8.41	8.41
Over-id. test p-value[1]	0.65	0.81	0.40	0.63	0.31	0.24	0.11	0.41	0.50	0.44

*, **, ***: 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. The regressions in this table use 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.

Table 5: Effect of "Threatened Status" on Test Participation ("All Students" and Various Subgroups)

Panel A: All Students	Reading (1)	Lang. Arts (2)	Math (3)	Science (4)	Soc. Studies (5)	AYP Test Part. (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
Panel B: White Students	Reading (7)	Lang. Arts (8)	Math (9)	Science (10)	Soc. Studies (11)	AYP Test Part. (12)
Failed AYP	-1.61	-1.83	-1.48	-2.09	-2.68	-1.47
	(2.67)	(2.68)	(1.08)	(3.24)	(4.16)	(1.11)
Observations	595	595	595	595	595	595
Bandwidth	5.71	5.71	5.71	5.71	5.71	5.71
Over-id. test p-value[1]	0.67	0.70	0.75	0.75	0.77	0.72
Panel C: Econ. Disadv. Students	Reading (13)	Lang. Arts (14)	Math (15)	Science (16)	Soc. Studies (17)	AYP Test Part. (18)
Failed AYP	-2.03	-2.30	-1.46	-3.08	-3.78*	-0.97
	(1.45)	(1.52)	(1.00)	(1.96)	(2.20)	(0.93)
Observations	125	125	125	125	125	125
Bandwidth	7.48	7.48	7.48	7.48	7.48	7.48
Over-id. test p-value[1]	0.30	0.31	0.13	0.08	0.09	0.12
Panel D: Special Ed. Students	Reading (19)	Lang. Arts (20)	Math (21)	Science (22)	Soc. Studies (23)	AYP Test Part. (24)
Failed AYP	-3.42**	-3.77**	-4.78**	-6.80**	-7.08**	-3.43**
	(1.67)	(1.69)	(2.10)	(2.99)	(3.50)	(1.64)
Observations	65	65	65	65	65	65
Bandwidth	8.41	8.41	8.41	8.41	8.41	8.41
Over-id. test p-value[1]	0.08	0.23	0.17	0.35	0.34	0.18

*, **, ***: 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.

Table 6: Effect of "Threatened Status" on WRCT Scores and Participation

	% Minimal	% Basic	% Proficient	% Advanced	% Prof/Adv	% Tested
	(1)	(2)	(3)	(4)	(5)	(6)
Failed AYP	1.84	5.80	-3.74	-3.90	-7.64	-1.09
	(2.16)	(4.82)	(3.31)	(5.24)	(6.51)	(4.45)
Observations	683	683	683	683	683	683
R^2	0.27	0.58	0.21	0.59	0.59	0.48
Bandwidth	7.67	7.67	7.67	7.67	7.67	7.67
	Using Indicators of Cutoffs Failed as Additional Instruments					
	% Minimal	% Basic	% Proficient	% Advanced	% Prof/Adv	% Tested
	(7)	(8)	(9)	(10)	(11)	(12)
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

Table 7: Effect of "Threatened Status" on Attendance and Graduation Rates

	Attendance	Attendance[2]	Graduation	AYP Other Indicator
	(1)	(2)	(3)	(4)
Failed AYP	-1.14	2.92*	5.33	6.69**
	(1.53)	(1.52)	(4.18)	(2.98)
Observations	1329	984	352	1329
R^2	0.32	0.47	0.20	0.05
Bandwidth	6.69	7.08	8.86	6.69
	Using Indicators of Cutoffs Missed as Additional Instruments			
	Attendance	Attendance[2]	Graduation	AYP Other Indicator
	(5)	(6)	(7)	(8)
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

Footnotes for tables 6 and 7: *, **, ***: 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. [2] Uses sample of schools where attendance matters (elementary and middle).

Table 8: Are Compositional Changes or Sorting Driving Results? Investigating Demographic Shifts Using a Regression Discontinuity Analysis

Panel A	% White	% Black	% Hispanic	% Asian	% Am. Indian
	(1)	(2)	(3)	(4)	(5)
	1.56	15.10	-8.46***	-3.82**	-0.44
	(23.41)	(23.87)	(3.18)	(1.61)	(2.15)
Observations	1343	1343	1343	1343	1343
R^2	0.09	0.06	0.04	0.01	0.01
Panel B	% 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	
R^2	0.00	0.05	0.09	0.04	
Panel C	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	
R^2	0.02	0.06	0.04	0.01	
Panel D	Am. Indians Counted	Limited English Prof. Counted	Special Ed. Counted	Econ. Disadv. Counted	
	(25)	(26)	(27)	(28)	
	-0.01	-0.07*	-0.06	-0.40**	
	(0.05)	(0.04)	(0.20)	(0.17)	
Observations	1343	1343	1343	1343	
R^2	0.03	0.02	0.10	0.11	

*, **, ***: 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 9: Investigating the Response of Schools That Missed AYP by Missing Reading and/or Math only: Effect on Percent of Students Scoring in Various Proficiency Categories

