

**The Impact of No Child Left Behind's Accountability Sanctions on School Performance:
Regression Discontinuity Evidence from North Carolina**

Thomas Ahn
University of Kentucky

Jacob Vigdor
Duke University and NBER

Preliminary Draft

February 2013

Abstract

Comparisons of schools that barely meet or miss criteria for adequate yearly progress (AYP) reveal that some sanctions built into the No Child Left Behind accountability regime exert positive impacts on students. Estimates indicate that the strongest positive effects associate with the ultimate sanction: leadership and management changes associated with school restructuring. We find some positive incentive effects in schools first entering the NCLB sanction regime, but no significant effects of intermediate sanctions. Further analysis shows that gains in sanctioned schools are concentrated among low-performing students, with the exception of gains from restructuring which are pervasive. We find no evidence that schools achieve gains among low-performing students by depriving high-performing students of resources.

*The authors gratefully acknowledge support from the Institute for Education Sciences, grant #R305A090019. We thank Sarah Crittenden Fuller, James Riddlesperger, and John Holbein for outstanding research assistance. We also thank Mike Lovenheim, Scott Imberman, and participants at the 2013 AEA meetings for helpful comments on a previous draft. Any opinions expressed in this paper are those of the authors and not of any affiliated institution.

1. Introduction

The No Child Left Behind Act of 2001 (NCLB) required schools receiving Federal Title I funding to track student performance, and to implement an escalating series of negative sanctions in the event performance fell below a state-established threshold. This paper analyzes the impact of these sanctions, and the threat of these sanctions, on student performance. We use school- and student-level data covering North Carolina public schools to compare student test score growth in schools subjected to varying sanctions. Since the NCLB sanction regime treated schools very differently if the performance of a subgroup of students lay on either side of an arbitrary threshold, we use regression discontinuity (RD) techniques to achieve econometric identification.

This paper contributes to a large and growing literature on educational accountability. Accountability regimes are intended to improve student performance by introducing incentives that some argue are absent in traditional public education systems. They rely on the existence of a suitable measure of student performance – ideally one correlated with long-run student well-being and difficult to manipulate by means other than providing high-quality education to all students – and sanctions, whether positive or negative, that sufficiently motivate educators to improve the quality of education provided. Previous literature, summarized below, has found both encouraging and cautionary evidence regarding the overall effects of educational accountability.

Our contribution assesses whether the sanctions incorporated into the nation's most comprehensive accountability regime are themselves positive or negative interventions, when assessed by their ability to yield significant year-over-year improvements in student test scores. On the one hand, the intent of NCLB is in some sense to force schools to implement interventions that improve education, potentially at some effort cost to educators. On the other

hand, the incentive effects of sanctions are maximized when they are costly to the sanctioned, and educators motivated to serve their students might be particularly moved by the opportunity to save them from harm.

At a moment in time when the future of No Child Left Behind is uncertain, our evidence may be of some use to policymakers interested in implementing the sanctions as stand-alone policy interventions outside the context of an accountability system. We specifically analyze effect heterogeneity by student subgroup and at varying points in time.

Our results indicate that the sanctions incorporated into NCLB had widely varying effects. On average, students attending schools that barely missed the cutoff for Adequate Yearly Progress (AYP) in a given year posted test score improvements at most slightly greater than those in schools that barely made the cutoff. Subsequent analyses stratifying schools by the sanction for which they are at risk reveals a significant mix of estimated effects. We find modest evidence of improvements for schools first entering the NCLB sanction regime, and more significant evidence of improvements among schools that become subject to the ultimate sanction in the system, restructuring incorporating leadership or management change. Intermediate sanctions, including mandatory tutoring services for low-income students, have no demonstrable effect.

Previous literature has raised concerns that a focus on proficiency will lead schools to reallocate resources away from higher-performing students, or more generally that incentives to focus on one group will result in a redirection of resources away from other groups. We assess the importance of these concerns in two ways. First, we determine whether sanctions for low proficiency result in poorer outcomes for students well above the proficiency threshold. While we do find that the beneficial impacts of early-stage sanctions are concentrated among lower-

performing students, we find no evidence of countervailing negative impacts on higher-performing students. In fact, the estimated impacts of restructuring appear to accrue throughout the test score distribution. North Carolina's independent accountability system, which focuses on test score growth rather than accountability, may have restrained any impulse to deprive high-performing students.

Second, we examine whether test score improvements in schools sanctioned for the performance of one specific subgroup are concentrated among members of that subgroup. A concentration of resources along these lines would be a rational response to the incentive. We find evidence of particularly strong effects within the targeted subgroup, with the exception of the restructuring sanction, which appears to generate improvements across the board.

Overall, our results suggest that accountability systems can have modest impacts on student performance, and if properly designed can in fact improve the performance of some students without harming others. The association of strongest, and broadest, effects with restructuring indicates that management and leadership issues are the most significant obstacles to strong performance in public schools.

2. Accountability Sanctions in Theory

Economic theory holds that competitive pressure forces firms to produce efficiently in order to survive in the marketplace. School accountability regimes follow from the notion that the market for primary and secondary education is not likely to mimic perfect competition. Monopoly power is the most commonly cited source of market failure in education (Friedman 1962). The question of whether education is in fact delivered more efficiently in markets with greater competition among producers is controversial and unsettled (Hoxby 2000; Rothstein

2007). Nonetheless, since the 1990s a number of interventions in the market for public education – school vouchers and charter schools most significant among them – have attempted to improve the efficiency of public education by introducing competition (Ladd, 2002). Research on the impacts of these interventions is also controversial and unsettled (Rouse 1998; Bettinger 2005; Bifulco and Ladd 2006; Angrist, Bettinger, and Kremer 2006; Sass 2006; Chakrabarti 2008a, 2008b).

Even in the absence of monopoly power, information asymmetries pose a more significant threat to the efficient operation of education markets. The value of educational services received is fully revealed many years after the services are provided. Short-term proxy measures, such as value-added measures based on standardized test scores, are available – but research has only recently verified that value-added measures correlate with longer-run measures of productivity (Chetty, Friedman, and Rockoff 2011). Value-added measures, which rely on longitudinal student data, are only rarely made available to the public; assessments of public schools more commonly rely on basic proficiency statistics which conflate the quality of educational production with the baseline characteristics of the students educated. Consumers of public education show evidence of being imperfectly informed, as simple informational interventions have been shown to significantly alter behavior (Hastings and Weinstein, 2008). There is also evidence of consumer reaction to simplified information about school performance, such as categorical rankings that are based on a publicly-available, continuous measure (Figlio and Lucas 2004; Martinez 2011).

Either market power or information asymmetries are sufficient to disrupt the consumer choice mechanism traditionally theorized to root out inefficiency in competitive markets. Accountability sanction regimes are intended to provide another mechanism to root out

inefficiency. The basic theoretical justification follows from standard principal-agent modeling: the principal (in this case a governmental agency) receives a signal of productivity from the agent (in this case a school) and applies a reward or punishment on the basis of that signal. Provided that the signal is sufficiently well-correlated with output, and output with effort, the agent is incentivized to engage in costly effort rather than shirk.

The educational setting introduces a series of complications to the simple principal-agent framework. Productivity signals are typically based on test scores, and test score statistics can be manipulated with a number of short-term strategies with dubious effects on students' long-term well-being (Jacob and Levitt 2003; Figlio and Winicki 2005; Figlio 2006). Test score statistics can also incentivize schools to focus attention on a small group of students at the expense of others (Reback 2008; Neal and Schanzenbach, 2010).

The design of sanctions themselves is also fraught with complication (Kane and Staiger 2002). Sanctions must be significant enough to incentivize agent behavior, but not so significant that they cause harm to the children being educated when implemented. For example, a threat to halve per-pupil funding in poor-performing schools may influence agent behavior, but the implementation of such a sanction might cause undue harm to future generations of students. These concerns are less salient when the sanctions are rewards rather than punishments, but systems of rewards are costlier to implement. In school accountability regimes, the ideal punishment is one that imposes significant utility costs on the agents without translating into degradation of the educational process. In fact, the goal of some punishments might be to force ineffective agents below their reservation utility, at which point they resign and are replaced by more-effective agents.

Despite the many possible pitfalls associated with the design and implementation of school accountability systems, prior research has found promising evidence of positive impacts on student outcomes (Ladd 1999; Hanushek and Raymond 2005; West and Peterson 2006). At the same time, research has illuminated some potential drawbacks of accountability systems, particularly to the extent they place low-performing schools at a disadvantage in the teacher labor market (Clotfelter et al. 2004).