						Using Indicators of Cutoff 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	1.58	-16.51***	4.71	10.21	14.92***	1.44	-10.25***	-1.55	10.37	8.82**
	(2.24)	(4.07)	(4.37)	(7.91)	(5.05)	(2.52)	(2.69)	(7.90)	(10.38)	(4.37)
Observations	110	110	110	110	110	110	110	110	110	110
R^2	0.33	0.22	0.22	0.62	0.48	0.33	0.43	0.25	0.60	0.54
Bandwidth	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00
Language Arts	-1.52	-20.14***	15.32**	6.34	21.66***	-3.20	-13.22***	11.22	5.20	16.42*
	(4.30)	(4.05)	(7.52)	(6.21)	(6.90)	(6.34)	(4.01)	(9.52)	(5.99)	(8.95)
Observations	110	110	110	110	110	110	110	110	110	110
R^2	0.25	.	.	0.33	0.21	0.24	0.16	0.09	0.34	0.29
Bandwidth	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00
Math	1.73	-6.32	-1.53	6.12	4.59	1.85	-9.21**	2.86	4.50	7.36
	(5.01)	(4.96)	(3.22)	(5.55)	(6.63)	(5.46)	(4.50)	(3.53)	(5.13)	(6.49)
Observations	110	110	110	110	110	110	110	110	110	110
R^2	0.61	0.13	0.43	0.50	0.60	0.61	0.05	0.43	0.50	0.60
Bandwidth	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00
Science	-11.80**	20.50***	-3.60	-5.10	-8.70*	-13.87**	16.20***	-0.71	-1.62	-2.33
	(5.05)	(4.50)	(5.13)	(7.26)	(4.81)	(6.76)	(4.00)	(5.10)	(6.56)	(5.94)
Observations	110	110	110	110	110	110	110	110	110	110
R^2	0.25	0.34	0.38	0.69	0.62	0.22	0.44	0.38	0.70	0.62
Bandwidth	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00
Social Studies	22.74***	-14.01***	-21.80***	13.07**	-8.73*	16.45***	-16.37***	-13.64**	13.56*	-0.08
	(6.52)	(4.28)	(6.55)	(5.08)	(4.80)	(5.07)	(5.95)	(6.11)	(6.99)	(3.39)
Observations	110	110	110	110	110	110	110	110	110	110
R^2	0.01	0.11	.	0.49	0.46	0.17	0.00	0.10	0.48	0.46
Bandwidth	10.00	10.00	10.00	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 passed test participation and "other indicator". Schools to the left missed AYP by just missing the reading/math cutoff. Schools to the right just made the reading/math cutoff.

Table 10: Investigating the Response of Schools That Missed AYP by Missing Reading and/or Math only: Effects on Test Participation, Attendance, Graduation, WRCT Scores and Participation

Panel A:	Test Participation						Other Indicators		
Test Part.	Read.	Lang. Arts	Math	Science	Soc. St.	AYP TP	Attend.	Attend(no grad)	AYP OI
Attend. and Grad.	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Failed AYP	-1.05	-1.05	-2.08	-3.28	-3.52	-1.03	-2.71	2.45**	3.65
	(2.18)	(2.18)	(2.17)	(2.52)	(2.24)	(2.18)	(1.77)	(1.09)	(12.21)
Observations	111	111	111	111	111	111	111	87	111
	Using Indicators of Cutoffs Missed as Additional Instruments								
Panel B:	Test Participation						Other Indicators		
Test Part.	Read.	Lang. Arts	Math	Science	Soc. St.	AYP TP	Attend.	Attend(no grad)	AYP OI
Attend. and Grad.	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)
Failed AYP	1.49	1.49	0.32	-0.44	-0.69	1.05	-2.18	2.43***	2.90
	(1.16)	(1.16)	(1.07)	(1.44)	(1.16)	(1.15)	(1.77)	(0.76)	(9.04)
Observations	111	111	111	111	111	111	111	87	111
Over-id. test p-value[1]	0.31	0.31	0.38	0.43	0.38	0.35	0.74	0.96	0.21

Panel C:	% Minimal	% Basic	% Proficient	% Advanced	% Prof/Adv	% Tested
WRCT	(19)	(20)	(21)	(22)	(23)	(24)
Failed AYP	3.03	9.77***	-0.25	-12.55**	-12.80***	-5.23
	(2.23)	(3.67)	(6.55)	(6.04)	(3.87)	(6.72)
Observations	59	59	59	59	59	59
	Using Indicators of Cutoffs Failed as Additional Instruments					
Panel D:	% Minimal	% Basic	% Proficient	% Advanced	% Prof/Adv	% Tested
WRCT	(25)	(26)	(27)	(28)	(29)	(30)
Failed AYP	1.37	5.84***	-4.21	-3.00	-7.21***	1.73
	(1.33)	(2.17)	(4.39)	(4.21)	(1.65)	(3.72)
Observations	59	59	59	59	59	59
Over-id. test p-value[1]	0.45	0.79	0.82	0.15	0.50	0.03

*, **, ***: 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. This RD design uses the sample of schools that passed test participation and "other indicator". Schools to the left missed AYP by just missing the reading/math cutoff. Schools to the right just made the reading/math cutoff.

Table 11: Investigating the Response of Schools That Missed AYP by Missing Test Participation only: Effect on Percent of Students Scoring in Various Proficiency Categories

	% 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 12: Effect of "Threatened Status" on Percent of Students Scoring in Various Proficiency Categories, Grades 4, 8, and 10

	Restricted Sample									
	Grade 4					Grade 8				
	% 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	4.11	-13.99*	6.94	2.94	9.89**	20.31	6.11*	-19.74	-6.68	-26.42
	(5.13)	(7.59)	(5.03)	(4.03)	(4.54)	(16.62)	(3.50)	(25.74)	(7.14)	(19.12)
Observations	691	691	691	691	691	446	446	446	446	446
Lang. Arts	-4.73***	-6.92*	8.92	2.72	11.64**	28.94	-1.41	-13.96	-13.58***	-27.53***
	(1.65)	(4.12)	(5.71)	(6.46)	(4.82)	(23.98)	(14.41)	(9.31)	(2.36)	(10.37)
Observations	691	691	691	691	691	446	446	446	446	446
Math	18.14	-8.47	-2.62	-7.04**	-9.66	24.98	-6.31	-10.38	-8.29	-18.67
	(16.38)	(5.61)	(13.80)	(4.39)	(12.48)	(19.71)	(8.60)	(15.42)	(5.12)	(11.47)
Observations	691	691	691	691	691	446	446	446	446	446
Science	5.07	4.98	-5.35	-4.70	-10.05	22.58**	-2.58	-13.03***	-6.97	-20.00***
	(3.64)	(8.23)	(8.86)	(4.35)	(11.38)	(11.12)	(11.94)	(4.44)	(5.91)	(2.56)
Observations	691	691	691	691	691	446	446	446	446	446
Social Studies	12.03	3.29	-9.84*	-5.47	-15.32	9.23	1.42	-1.05	-9.61	-10.65
	(7.89)	(3.21)	(5.54)	(5.77)	(10.59)	(6.39)	(0.92)	(12.30)	(6.42)	(6.72)
Observations	691	691	691	691	691	446	446	446	446	446

	Grade 10				
	% Min	% Basic	% Prf	% Adv	% At/Abv Prf
	(11)	(12)	(13)	(14)	(15)
Reading	0.93	3.48	-6.94	2.53	-4.41
	(4.88)	(7.25)	(4.67)	(8.26)	(10.82)
Observations	345	345	345	345	345
Lang. Arts	8.33	-3.41	-2.50	-2.42	-4.92
	(7.04)	(3.02)	(7.44)	(3.77)	(8.25)
Observations	345	345	345	345	345
Math	2.08	2.54	-8.25	3.62	-4.62
	(5.54)	(3.10)	(6.88)	(4.78)	(6.21)
Observations	345	345	345	345	345
Science	-0.66	7.00*	-4.09	-2.24	-6.34
	(4.48)	(4.09)	(7.12)	(4.77)	(6.42)
Observations	345	345	345	345	345
Social Studies	1.50	-3.20**	2.86	-1.16	1.71
	(6.11)	(1.50)	(5.25)	(4.29)	(6.28)
Observations	345	345	345	345	345

*, **, ***: 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 13: Did Competition Matter? Examining the Effect on Percent of Students Scoring in Various Proficiency Categories

	% Minimal	% Basic	% Proficient	% Advanced	% at/above Prof
	(1)	(2)	(3)	(4)	(5)
	Reading				
AYPfail	3.47	-4.99	-7.81	9.33*	1.52
	(6.12)	(3.64)	(6.51)	(5.18)	(6.70)
AYPfail * count	-5.51**	-3.76	12.52***	-3.24	9.28**
	(2.60)	(2.32)	(3.54)	(2.66)	(4.67)
Observations	1294	1294	1294	1294	1294
Bandwidth	6.69	6.69	6.69	6.69	6.69
	Language Arts				
AYPfail	0.77	-6.64*	2.62	3.25	5.87
	(3.92)	(3.49)	(5.29)	(8.07)	(6.96)
AYPfail * count	-4.36	-6.41***	5.18	5.60**	10.77**
	(3.37)	(2.06)	(5.01)	(2.34)	(5.25)
Observations	1294	1294	1294	1294	1294
Bandwidth	6.69	6.69	6.69	6.69	6.69
	Math				
AYPfail	7.86	-4.24	-1.61	-2.01	-3.62
	(7.18)	(3.44)	(4.53)	(3.62)	(6.46)
AYPfail * count	3.61	-3.76***	-7.28**	7.43**	0.85
	(2.99)	(1.07)	(3.06)	(2.09)	(2.89)
Observations	1294	1294	1294	1294	1294
Bandwidth	6.69	6.69	6.69	6.69	6.69
	Science				
AYPfail	-0.03	2.48	1.47	-3.92	-2.45
	(5.65)	(4.51)	(9.08)	(3.65)	(7.54)
AYPfail * count	-11.88***	8.19***	9.64**	-5.96***	3.68
	(4.53)	(1.62)	(4.72)	(2.08)	(5.08)
Observations	1294	1294	1294	1294	1294
Bandwidth	6.69	6.69	6.69	6.69	6.69
	Social Studies				
AYPfail	11.70	-5.09*	-8.45	1.84	-6.61
	(11.07)	(2.90)	(5.33)	(9.25)	(10.09)
AYPfail * count	-8.31***	-3.57*	-7.32***	19.20***	11.88***
	(3.16)	(2.04)	(2.49)	(3.89)	(4.47)
Observations	1294	1294	1294	1294	1294
Bandwidth	6.69	6.69	6.69	6.69	6.69