3. No Child Left Behind in Practice

The accountability sanction regime introduced by the No Child Left Behind Act of 2001 assessed schools on the basis of the rate at which students attain state-defined proficiency standards. Assessments are based not only on the overall performance of the student body but also the performance of specific subgroups of students, defined by race, participation in the Federal free and reduced price lunch program, English proficiency, and special education status, provided that a school's enrollment of students in that subgroup exceeds a state-defined threshold. The "adequate yearly progress" (AYP) standard is met if proficiency rates for each of these subgroups exceeds a predetermined threshold.¹ The AYP standard is assessed independently in the subjects of reading and math.

Schools may also qualify for the AYP standard if subgroups falling below the predetermined threshold meet the requirements for one of two exceptions. The *confidence interval* exception effectively allows subgroups to be considered as meeting the target if their observed proficiency rate is close enough to the predetermined threshold that the statistical hypothesis that the "population" rate equals or exceeds the threshold cannot be rejected. The

¹ The AYP standard also incorporates a participation requirement, mandating that at least 95% of enrolled students are tested.

safe harbor exception exempts schools from negative sanctions if the subgroups with subpar proficiency have demonstrated a tangible improvement from one year to the next.

In the event a school fails to qualify for AYP either through straight examination of proficiency or either exception, it is subjected to an escalating series of negative sanctions. The sanction regime is summarized in Table 1. After failing to meet the AYP standard in the same subject for two consecutive years, schools are required to offer their students the opportunity to transfer to a higher-performing school in the same district. After a third failure, schools must offer supplemental education services, such as tutoring, to qualifying students. Should the school fail a fourth time, more fundamental changes in leadership or staffing are required. Further failures result in the school being required to develop and implement a more comprehensive restructuring plan. Schools that have entered the NCLB sanction system may exit after meeting the AYP standard for two consecutive years.

The NCLB sanction regime does not incorporate supplemental funds. Rather, participation in the regime is required for those schools who wish to maintain Federal Funding through the Title I program. In practice, many states have implemented the NCLB regime system in non-Title I schools as well. The imposition of sanctions thus might be hypothesized to have either a positive or negative impact. If sanctions require schools to reallocate resources away from more-productive uses, then they might have a negative effect. Sanctions might, however, result in improved efficiency – in fact, that is the basic rationale for the regime.

It is important to note that the existence of positive or negative effects of any particular sanction is neither necessary nor sufficient for the NCLB regime to have a positive effect on educational outcomes. In part this is because the regime is hypothesized to have systemic effects on all schools, whether exposed to sanctions or not. Additionally, sanctions with negative effects

might be productive if they induce schools to take extraordinary actions to avoid them. While we might find negative impacts of the sanction in such a scenario, this negative effect could theoretically be partially, fully, or more than fully offset by the efforts that schools undertook to avoid it.

While this research cannot in itself address the basic question of whether NCLB has been beneficial, information on the benefit or harm done by individual sanctions is still potentially useful to policymakers interested in implementing interventions along the lines of those incorporated into the sanction regime.

Previous investigations of the impact of No Child Left Behind have reported mixed results. Difference-in-difference style analysis inferring the total impact of the law by comparing trends in states with and without pre-existing accountability programs yields some positive results (Dee and Jacob 2011). An early evaluation of sanctions imposed in Florida finds no impact, however (West and Peterson 2006).

4. Data and Methods

We seek to analyze whether the imposition, or mere threat, of NCLB sanctions produces significant improvements or declines in student performance. To do so, we make use of student and school-level data from North Carolina, covering the period from inauguration of the NCLB system in the 2002/03 school year to 2009/10. These data are a mix of public records and anonymized data made available to researchers by the North Carolina Educational Research Data Center (NCERDC). To analyze the impact of sanctions on student performance, we use longitudinally linked data on students who attend the same school in consecutive years. This requires us to focus on schools that serve students in grades 3-8.

At present, AYP results from five school years are combined with test score results from those years and the following years to form the heart of our dataset. The dataset consists of over 8,000 school-by-year observations and over 1.7 million individual-level computations of test score gains.²

4.1 Regression Discontinuity

A simple comparison of student performance across schools subject or not subject to sanctions in a given year will be prone to yielding misleading results, since the sanctions are applied exclusively to schools in which proficiency rates are relatively low. To overcome this obstacle, we take advantage of the structure of the NCLB sanction system, in which schools with very similar patterns of student proficiency might be subject to very different sanctions if their small differences place them on either side of the AYP threshold. Under certain assumptions, this regression discontinuity analysis reveals an estimate of the average treatment effect of interventions applied at the AYP threshold, local to the set of schools that find themselves close to that threshold.

In practice, it is not trivial to identify the set of schools whose performance places them within an arbitrarily small neighborhood around the AYP threshold. Schools can miss AYP in a subject if even one student subgroup posts a sufficiently low proficiency rate. Moreover, because of the confidence interval and safe harbor exemptions, schools can make AYP even when numerous subgroups fail to post proficiency rates that place them above the state-defined standard.

² Sample sizes for reading and math analyses vary because in some small schools in some years we are unable to match test score performance for any students from one year to the next.

To identify schools that barely meet or miss the AYP standard, we first compute for each subgroup in each school an “effective” proficiency threshold, defined as follows:

$$\text{Effective threshold} = \min(\text{boundary for CI exemption}, \text{boundary for SH exemption})$$

Where CI and SH stand for the two exemptions made available to schools. The boundary for the CI exemption is determined by the confidence interval formula employed by the state and the number of students assessed. The boundary for the SH exemption is determined by the subgroup’s performance in the prior academic year.

We then compare the actual proficiency rate exhibited by a subgroup within a school with the effective threshold for that subgroup:

$$\text{Subgroup gap} = \text{actual proficiency} - \text{effective threshold}$$

The assignment variable for our RD analysis is defined as the minimum subgroup gap, across that set of subgroups with a sufficiently large number of tested students. In North Carolina, the threshold is set at 40 students. When the assignment variable is positive, the school qualified for AYP; when it is negative the school fails to make AYP. The cutoff for AYP, which we refer to as the minimum effective threshold, is zero.

Note that the assignment variable is not necessarily determined by the subgroup that performs worst in absolute terms. The worst-performing subgroup may qualify for AYP by an exemption in situations where the school as a whole fails to qualify. This would occur when the worst-performing group is numerically small and qualifies under the confidence interval

exemption (and the next-worst performing group is large and does not), or when the worst-performing group has shown significant improvement over the prior year, thereby qualifying under safe harbor (and the next-worst performing group does not).

To assess the impact of exposure to NCLB sanctions and threats, we evaluate the performance of students in the affected school in the subsequent academic year. The dependent variable of interest is the change in standardized test score over the period in which the treatment is administered. Thus, when a school meets or misses the AYP cutoff in year t , we examine the effects by observing growth in test scores between the end of year t and the end of year $t+1$. In our baseline specifications, we include only those students observed attending the school in question in both years. In alternate specifications, we include students who attend the school in year t regardless of the school attended in year $t+1$, so long as the students remain enrolled in any North Carolina public school. This will permit us to include the hypothesized treatment effect of offering transfers, and other family school choice behavior occasioned by the failure of a child's school to make AYP. We caution, however, that these short run effects may be dominated by disruption effects that dissipate over time (Hanushek, Kain and Rivkin 2004).

Regression discontinuity analysis can be performed either parametrically or nonparametrically. For our analyses, we use the Hahn, Todd, and van der Klaauw (2001) nonparametric specification, which entails the estimation of local linear regressions to fit a smooth function to either side of the discontinuity. While there is no explicit functional form choice associated with the local linear regression, in practice the shape of the smooth function is heavily influenced by choice of bandwidth in the local linear regression – put simply, the number of data points used to estimate a slope in the neighborhood of each individual data point. We report results for a variety of bandwidths.

The nonparametric analysis is complicated by the nature of the data: we have individual-level records for purposes of computing treatment effects, but assignment to the treatment occurs at the school rather than individual level. In our analyses below, we address this problem by collapsing the individual-level data to the school level, weighting observations in the school-level estimation by the number of individual-level observations used in computing school-specific means of the dependent variable and covariates.³

4.2 Analyzing the Impact of Specific Sanctions

Our primary goal is to assess the impact of individual NCLB sanctions. As such, a simple RD analysis on the entire dataset is unlikely to be informative, as it represents the average effect of a number of different threats and sanctions. For most of our analyses, then, we focus our attention on sets of school/year observations at risk for the same sanction or threat. For example, to analyze the impact of being threatened with the requirement to offer transfers, we will begin with the set of schools with no history of missing AYP prior to year t , and compare student performance in year $t+1$ across schools that make or miss AYP in year t .

While restriction to schools with similar AYP histories can help refine the sanction or threat under analysis, we are still left with the difficulty that each failure to make AYP after the first results in the simultaneous imposition of one sanction and the threat of the next sanction in the sequence shown in Table 1. Thus in the absence of additional information it is impossible to separate the effect of offering transfers from the effect of the threat of offering tutoring.