*, **, ***: 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 14: Did Competition Matter? Examining the Effect on Test Participation, Attendance, and Graduation Rates

Panel A	Test Participation					
	Reading	Language Arts	Math	Science	Soc. Studies	AYP Test Part.
	(1)	(2)	(3)	(4)	(5)	(6)
AYPfail	-0.78	-0.77	-0.72	-2.25	-2.93	-0.26
	(1.18)	(1.21)	(1.35)	(2.04)	(2.38)	(0.90)
AYPfail * count	4.44***	4.81***	5.00***	4.97***	5.56***	4.37***
	(1.09)	(1.06)	(1.32)	(1.39)	(1.66)	(1.08)
Observations	1295	1295	1295	1295	1295	1295
Bandwidth	6.69	6.69	6.69	6.69	6.69	6.69

Panel B	Attendance and Graduation			
	Attend.	Attend. [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. All regressions include controls for 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.

A1: First Stage Regressions

	All Criteria		Failed AYP by Failing Rdg/Mth		Failed AYP By Failing Test Part.
	(1)	(2)	(3)	(4)	(5)
F	0.77***	0.43**	0.46***	0.66*	0.70***
	(0.17)	(0.19)	(0.11)	(0.34)	(0.18)
Missed Reading Cutoff		0.11		-0.24	
		(0.16)		(0.25)	
Missed Math Cutoff		0.14		-0.13	
		(0.21)		(0.30)	
Missed Test Participation Cutoff		0.39*			
		(0.20)			
Missed Other Indicator Cutoff		0.19			
		(0.30)			
Observations	1328	1328	110	110	1296
R^2	0.42	0.44	0.30	0.32	0.38
Bandwidth	6.69	6.69	10.00	10.00	10.00
F-test of excl. instr.[1]	21.12	6.14	16.81	13.05	15.45

A2: Examining Effect of "Threatened Status" after Controlling for Subgroup Accountability

						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	1.84	-5.95	-6.51	10.61**	4.10	-0.84	-5.05	-7.22	13.12**	5.89
	(5.63)	(4.49)	(6.06)	(5.39)	(5.64)	(3.59)	(3.31)	(5.84)	(5.23)	(5.12)
Observations	1328	1328	1328	1328	1328	1328	1328	1328	1328	1328
R^2	0.28	0.29	0.16	0.43	0.35	0.27	0.29	0.16	0.43	0.35
Bandwidth	6.69	6.69	6.69	6.69	6.69	6.69	6.69	6.69	6.69	6.69
Lang. Arts	-1.00	-7.50**	3.55	4.95	8.50	-3.86	-7.04**	5.25	5.65	10.89*
	(3.36)	(3.32)	(5.80)	(7.90)	(5.84)	(3.03)	(2.98)	(5.65)	(6.81)	(5.56)
Observations	1328	1328	1328	1328	1328	1328	1328	1328	1328	1328
R^2	0.31	0.22	0.06	0.21	0.33	0.30	0.22	0.06	0.21	0.32
Bandwidth	6.69	6.69	6.69	6.69	6.69	6.69	6.69	6.69	6.69	6.69
Math	8.64	-5.62	-2.15	-0.87	-3.02	1.26	-5.79*	2.58	1.95	4.53
	(9.10)	(3.40)	(6.24)	(3.29)	(7.41)	(5.76)	(3.00)	(4.14)	(3.28)	(5.09)
Observations	1328	1328	1328	1328	1328	1328	1328	1328	1328	1328
R^2	0.51	0.16	0.17	0.30	0.47	0.51	0.16	0.16	0.30	0.46
Bandwidth	6.69	6.69	6.69	6.69	6.69	6.69	6.69	6.69	6.69	6.69
Science	-2.92	3.66	3.36	-4.09	-0.74	-2.96	-2.50	7.62	-2.17	5.45
	(4.16)	(5.69)	(8.07)	(3.55)	(7.15)	(4.07)	(3.53)	(6.26)	(3.37)	(5.30)
Observations	1328	1328	1328	1328	1328	1328	1328	1328	1328	1328
R^2	0.35	0.48	0.18	0.39	0.52	0.35	0.49	0.17	0.40	0.51
Bandwidth	6.69	6.69	6.69	6.69	6.69	6.69	6.69	6.69	6.69	6.69
Social Studies	10.61	-6.11**	-10.50	6.00	-4.50	3.14	-5.21***	-5.82	7.88	2.07
	(10.27)	(3.05)	(6.70)	(7.06)	(8.64)	(4.93)	(1.81)	(4.15)	(6.12)	(4.93)
Observations	1328	1328	1328	1328	1328	1328	1328	1328	1328	1328
R^2	0.20	0.38	0.16	0.31	0.37	0.21	0.39	0.16	0.31	0.35
Bandwidth	6.69	6.69	6.69	6.69	6.69	6.69	6.69	6.69	6.69	6.69

Note for tables A1 and A2: *, **, ***: 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, real per pupil expenditures and whether each subgroup was accountable.