In practice, we take advantage of a waiver to standard NCLB policy that was offered to seven North Carolina districts beginning in the 2006/07 school year, and expanded statewide

³ We also estimate models using individual-level data, using a bootstrapping procedure to approximate clustered standard errors. While point estimates are comparable with both procedures, the clustered standard errors are larger, consistent with the notion that clustering is a conservative solution to the problem of grouped data.

beginning in 2008/09.⁴ The waiver permitted affected districts to reverse the order of the first two NCLB sanctions, implying that schools would be required to offer supplemental services first and transfers second. Thus, under the assumption that the effects of each threat and sanction are time invariant, we can uniquely identify not only the impact of being threatened with offering transfers, but the threat of offering supplemental services, the transfers themselves, the supplemental services themselves, and the threat of corrective action.

Of course, there is no guarantee that these effects will be time-invariant. Revelation of information about the effects of sanctions might well alter responses to the threat of those sanctions.

4.3 RD diagnostics

A standard concern with regression discontinuity analysis is that agents may have some capacity to manipulate the assignment variable. This would appear to be a valid concern in this application, as the effective thresholds for each subgroup are in theory calculable at the beginning of each school year. There is at least some evidence that school personnel might manipulate testing data in a high-stakes setting (Jacob and Levitt, 2003).

To assess this and other potential threats to validity, Figures 1-3 perform basic diagnostics appropriate to RD analysis (Imbens and Lemieux 2008; Lee and Lemieux, 2010). Figure 1 begins by showing a basic density plot of our data, pooled across all available years, for reading and math. In these plots, we would be particularly concerned if we found evidence of a mass on the right hand side of the minimum effective threshold relative to the left hand side.⁵

⁴ The seven districts are Burke County, Cumberland County, Durham County, Guilford County, Moore County, Northhampton County, and Pitt County. Three of these districts are among the state's ten largest.

⁵ It should be noted that this analysis differs from a typical RD analysis in one important respect: some of the schools "at risk" for sanctions have already had sanctioned applied. To the extent that there is a positive treatment

The figure demonstrates that the AYP cut point falls very nearly in the middle of the observed data, and that the density is in fact slightly greater to the immediate right of the cut point. The histogram bar to the immediate right is about 20% higher than the one immediately to the left. A formal statistical test for a significant break in density at the cut point fails to reject the null hypothesis of no difference (test statistic 0.094, standard error 0.079, see McCrary 2008).

To further investigate, Figure 2 plots average student characteristics for schools as a function of the minimum subgroup gap. Included characteristics include percent nonwhite, percent receiving free or reduced lunch, and percent limited English proficiency. While it is clear that schools with lower performance tend to serve more disadvantaged students, there is no evidence of a significant trend or break in trend in any of these student characteristics at the AYP threshold.

Finally, Figure 3 verifies that there is indeed a discontinuous change in the probability of treatment at the point where schools cross the AYP threshold. The magnitude of the probability, revealed in nonparametric specifications based on local linear regression, is 0.854 for reading and 0.828 for math. The “fuzziness” of the discontinuity, or the imperfections in our ability to forecast which schools will make or miss AYP in a given year, stems from the fact that factors other than test score performance are incorporated into AYP designations. Schools must meet attendance and test participation benchmarks.

Following our discussion of results below, we address an additional potential concern with regression discontinuity analysis in this application. Given that schools must exhibit low performance to miss AYP, there is some concern that simple mean reversion may lead us to

effect of any sanction, this may lead to a scenario where the density to the right of the discontinuity exceeds that on the left. Coupled with the insignificant McCrary test statistic reported below, this increases our confidence that manipulation is not a significant concern in these data.

estimate spurious positive impacts on test score growth in the following year. In theory, the RD design should address this concern, as we have no reason to think that the degree of mean reversion would jump discontinuously at the AYP threshold. Nonetheless, we report the estimates below of pseudo-RD analyses designed to determine whether mean reversion is a concern.

5. Results

In all results tables below, the signs obtained from RD analysis have been reversed, so that we might interpret the reported effects as the impact of being exposed to NCLB sanctions or threats, rather than the impact of avoiding them.

5.1 The main effect of failure to make AYP

Table 2 reports regression discontinuity estimates of the effect of meeting AYP criteria using all school-year observations. As noted above, these estimates conflate a wide variety of NCLB sanctions and threats. We report estimates derived from local linear regressions using the “optimal” bandwidth (Imbens and Kalyanaraman 2009), but note in the table whether results are statistically significant in specifications using twice or half this bandwidth. These specifications, as well as those reported below, condition on student race, free/reduced price lunch participation, and English proficiency status.

For reading test scores, estimation with optimal bandwidth yields an exceptionally small point estimate, on the order of two-hundred-thousandths of a standard deviation. The effect is estimated precisely enough to rule out any positive or negative impacts greater than one percent of a standard deviation. Estimation using alternate bandwidths from 50% to 300% of the optimal

value yields similarly insignificant results. The point estimate remains extremely small when we include students who transfer away from the school in the sample.

When analyzing math test scores, evidence is suggestive of a positive impact of failure to make AYP. Using the optimally selected bandwidth, the point estimate suggests that students in schools that miss AYP post test score gains of 2% of a standard deviation relative to students that do not make AYP. The effect is marginally significant, with a p -value of 0.057. Estimation with larger bandwidths yields larger and more significant results, while estimation with smaller bandwidths produces smaller estimates – in some cases reverse-signed – that are statistically insignificant. Introducing students who transfer out in the sample increases the point estimate slightly. Ultimately, then, we consider the math results suggestive but fragile.

5.2 Effects of specific sanctions and threats

The consequences of failing to make AYP vary dramatically depending on a school's history, and for this reason aggregate effects such as those reported above are quite likely to misrepresent the impact of exposure to particular sanctions, or the threats of those sanctions. This section summarizes the results of a number of additional nonparametric RD specifications, which by restricting the sample to students in schools with comparable AYP histories report the local average treatment effect of exposure to a certain combination of sanctions and/or threats of sanctions.

The estimates in Table 3 are based on sample restrictions to those schools not currently subject to any threats or sanctions in the NCLB regime, because they have consistently met the AYP criteria since the inception of the policy or for two consecutive years. Among these schools, the failure to make AYP exposes them to the threat of the first sanction in the regime,

which varies by year and district in North Carolina. These schools face an incentive to make AYP in the following year in order to avoid the imposition of this first sanction.

Across the board, point estimates suggest modest positive impacts of failing to meet the AYP standard for the first time. For reading test scores, the estimated effects are 2% of a standard deviation or below, and fail to reach statistical significance at standard bandwidth whether considering exposure to the tutoring or transfer threats. The point estimates are slightly larger in magnitude with p -values of 0.074 when using a large bandwidth. Including students who transfer away from the school in the sample leaves point estimates within the range already established in other specifications.

Math test scores appear to improve more substantially when exposed to the first NCLB threat. The effect is most pronounced with exposure to the transfer threat, where point estimates indicate a statistically significant 4.7% of a standard deviation improvement in test scores. This particular result is sensitive to bandwidth choice; larger and more significant at higher levels, but reverse-signed and insignificant at half the optimal bandwidth. By contrast, the impact of exposure to the tutoring threat is somewhat smaller at 2.8% of a standard deviation, but statistically significant and robust to bandwidth choice. In both cases, including students who transfer away from the school in the sample yields larger point estimates, suggesting that the failure to make AYP – even when not attached to a particular sanction – leads some families to make school choices resulting in better math instruction for their children. Intriguingly, as we will see below, this effect is not apparent in specifications when the failure to make AYP leads to the transfer sanction.

The estimates in Table 4 focus on schools that are under threat of the first sanction in the AYP regime. These schools have missed the AYP criteria in a particular subject in a recent year.

Comparison of those who make or miss AYP in the reference year reveals the local treatment effect of being exposed to the first sanction and threatened with the second, relative to being threatened with that sanction.

Normally, the first sanction applied in the NCLB system is a requirement to offer students transfers to higher-performing public schools in the same district. Students in schools required to offer transfers – and threatened with further sanctions – post modest improvements in standardized reading test scores, relative to students in schools that barely avoid this sanction. The effect is not statistically significant using the standard bandwidth selection, with a p -value of 0.149. Across numerous alternate bandwidths, the point estimate is similar in magnitude while never attaining statistical significance. In math, point estimates are of a similar order of magnitude, but attain significance at the 10% level when using bandwidths at least twice the optimal level. Evidence thus suggests some possible modest positive effects of being exposed to the transfer sanction and supplemental education services threat. As noted above, introducing transferred students into the sample does not produce a significant change in point estimates. This suggests that the mechanism of improving student performance through access to better public schools is at best canceled out by the negative impact of switching schools in the short run. Together with the results in Table 3, they suggest that savvy parents exercise school choice options in response to the initial information that their school has failed to make AYP, rather than the explicit offer of transfers.

In district/year observations where the order of sanctions was reversed, failure to make AYP a second time exposed schools to the supplemental education services sanction and the threat of offering transfers. Exposure to the tutoring sanction and transfer threat has no discernible impact on reading or math test scores, with or without transferred students in the

sample. The associated standard errors are relatively large, as by the time the sanction reversal was implemented most schools at high risk for entry into the NCLB sanction system were already beyond the level of the first sanction.

Table 5 show estimates derived from a sample of schools that are already subject to one NCLB sanction and whose exposure to a second sanction, along with the threat of the third in the sequence, depends on their performance in the reference year. The RD estimates thus reveal the local average treatment effect of the second sanction, along with the threat of the third, relative to persisting in a state where only the first sanction is in place.

In the standard NCLB sequence, requirements to offer supplemental education services are layered on top of transfers when schools miss AYP a third time. Estimates of the effect of introducing this sanction (and the threat of further action) are statistically insignificant in both math and reading, with or without transferred students included in the sample. In districts and time periods with a reversed sanction sequence, exposure to transfers as a second sanction associates with negative effects, marginally significant and robust in the case of math test scores, less robust in the case of reading. At the optimal bandwidth, point estimates suggest that introducing the transfer sanction, along with the threat of further action, leads to a 4.4% of a standard deviation reduction in math test scores. Including transferred students in the analysis actually produces more negative, and statistically significant, point estimates – indicating once again that the offer of transfers fails to yield short-term improvements among the students who take them up.

To summarize results to this point, the strongest association between failure to make AYP and subsequent test score performance occurs among those schools not yet exposed to any actual sanctions. There is some modest evidence of positive impacts associated with introducing

transfers as a first sanction, but in districts that implemented transfers as a second sanction the impact is strongly negative. Put together, the results indicate that the threat of exposure to the tutoring sanction – rather than the implementation of any sanction per se – has the strongest effect on student test scores.

Table 6 reports the results of exposure to higher-order sanctions in the NCLB regime. Beyond the first two years, it is no longer possible to distinguish the impact of exposure to one sanction from the impact of exposure to the threat of the next sanction. Nonetheless, these specifications produce intriguing findings.⁶

Upon failing to make AYP a fourth time, schools are required to take “corrective action.” Point estimates associate this sanction with test score improvements in both reading and math, but neither are statistically significant using either the default bandwidth or close alternatives.⁷

Schools that fail to make AYP a fifth time are required to formulate a restructuring plan, which they will be required to implement in the event they are unsuccessful a sixth time. Results indicate that schools exposed to this sanction exhibit declines in reading test score performance of nearly 6% of a standard deviation, a result that is both close to the standard statistical significance threshold ($p=0.053$ using the optimal bandwidth) and fairly robust across bandwidth selections. Results for math test scores are statistically insignificant across a wide range of bandwidths.

Finally, schools that fail to make AYP a sixth time are required to implement their restructuring plans. The estimated effects of restructuring on reading scores are sensitive to

⁶ We do not report the results of alternate specifications including transferred students here, as there are no noteworthy deviations from the included results.

⁷ Separate estimation for those schools that implemented sanctions in the standard order – transfers first, then tutoring – reveals larger, statistically significant impacts on both reading and math test score gains, with coefficients on the order of 7-8% of a standard deviation. In the full sample, these gains are offset by more modest and in some cases reverse-signed impacts among those schools that implemented tutoring first. This may in fact be more of a timing effect than a sanction order effect – those schools that reached the corrective action stage earlier in the process may have been better able to forecast which corrective actions would improve test scores.

bandwidth choice, with the optimal bandwidth yielding an estimated 2.6% of a standard deviation improvement ($p=0.066$), and smaller bandwidths producing larger and more significant estimates. Math results are more consistently robust, with point estimates indicating a statistically significant 5.5% of a standard deviation improvement in test scores at the optimal bandwidth, and only modest variation at alternate bandwidths.

In summary, across all the sanctions in the NCLB regime, the ultimate penalty, implementation of a restructuring plan, shows the strongest evidence of positive test score impacts across the board.⁸ No other sanction yields estimates of the same magnitude in both subjects; the threat of exposure to the sanction system comes closest.

5.3 Effect heterogeneity: by initial student test performance

One concern with school accountability systems based on proficiency rates is that they provide incentives for schools to focus only on those students near the proficiency threshold, focusing less attention on those either well above or below the cutoff (Reback 2008; Neal and Schanzenbach 2010). In theory, the treatment effect estimates presented above could represent a mix of very positive impacts for students near the threshold and very negative impacts for students far from it. Table 7 assesses this concern by reporting treatment effect estimates specific to students' initial test score quartile.⁹ An exclusive focus on moving students above the proficiency threshold would lead schools to focus on the bottom two quartiles, shifting resources away from students in higher quartiles.

⁸ As noted above, exposure to the corrective action sanction appears to have even stronger positive impacts among schools that proceed through the sanction system in the standard order.

⁹ Table 7 omits analysis of school/year observations subject to reversed order of the first two accountability sanctions. Results in this subsample are uniformly insignificant.

There is at least some evidence to indicate that schools respond to accountability sanctions by focusing more resources on lower-performing students. In schools barely missing the AYP criteria for the first time, students with below-average initial test scores post math test score increases 5% of a standard deviation greater, relative to below-average students in schools barely meeting the AYP criteria. Point estimates suggest a diminished effect among students in the second-highest quartile, and no effect whatsoever among the highest-performing students.

There is also evidence that below-average students post significant reading test score increases in schools that barely miss the AYP criteria for a second consecutive year, relative to those that miss the first time but barely make it the second. These estimates are even greater in magnitude, up to 11% of a standard deviation, and once again diminish at higher initial achievement levels, to the point where top-quartile students have comparable performance on either side of the discontinuity.

Table 6 above showed that students post significantly higher test score gains in schools required to undergo restructuring, after missing AYP for a sixth consecutive year. The final row in Table 7 shows that the beneficial effects of restructuring, estimated somewhat imprecisely here, appear to be spread fairly evenly throughout the initial test score distribution. Math test score gains, in particular, accrue at the rate of 4-6% of a standard deviation across all four quartiles. Point estimates are in fact slightly larger for students with higher initial test scores; estimates for the lower quartiles are not statistically distinguishable from zero. Point estimates for reading test score gains by quartile are positive across the board, but never significantly different from zero.

There is one set of anomalous results worth mentioning. Lower-performing students in schools required to formulate, but not implement, a restructuring plan post significantly worse

test score improvements than their counterparts in schools that avoid this sanction. This pattern replicates the more basic results in Table 6 above. It is possible that schools tasked with formulating a restructuring plan must devote resources to that activity that would otherwise be productive in the instruction of reading to low-performing students. And of course, it is also possible that this anomaly is just that – attributable to an outlier in the data. Aside from this anomaly, the remaining results in Table 7 reveal no significant patterns of note.

Overall, then, our results partly support those of existing literature showing that schools focus resources on students near the proficiency threshold when subject to accountability pressure. In contrast to this prior literature, we find no evidence that accountability leads schools to “leave behind” their highest-performing students; results for these students are near zero but never significantly negative. Moreover, we find that school restructuring leads to broad-based improvements in math test scores.

The lack of negative effects for inframarginal students might reflect the simultaneous operation of North Carolina’s own accountability system, which is based on test score increases rather than proficiency and offers cash bonuses to personnel in high-performing schools. Given tradeoffs between qualifying for AYP by increasing proficiency and qualifying for a bonus by promoting test score growth across the initial test score distribution, school personnel may have responded rationally by reallocating resources only to the point where high-performing students were not harmed on the margin. High-performing students may also have avoided relative test score declines to the extent their parents substituted for resources allocated away from them within the school system. To the extent that this occurred, the lack of test score declines among high-performers should not be considered evidence that the gains to low-performing students came at no cost.

5.4 Effect heterogeneity: among critical subgroups

For many schools in our sample, the failure to make AYP in a given year can be attributed to the performance of a single subgroup, such as Limited English Proficiency students or students receiving free or reduced price lunch. In these schools, there is a clear incentive to focus efforts on improving the performance of just that subgroup. If the improvement is sufficiently large, the group's proficiency rate does not even need to make the NCLB target, under the safe harbor exemption. Table 8 examines the impact of failure to make AYP, using regression discontinuity analysis, on a sample comprised exclusively of students belonging to the subgroup responsible for the school's failure to make AYP.