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

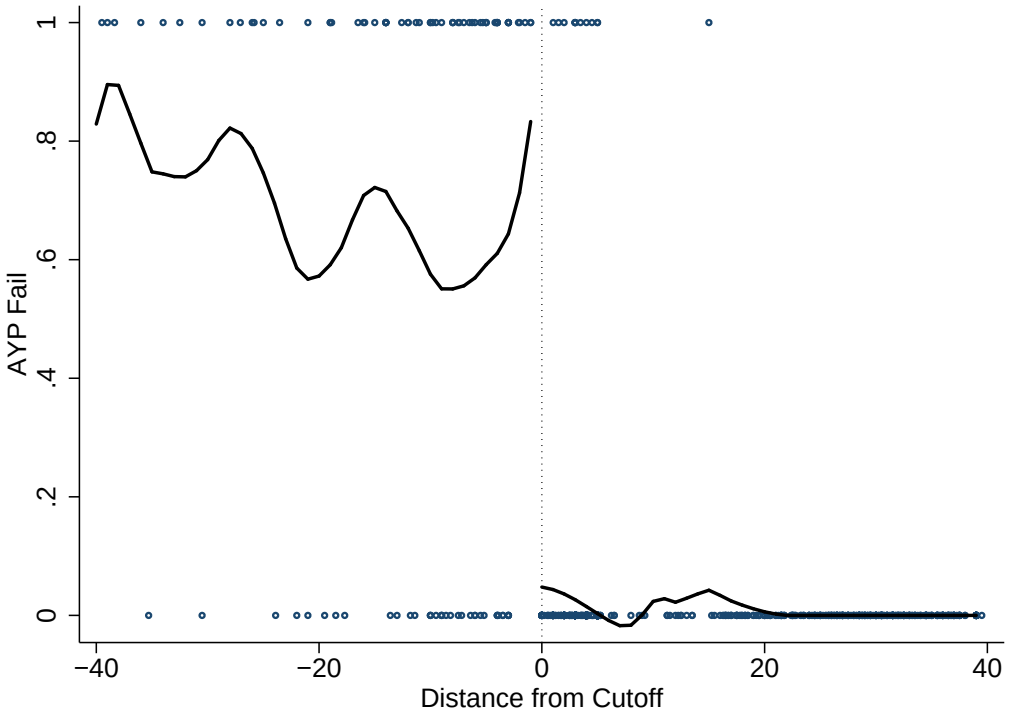


Figure 2: Testing Validity of the Regression Discontinuity Strategy: Are Pre-Program Characteristics Smooth through the Cutoff?