Consistent with a pattern of rational focus on the problematic subgroup, point estimates of the effects of failure to make AYP, estimated using the Imbens-Kalyanaraman optimal bandwidth, tend to be more positive than analogous estimates for the full sample. The pattern holds in 11 of 14 cases, comparing Table 8 point estimates to those in earlier tables. The overall impact of failure to make AYP on the math scores of the problematic subgroup, for example, is twice the magnitude of the effect in the overall population. Despite the larger magnitudes of these coefficients, they only rarely achieve statistical significance. The underlying sample size of students in the problematic subgroup is, of course, considerably smaller than the sample of students in schools overall, leading to larger standard errors.

The three exceptions to the larger-effect-size pattern in Table 8 include one case where the Table 8 coefficient is wrong-signed relative to its earlier counterpart. In a second case, both the original coefficient and the Table 8 coefficient are negative-signed, with the Table 8 version bearing a larger magnitude. The final case pertains to the effect of the ultimate sanction, school

restructuring, on math achievement. Here, the point estimate in the full population is larger than the effect in the problematic subgroup – the only case where point estimates have the same sign and the point estimate in the problematic subgroup is closer to zero. While it may be inappropriate to assign much interpretation to this pattern, given the imprecision of the underlying estimates, this is consistent with the notion that mandated changes in school leadership among schools that have persistently failed to make AYP yields benefits to all students, and not only those who would be targeted by a rational administrator.

5.5 Effect heterogeneity: SES effects among eligible students

Supplemental education services were required to be offered to students eligible for free or reduced price lunch in schools that had missed AYP in the same subject for three consecutive years (two years following North Carolina’s reversal of sanction order). Tables 4 and 5 above failed to report any evidence that SES led to improved test score outcomes; in fact the preponderance of point estimates are negative. Table 8 re-examines the effects of SES by restricting attention to those students actually eligible to receive the services. Even in this subsample, there is no evidence of a significant impact of SES on student performance.

5.6 Falsification Tests

As noted above, our procedure may yield spurious positive effects of missing the AYP cutoff if the tendency to mean revert towards higher performance is, for any reason, significantly greater on the left-hand side of the discontinuity point. To assess this possibility, we estimated two sets of pseudo-regression discontinuity analyses. In the first set, we estimated our models using a cut point 15 percentage points above the actual proficiency cutoff in each school. This

point lies towards the upper tail of the proficiency distribution as shown in Figure 1, implying that a pattern of discontinuous mean reversion would produce spurious negative results. The second set uses a cut point 15 percentage points below the actual cutoff, closer to the lower tail of the distribution. Discontinuous mean reversion would produce spurious positive results in this sample.

Table 10 presents our basic optimal-bandwidth results from preceding tables, along with results using an identical procedure at the pseudo-discontinuity points. These point estimates will tend to be large in absolute magnitude, because they are derived from a Wald estimator with a relatively small denominator. That is to say, there is not much evidence of a discontinuity in treatment at either of these points, so any effect observed in the outcome variable will be increased substantially. In analyzing these results, then, we are more interested in patterns of sign and significance than the actual value of the point estimates *per se*.

Significance patterns reveal few signs of concern. Across forty falsification tests, we obtain 5 point estimates significant at the 10% level, 3 at the 5% level, and one at the 1% level. In each case, these exceed the expected number of false positives under the null hypothesis of no effects, but by no more than one. Moreover, the signs of the significant coefficients do not always concord with the hypothesized effects of mean reversion. In the first set of tests, where we would be concerned about spurious negative mean reversion, only two of three significant coefficients are in fact negative. In the second test, where our concern is positive mean reversion, both significant coefficients are in fact negative.

Looking more broadly, the first set of falsification tests yield 12 negative point estimate in 20 tests, a plausible outcome under the null hypothesis of no effect in any test. The second set provides an even split across positive and negative point estimates. In summary, using two sets

of regression discontinuity analyses where we might suspect mean reversion to be an even greater concern, we find no evidence to validate these concerns.

6. Conclusions

Previous research has raised a number of concerns about school accountability regimes. Attaching high stakes to test scores, and in particular to proficiency rates on tests, may lead to a number of resource allocation decisions that have the short-term impact of avoiding sanctions but no impact, or in some cases even a negative impact, on long-term measures of student productivity or well-being. While this study is incapable of providing a holistic estimate of the impact of accountability relative to a counterfactual of no accountability (see Hanushek and Raymond 2005 or Dee and Jacob 2011 for such exercises), we assess whether the sanctions associated with the nation's most comprehensive accountability system impact a fundamental measure of student learning – year-over-year improvement in test scores.

We find evidence that schools respond to the threat of initial sanctions in ways that contribute to student learning, and that the ultimate sanction in the NCLB system – restructuring – yields even greater contributions to learning. Consistent with prior research, we find some evidence that these gains are concentrated among students at or below the proficiency threshold. Unlike prior research, we find that these gains are not accompanied by significant losses among students well above the proficiency threshold – which may illustrate the importance of a secondary accountability system measuring growth rather than accountability. We also find evidence that schools focus in particular on improving the performance of the subgroup that caused the school to miss AYP.

The strong positive effects of restructuring – which appear to be broad, rather than focused on the lowest-performing students – indicate that school management or leadership problems constitute the single greatest obstacle to improved student performance. Indeed, the existence of management or leadership problems might help explain why lesser sanctions – including a targeted tutoring intervention associated with no positive effects even among the targeted students – have no demonstrable effect. School leaders who cannot formulate strategies to improve performance cannot be expected to react constructively to incentives to do so. The association of positive effects with entry into the NCLB sanction system might also be interpreted as evidence of effective leadership reactions – among those schools at the threshold of AYP, the ones with effective leadership are the ones who find ways to increase proficiency rates and exit the sanction system, leaving only schools with less-effective leaders behind.

Even with this admittedly speculative interpretation of the results, one should not jump to the conclusion that a No Child Left Behind-style sanction regime is an effective way to identify schools in need of leadership change and implement that change. Presumably, more targeted strategies could achieve the same results without the need to wait for six years' worth of standardized test results. These results do highlight the need for further research into the nature of effective school leadership and management.

References

- Angrist, J., E. Bettinger, and M. Kremer (2006) "Long-Term Consequences of Secondary School Vouchers: Evidence from Administrative Records in Colombia." *American Economic Review* 96(3): 847-862.
- Bettinger, E.P. (2005) "The Effect of Charter Schools on Charter Students and Public Schools." *Economics of Education Review* 24(2): 133-147.
- Bifulco, R. and H.F. Ladd (2006) "The Impacts of Charter Schools on Student Achievement: Evidence from North Carolina." *Education Finance and Policy* 1(1):50-90.

- Chakrabarti, R. (2008a) “Can Increasing Private School Participation and Monetary Loss in a Voucher Program Affect Public School Performance? Evidence from Milwaukee.” *Journal of Public Economics* 92 (5-6) 1371-1393.
- Chakrabarti, R. (2008b) “Impact of Voucher Design on Public School Performance: Evidence from Florida and Milwaukee Voucher Programs.” FRB of New York Staff Report #315.
- Chetty, R., J.N. Friedman, and J.E. Rockoff (2011) “The Long-Term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood.” NBER Working Paper #17699.
- Clotfelter, C.T., H.F. Ladd, J.L. Vigdor, and R.A. Aliaga (2004) “Do School Accountability Systems Make It More Difficult for Low-Performing Schools to Attract and Retain High-Quality Teachers?” *Journal of Policy Analysis and Management* 23(2): 251-271.
- Dee, T.S. 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, D.N. and M.E. Lucas (2004) “What’s In A Grade? School Report Cards and the Housing Market.” *American Economic Review* 94(3): 591-604.
- Figlio, D.N. and J. Winicki (2005) “Food for Thought: The Effects of School Accountability Plans on School Nutrition.” *Journal of Public Economics* 89(2-3): 381-394.
- Figlio, D.N. (2006) “Testing, Crime, and Punishment.” *Journal of Public Economics* 90(4-5): 837-851.
- Friedman, M. (1962) *Capitalism and Freedom*. Chicago: The University of Chicago Press.
- Hahn, J., P. Todd, and W. van der Klaauw (2001) “Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design.” *Econometrica* 69(1): 201-209.
- Hanushek, E.A., J.F. Kain, and S.G. Rivkin (2004) “Disruption versus Tiebout Improvement: The Costs and Benefits of Switching Schools.” *Journal of Public Economics* 88: 1721-1746.
- Hanushek, E.A. and M.E. Raymond (2005) “Does School Accountability Lead to Improved Student Performance?” *Journal of Policy Analysis and Management* 24(2): 297-327.
- Hastings, J.S. and J.M. Weinstein (2008) “Information, School Choice, and Academic Achievement: Evidence from Two Experiments.” *Quarterly Journal of Economics* 123(4): 1373-1414.
- Hoxby, C.M. (2000) “Does Competition Among Public Schools Benefit Students and Taxpayers?” *American Economic Review* 90(5): 1209-38.
- Imbens, G.W. and K. Kalyanaraman (2009) “Optimal Bandwidth Choice for the Regression Discontinuity Estimator.” NBER Working Paper #14726.
- Imbens, G.W. and T. Lemieux (2008) “Regression Discontinuity Designs: A Guide to Practice.” *Journal of Econometrics* 142(2): 615-635.

- Jacob, B.A. and S.D. Levitt (2003) "Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating." *Quarterly Journal of Economics* 118(3): 843-877.
- Ladd, H.F. (2002) "School Vouchers: A Critical View." *Journal of Economic Perspectives* 16(4): 3-24.
- Ladd, H.F. (1999) "The Dallas School Accountability and Incentive Program: An Evaluation of its Impacts on Student Outcomes." *Economics of Education Review* 18(1): 1-16.
- Lee, D.S. and T. Lemieux (2010) "Regression Discontinuity Designs in Economics." *Journal of Economic Literature* 48(2): 281-355.
- Kane, T.J. and D.O. Staiger (2002) "The Promise and Pitfalls of Using Imprecise School Accountability Measures." *Journal of Economic Perspectives* 16(4): 91-114.
- Martinez, E. (2011) "Do Housing Prices Account for School Accountability?" University of South Florida working paper.
- McCrary, J. (2008) "Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test." *Journal of Econometrics* 142(2): 698-714.
- Neal, D. and D. Schanzenbach (2010) "Left Behind By Design: Proficiency Counts and Test-Based Accountability." *Review of Economics and Statistics* 92(2): 263-283.
- Reback, R. (2008) "Teaching to the Rating: School Accountability and the Distribution of Student Achievement." *Journal of Public Economics* 92(5-6): 1394-1415.
- Rothstein, J. (2007) "Does Competition Among Public Schools Benefit Students and Taxpayers? Comment." *American Economic Review* 97(5): 2026-2037.
- Rouse, C.E. (1998) "Private School Vouchers and Student Achievement: An Evaluation of the Milwaukee Parental Choice Program." *Quarterly Journal of Economics* 113(2): 553-602.
- Sass, T.R. (2006) "Charter Schools and Student Achievement in Florida." *Education Finance and Policy* 1(1): 91-122.
- West, M.R. and P.E. Peterson (2006) "The Efficacy of Choice Threats Within School Accountability Systems: Results from Legislatively Induced Experiments." *Economic Journal* 116(510): C46-C62.

Table 1: The NCLB sanction regime

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

Table 2: Regression Discontinuity Estimates of the Effect of Missing AYP

	Reading (n=8,264)	Math (n=8,266)
At optimal bandwidth	-0.00002 (0.0056)	0.0180* (0.0095)
Half optimal bandwidth	0.0061 (0.0071)	-0.0117 (0.0137)
Twice optimal bandwidth	-0.0025 (0.0053)	0.0282 *** (0.0081)
At optimal bandwidth, including transferred students	-0.0004 (0.431)	0.022** (0.011)

Note: Standard errors in parentheses. Unit of observation is school/year, weighted by number of students used to compute dependent variable. Dependent variable is mean year-over-year change in standardized test score in the indicated subject between the year of AYP status determination and the following year. Bandwidth determination is by the Imbens-Kalyanaraman algorithm. Specifications factor out school-level aggregate characteristics including percent minority, percent limited English proficient, and percent free/reduced lunch.

*** denotes an estimate significant at the 1% level; ** the 5% level; * the 10% level.

Table 3: RD estimates of the impact of exposure to the threat of the first NCLB sanction

First sanction is transfers	Reading (n=3,241)	Math (n=3,244)
At optimal bandwidth	0.0203 (0.0140)	0.0473** (0.0204)
Half optimal bandwidth	0.0086 (0.0218)	-0.0197 (0.0322)
Twice optimal bandwidth	0.0200* (0.0111)	0.0681*** (0.0170)
At optimal bandwidth, including transferred students	0.0160 (0.0118)	0.0759*** (0.017)
First sanction is SES	Reading (n=958)	Math (n=957)
At optimal bandwidth	0.0161 (0.0136)	0.0285* (0.0163)
Half optimal bandwidth	0.0119 (0.0170)	0.0350* (0.0204)
Twice optimal bandwidth	0.0227* (0.0127)	0.0241* (0.0147)
At optimal bandwidth, including transferred students	0.0221* (0.0131)	0.0338** (0.0153)

Note: Standard errors in parentheses. Unit of observation is school/year, weighted by number of students used to compute dependent variable. Dependent variable is mean year-over-year change in standardized test score in the indicated subject between the year of AYP status determination and the following year. Bandwidth determination is by the Imbens-Kalyanaraman algorithm. Specifications factor out school-level aggregate characteristics including percent minority, percent limited English proficient, and percent free/reduced lunch.

*** denotes an estimate significant at the 1% level; ** the 5% level; * the 10% level.

Table 4: RD estimates of the impact of exposure to the first NCLB sanction/threat of second

First sanction is transfers (second SES)	Reading (n=1,486)	Math (n=1,486)
At optimal bandwidth	0.0178 (0.0124)	0.0144 (0.0181)
Half optimal bandwidth	0.0139 (0.0176)	0.0076 (0.0254)
Twice optimal bandwidth	0.0185 (0.0106)	0.0256 (0.015)
At optimal bandwidth, including transferred students	0.0194 (0.0105)	0.0092 (0.0200)
First sanction is SES (second transfers)	Reading (n=177)	Math (n=177)
At optimal bandwidth	-0.0064 (0.0377)	-0.0069 (0.0337)
Half optimal bandwidth	-0.0291 (0.0445)	-0.0184 (0.0423)
Twice optimal bandwidth	-0.0126 (0.0353)	0.0069 (0.0306)
At optimal bandwidth, including transferred students	0.0055 (0.0381)	-0.0056 (0.0314)

Note: Standard errors in parentheses. Unit of observation is school/year, weighted by number of students used to compute dependent variable. Dependent variable is mean year-over-year change in standardized test score in the indicated subject between the year of AYP status determination and the following year. Bandwidth determination is by the Imbens-Kalyanaraman algorithm. Specifications factor out school-level aggregate characteristics including percent minority, percent limited English proficient, and percent free/reduced lunch.

*** denotes an estimate significant at the 1% level; ** the 5% level; * the 10% level.

Table 5: RD estimates of the impact of exposure to the second NCLB sanction/threat of third

Second sanction is SES (transfers in place)	Reading (n=727)	Math (n=727)
At optimal bandwidth	-0.0244 (0.0268)	-0.0219 (0.0368)
Half optimal bandwidth	-0.0519 (0.0398)	-0.064 (0.0546)
Twice optimal bandwidth	-0.0234 (0.0217)	-0.0409 (0.0394)
At optimal bandwidth, including transferred students	-0.0191 (0.0259)	-0.0384 (0.0473)
Second sanction is transfers (SES in place)	Reading (n=310)	Math (n=310)
At optimal bandwidth	-0.0289 (0.0244)	-0.0442 (0.0243)
Half optimal bandwidth	0.005 (0.0327)	-0.0495 (0.029)
Twice optimal bandwidth	-0.0429** (0.0208)	-0.0394 (0.0234)
At optimal bandwidth, including transferred students	-0.0495** (0.0216)	-0.0528* (0.0261)

Note: Standard errors in parentheses. Unit of observation is school/year, weighted by number of students used to compute dependent variable. Dependent variable is mean year-over-year change in standardized test score in the indicated subject between the year of AYP status determination and the following year. Bandwidth determination is by the Imbens-Kalyanaraman algorithm. Specifications factor out school-level aggregate characteristics including percent minority, percent limited English proficient, and percent free/reduced lunch.

*** denotes an estimate significant at the 1% level; ** the 5% level; * the 10% level.

Table 6: RD estimates of the impact of higher order sanctions and threats

Corrective action, threat of restructuring plan	Reading (n=549)	Math (n=549)
At optimal bandwidth	0.035 (0.0261)	0.0231 (0.0247)
Half optimal bandwidth	0.0337 (0.0392)	-0.0065 (0.0319)
Twice optimal bandwidth	0.0148 (0.0213)	0.0311 (0.0215)
Restructuring plan, threat of restructuring	Reading (n=442)	Math (n=442)
At optimal bandwidth	-0.0558** (0.0288)	0.0278 (0.0258)
Half optimal bandwidth	-0.052 (0.0466)	0.0486 (0.0339)
Twice optimal bandwidth	-0.0937* (0.0516)	0.0197 (0.0238)
Restructuring	Reading (n=367)	Math (n=367)
At optimal bandwidth	0.0264* (0.0144)	0.0547** (0.0263)
Half optimal bandwidth	0.0562*** (0.0202)	0.0664* (0.035)
Twice optimal bandwidth	0.0205 (0.0132)	0.0609** (0.0242)

Note: Standard errors in parentheses. Unit of observation is school/year, weighted by number of students used to compute dependent variable. Dependent variable is mean year-over-year change in standardized test score in the indicated subject between the year of AYP status determination and the following year. Bandwidth determination is by the Imbens-Kalyanaraman algorithm. Specifications factor out school-level aggregate characteristics including percent minority, percent limited English proficient, and percent free/reduced lunch.