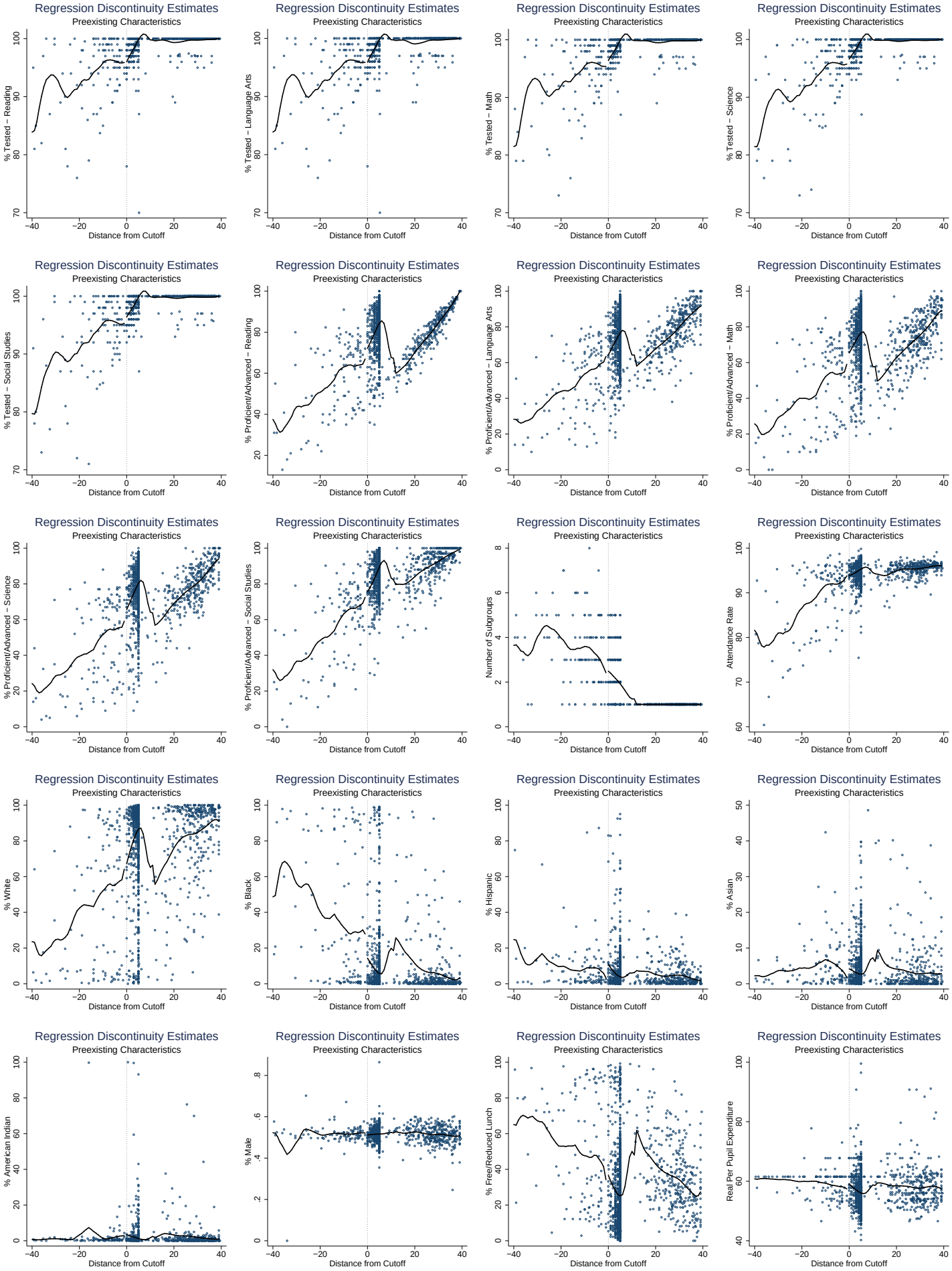


Figure 3: Testing Validity of the Regression Discontinuity Analysis: Is there a Discontinuity in the Density of the Running Variable at the Cutoff?