*** denotes an estimate significant at the 1% level; ** the 5% level; * the 10% level.

Table 7: RD estimates of the effect of NCLB sanctions and threats on test score gains, by quartile of initial test score

Sanction/ Threat	Reading				Math			
	Lowest	2 nd	3 rd	Highest	Lowest	2 nd	3 rd	Highest
General Impact of Missing AYP	-0.0023 (0.0130)	-0.0137 (0.0138)	-0.0026 (0.0103)	0.0019 (0.0066)	0.0363*** (0.0116)	0.0146 (0.0103)	0.0194* (0.0103)	0.0015 (0.0114)
None/ Transfers	-0.0054 (0.0388)	-0.0251 (0.0375)	-0.0036 (0.0247)	0.0014 (0.0173)	0.0548** (0.0272)	0.0506** (0.0236)	0.0308 (0.0244)	0.0039 (0.0261)
Transfers/ SES	0.1045** (0.0448)	0.1127*** (0.0421)	0.0506* (0.0277)	0.0195 (0.0200)	0.0404 (0.0322)	0.0237 (0.0299)	-0.0077 (0.0308)	0.0028 (0.0239)
SES/ Corrective Action	-0.0897 (0.0707)	-0.0808 (0.0603)	0.0050 (0.0348)	-0.0065 (0.0223)	0.0007 (0.0382)	0.0160 (0.0413)	0.0282 (0.0363)	0.0090 (0.0391)
Corrective Action/ Restructuring Plan	0.0378 (0.0373)	0.0427 (0.0444)	0.0305 (0.0256)	0.0222 (0.0182)	0.0360 (0.0259)	0.0042 (0.0294)	0.0531* (0.0311)	-0.0021 (0.0299)
Restructuring Plan/ Restructuring	-0.0788** (0.0379)	-0.0936** (0.0387)	-0.0373 (0.0290)	-0.0349 (0.0231)	0.0424 (0.0346)	-0.0000 (0.0300)	0.0078 (0.0341)	0.0308 (0.0380)
Restructuring/ None	0.0346 (0.0217)	0.0232 (0.0194)	0.0087 (0.0179)	0.0172 (0.0206)	0.0418 (0.0299)	0.0424 (0.0298)	0.0598** (0.0298)	0.0605* (0.0313)

Note: Standard errors in parentheses. Unit of observation is school/year, weighted by number of students used to compute dependent variable. Dependent variable is mean year-over-year change in standardized test score in the indicated subject between the year of AYP status determination and the following year. All estimates utilize the optimal bandwidth as determined by the Imbens-Kalyanaraman algorithm. Specifications factor out school-level aggregate characteristics including percent minority, percent limited English proficient, and percent free/reduced lunch.

*** denotes an estimate significant at the 1% level; ** the 5% level; * the 10% level.

Table 8: RD estimates of the effect of NCLB sanctions on students in the critical subgroup

Sanction/ Threat	Reading	Math
General Impact of Missing AYP	0.009 (0.025)	0.038** (0.015)
None/ Transfers	0.043** (0.022)	0.053 (0.036)
Transfers/ SES	0.029 (0.049)	-0.043 (0.045)
SES/ Corrective Action	0.025 (0.032)	0.040 (0.037)
Corrective Action/ Restructuring Plan	0.068 (0.062)	0.044 (0.046)
Restructuring Plan/ Restructuring	-0.203 (0.138)	0.095** (0.038)
Restructuring/ None	0.044 (0.075)	0.043 (0.042)

Note: Standard errors in parentheses. Unit of observation is school/year, weighted by number of students used to compute dependent variable. Dependent variable is mean year-over-year change in standardized test score in the indicated subject between the year of AYP status determination and the following year. All estimates utilize the optimal bandwidth as determined by the Imbens-Kalyanaraman algorithm. Specifications factor out school-level aggregate characteristics including percent minority, percent limited English proficient, and percent free/reduced lunch.

*** denotes an estimate significant at the 1% level; ** the 5% level; * the 10% level.

Table 9: RD estimates of the impact of SES, Free/Reduced Lunch Eligible Students Only

	Reading	Math
SES implemented as second sanction	-0.0393 (0.0501)	0.0437 (0.0382)
SES implemented as first sanction	-0.0040 (0.0454)	0.0083 (0.0357)

Note: Standard errors in parentheses. Unit of observation is school/year, weighted by number of students used to compute dependent variable. Dependent variable is mean year-over-year change in standardized test score in the indicated subject between the year of AYP status determination and the following year. All estimates utilize the optimal bandwidth as determined by the Imbens-Kalyanaraman algorithm. Specifications factor out school-level aggregate characteristics including percent minority, percent limited English proficient, and percent free/reduced lunch.

*** denotes an estimate significant at the 1% level; ** the 5% level; * the 10% level.

Table 10: Summary of Falsification Tests

Sanction/ Threat	Actual result		Pseudo-RD at 15% above threshold		Pseudo-RD at 15% below threshold	
	Math	Reading	Math	Reading	Math	Reading
Overall	0.018*	0.0002	-1.09***	0.061	0.0097	0.130
None/ Transfers	0.047**	0.021	-13.26	-0.864	-0.040	-0.036
None/ SES	0.029*	0.016	-0.042	-0.844	0.083	0.032
Transfers/ SES	0.014	0.018	0.004	-0.049	0.068	0.258
SES/ Transfers	-0.007	-0.006	0.328	0.333	-0.586	-0.628**
SES/ Corrective Action	-0.022	-0.024	-0.090	0.007	0.125	-0.111
Transfers/ Corrective Action	-0.044	-0.029	0.132*	0.046	-0.077	0.121
Corrective Action/ Restructuring Plan	0.023	0.035	-0.038	-0.038	-0.132	-0.114
Restructuring Plan/ Restructuring	0.028	-0.056**	1.86	-0.065	-.423*	-0.392
Restructuring/ None	0.055**	0.026*	-0.125**	-0.048	0.034	0.293

Note: Unit of observation is school/year, weighted by number of students used to compute dependent variable. Dependent variable is mean year-over-year change in standardized test score in the indicated subject between the year of AYP status determination and the following year. All estimates utilize the optimal bandwidth as determined by the Imbens-Kalyanaraman algorithm. Specifications factor out school-level aggregate characteristics including percent minority, percent limited English proficient, and percent free/reduced lunch.