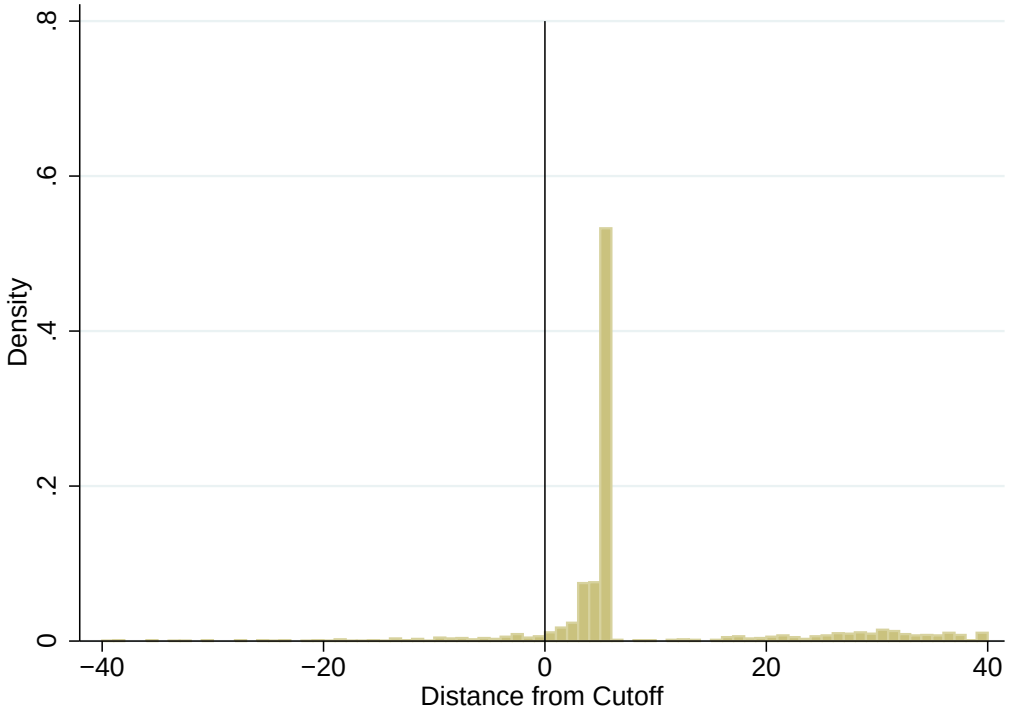


Figure 4: Distribution of Test Participation in 2001-02

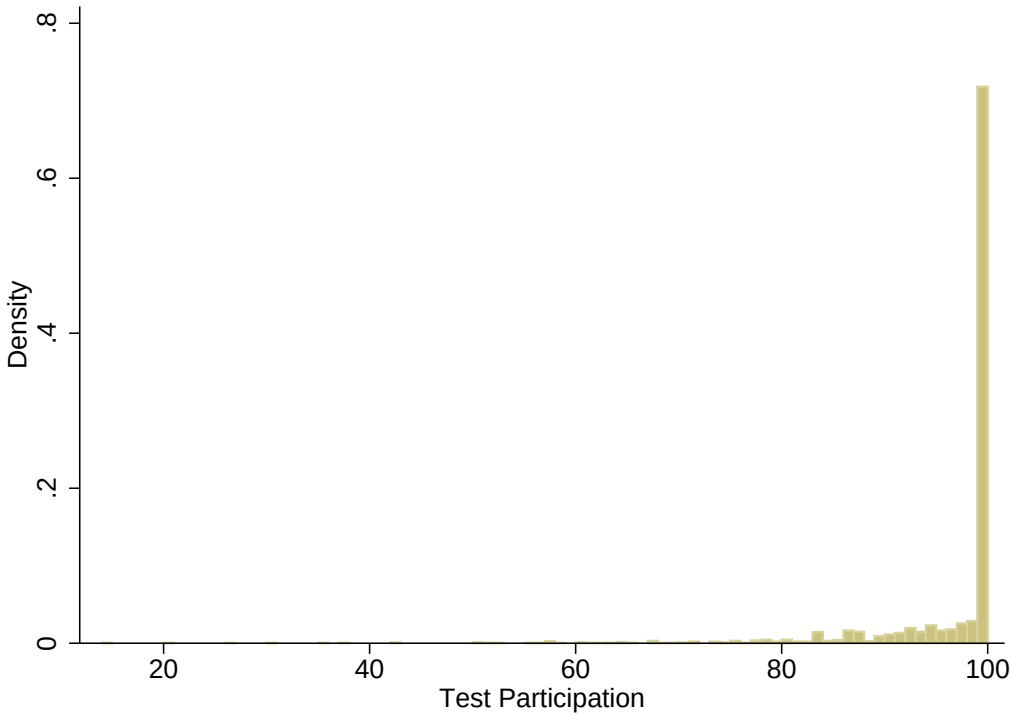


Figure 5: Plotting Means of Pre-Program Characteristics

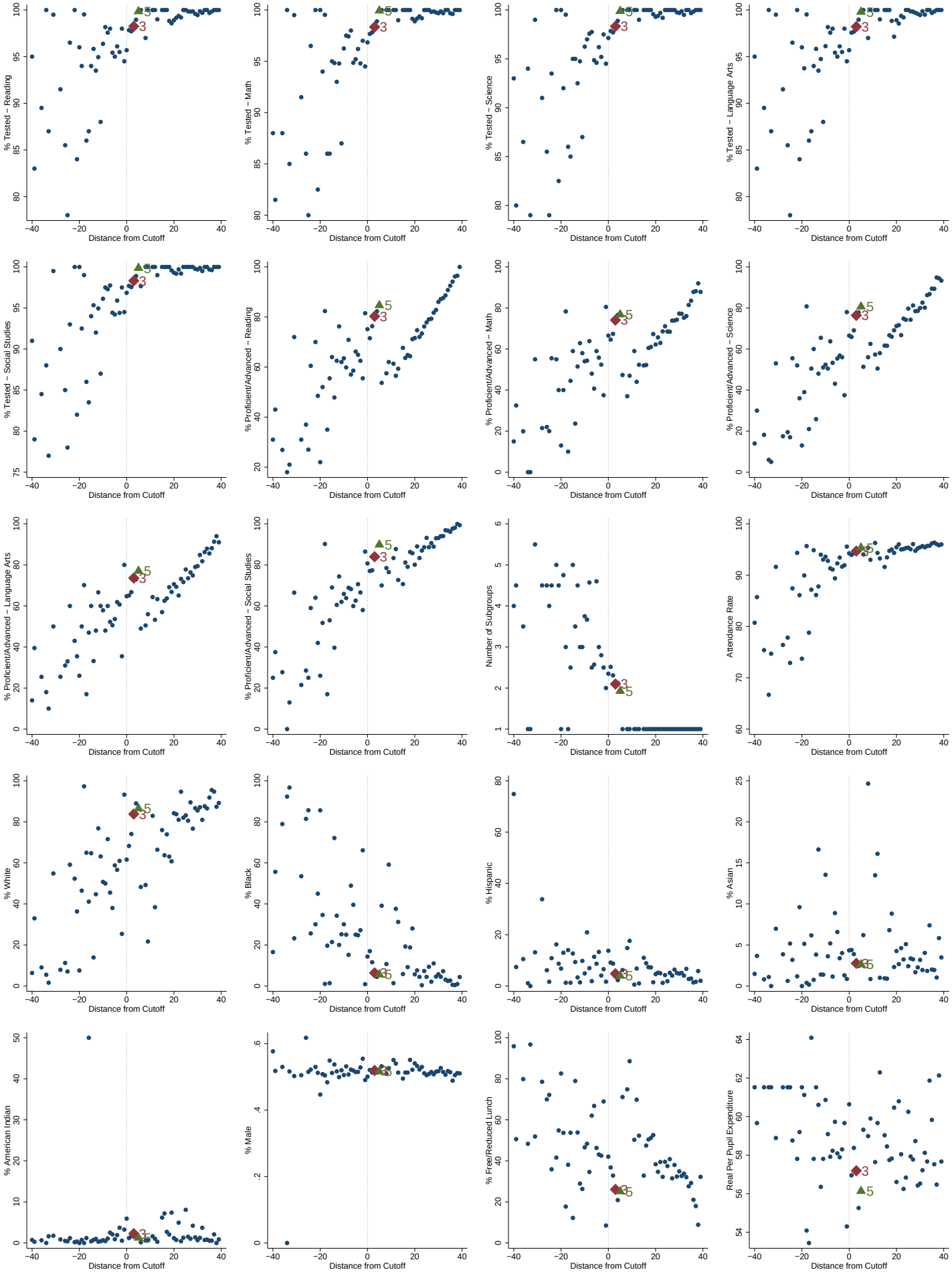
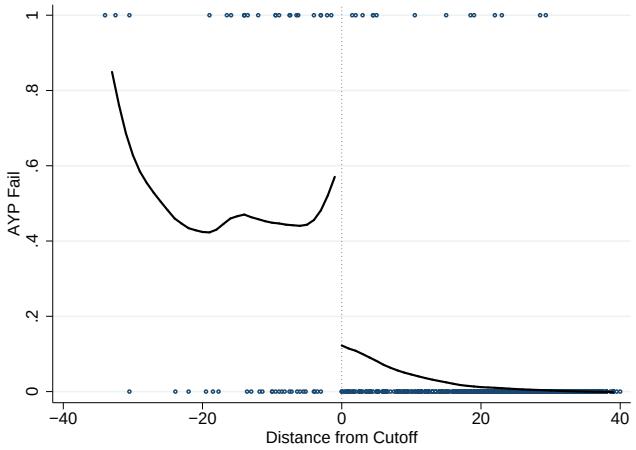
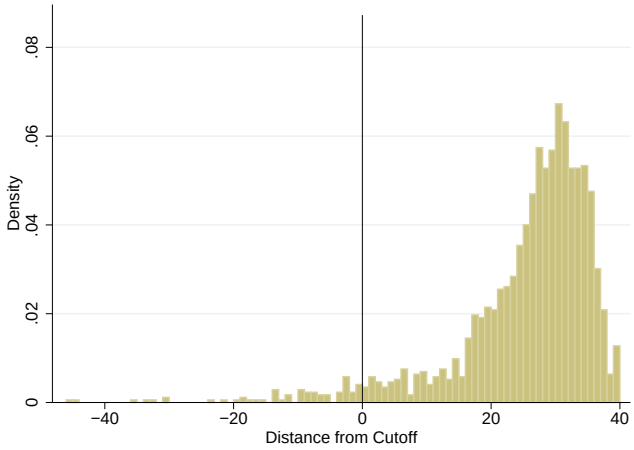


Figure 6: Relationship Between Treatment Status and Running Variable (Distance from Reading/Math Cutoff)



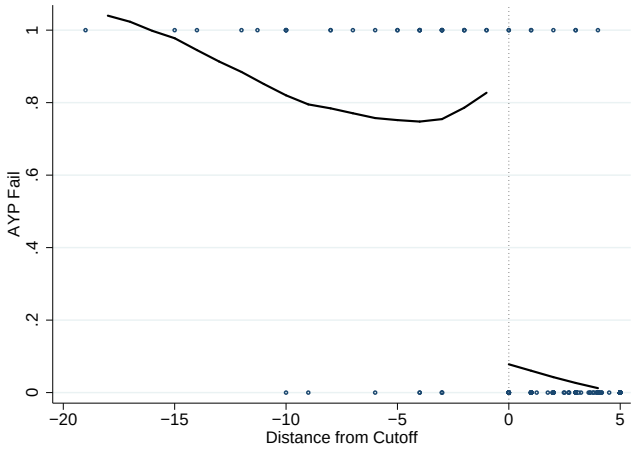
The running variable is distance from the reading/math cutoff in the sample of schools that made test participation and "other indicator".

Figure 7: Is there a Discontinuity in the Density of the Running Variable at the Cutoff?



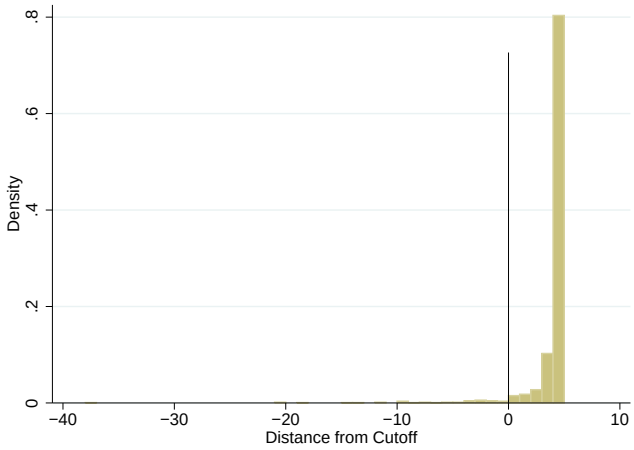
The running variable is distance from the reading/math cutoff in the sample of schools that made test participation and "other indicator".

Figure 8: Relationship Between Treatment Status and Running Variable (Distance from Test Participation Cutoff)



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

Figure 9: Is there a Discontinuity in the Density of the Running Variable at the Cutoff?



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