*** denotes an estimate significant at the 1% level; ** the 5% level; * the 10% level

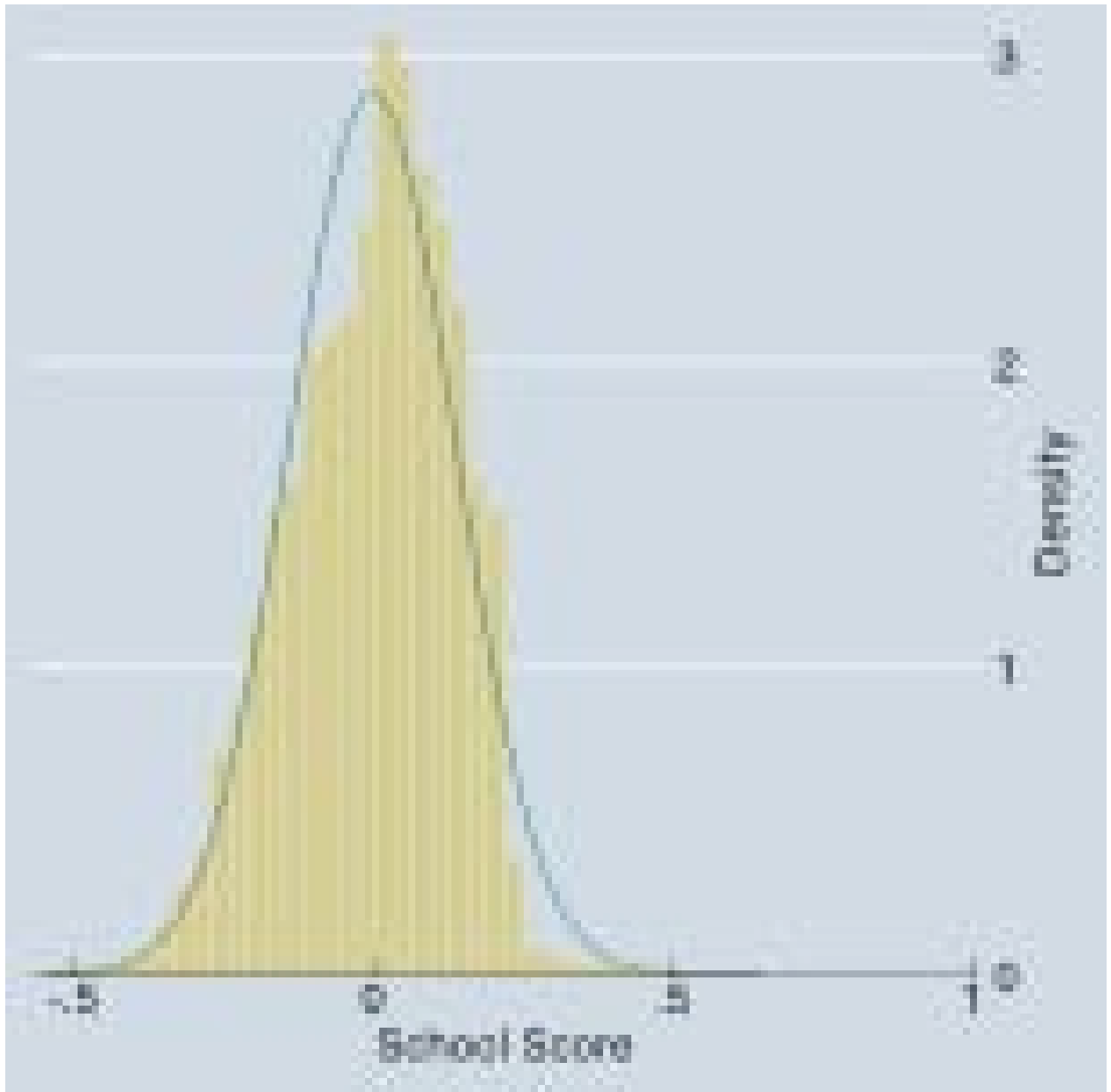


Figure 1: Density plot by assignment variable

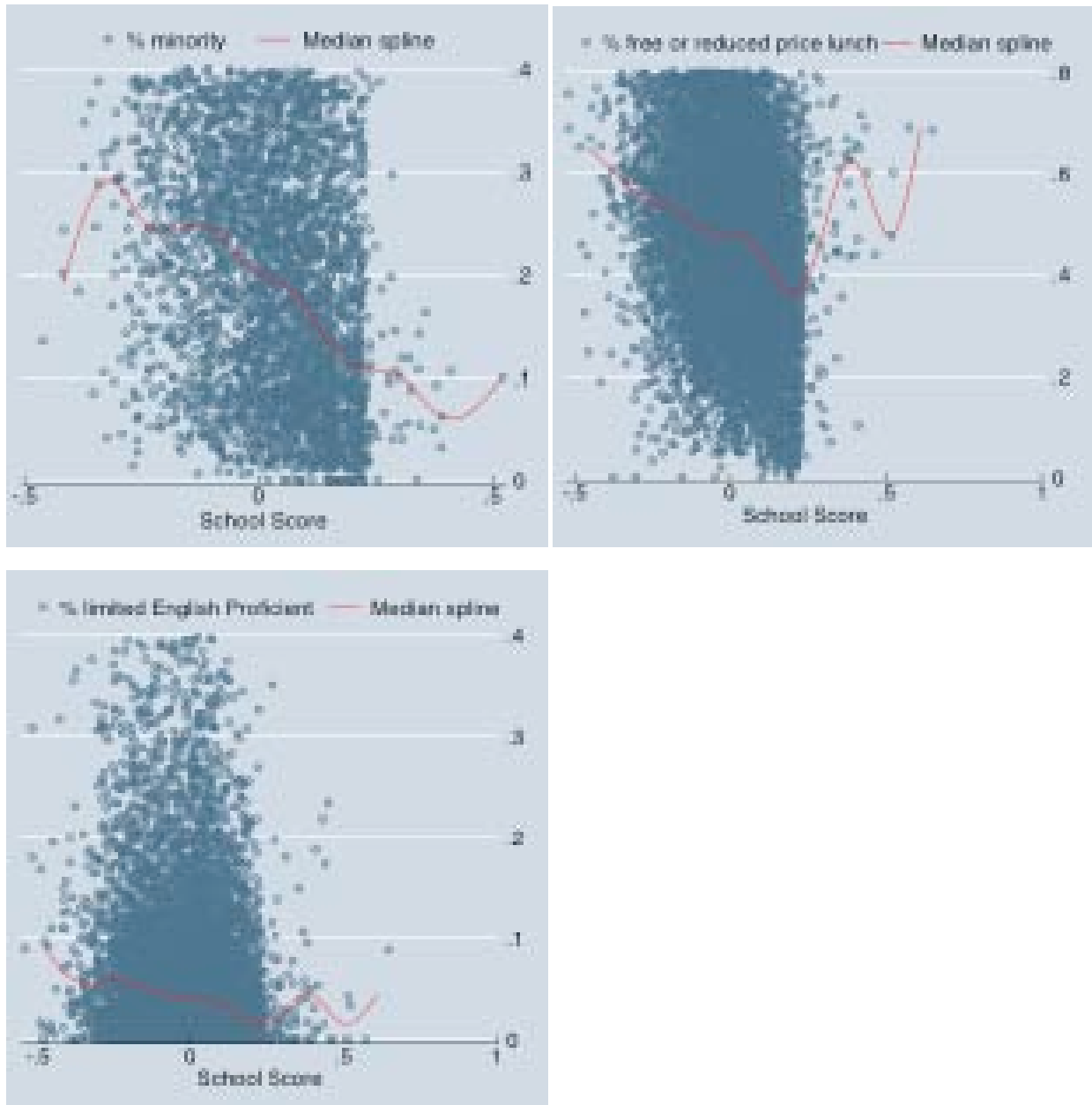
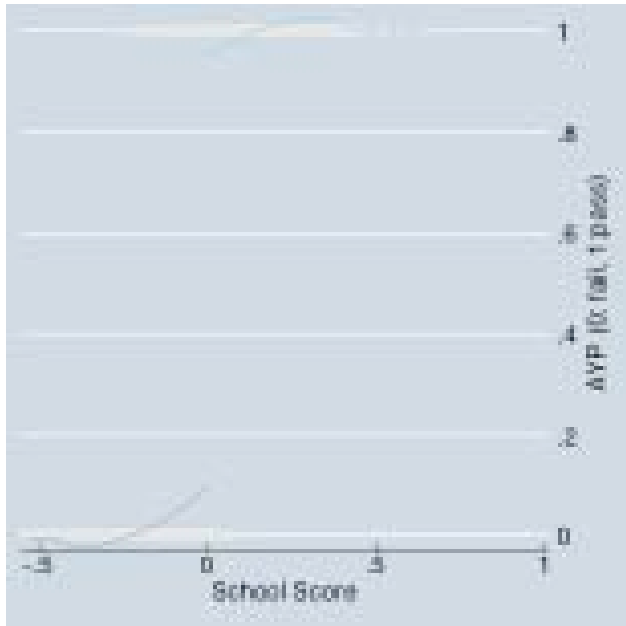
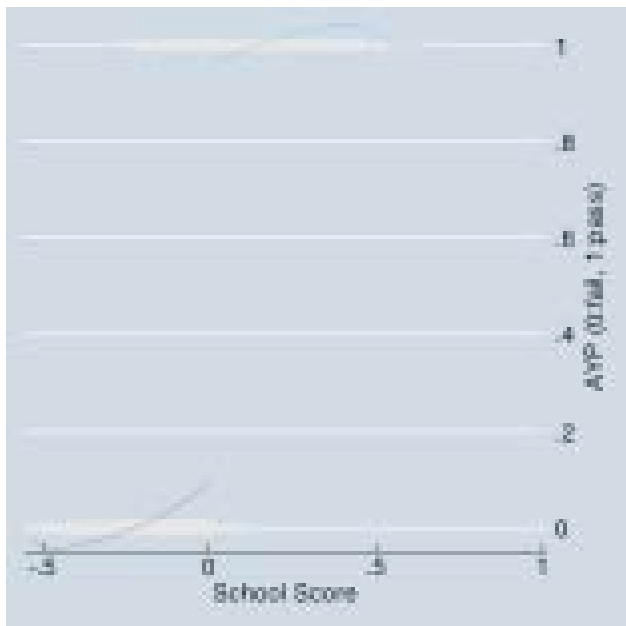


Figure 2: Covariates by assignment variable



A. Reading



B. Math

Figure 3: Probability of meeting the AYP criteria by assignment variable