NBER WORKING PAPER SERIES

SCHOOL PERFORMANCE, ACCOUNTABILITY AND WAIVER REFORMS: EVIDENCE FROM LOUISIANA

Thomas Dee
Elise Dizon-Ross

Working Paper 23463
http://www.nber.org/papers/w23463

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
June 2017

We would like to acknowledge financial support from the Spencer, Walton, and WT Grant Foundations. We also express appreciation for comments provided by seminar participants at Stanford University and by participants at the AEFP and APPAM research conferences. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

School Performance, Accountability and Waiver Reforms: Evidence from Louisiana
Thomas Dee and Elise Dizon-Ross
NBER Working Paper No. 23463
June 2017
JEL No. H70,I2

ABSTRACT

States that received federal waivers to the No Child Left Behind (NCLB) Act were required to implement reforms in designated "Focus Schools" that contribute to achievement gaps. In this study, we examine the performance effects of such "differentiated accountability" reforms in the state of Louisiana. The Focus School reforms in Louisiana emphasized school-needs assessments and aligned technical assistance. These state reforms may have also been uniquely high-powered because they were linked to a new letter-based school-rating system. We examine the impact of these reforms in a sharp regression discontinuity (RD) design based on the assignment of schools to Focus status. We find that, over each of three years, Louisiana's Focus School reforms had no measurable impact on school performance. We discuss evidence that these findings may reflect policy uncertainty and implementation fidelity at the state and local level.

Thomas Dee
Stanford University
520 Galvez Mall, CERAS Building, 5th Floor
Stanford, CA 94305-3084
and NBER
tdee@stanford.edu

Elise Dizon-Ross
Stanford University
520 Galvez Mall, CERAS Building, 5th Floor
Stanford, CA 94305-3084
elised@stanford.edu

1. INTRODUCTION

Educational achievement gaps in the United States are large and continue to persist despite national attention and multiple policy efforts to close them. According to the National Assessment of Educational Progress (NAEP), the achievement gap between black and white students on the 8th grade NAEP for math corresponds to roughly three years of schooling. A U.S. Department of Education study indicates that just over half of this gap was attributable to *within-school* sources, while approximately 16 percent was attributable to between-school gaps and the remaining 32 percent was indeterminate (Bohrnstedt et al, 2015). Such achievement gaps persist not just between races, but increasingly between high- and low-income students as well (Reardon, 2011).

For over a decade, one of the key elements of federal education policy has been a focus on trying to reduce these achievement gaps. Under No Child Left Behind (NCLB), the federal government mandated accountability reforms that, for the first time, would identify the achievement of individual subgroups of students and sanction those schools that failed to make progress improving the outcomes of their lowest-performing students. The law was scheduled for revision in 2007, but with Congress unable to collectively reauthorize a new version of NCLB, the Department of Education in 2011 introduced flexibility waivers for which states could apply in order to avoid being held to the strictest requirements of the law. In exchange, states had to implement a set of reforms, one component of which was a system of "differentiated" accountability that would identify low-performing schools contributing to achievement gaps, known as Focus Schools, and target them for intervention. The majority of states applied and received a waiver.

In this paper, we examine Focus School reforms conducted under the NCLB waiver in Louisiana. Louisiana's implementation of this signature federal achievement-gap reform is uniquely interesting because it coincided with a state consequential-accountability mechanism that has been shown to be effective elsewhere (i.e., school letter grades). The treatment contrast that we study in Louisiana is a simultaneous combination of both Focus School and letter grade reforms, which is arguably a stronger treatment than the waivers themselves required and distinct from Focus School systems in other waiver states.

We use a regression discontinuity for our analysis, leveraging the sharp discontinuity in Louisiana's assignment of schools to Focus School status based on a baseline performance

measure. This identification strategy allows us to make a causal estimate of the effect on school outcomes of being identified as a Focus School and receiving the corresponding interventions, in combination with the impact of receiving a low letter grade. We find no evidence that being assigned to the Focus School "treatment" led to improvements in student test scores or schools' performance rating relative to other low-performing schools. In fact, three years after the start of the intervention for the first cohort of Focus Schools, these schools appear to be doing somewhat worse relative to other low-performing schools, though these effects are largely not statistically significant.

Our findings are particularly interesting given the fact that the recently passed reauthorization of ESEA, the Every Student Succeeds Act, provides guidelines on school accountability systems that closely mirror those in the NCLB flexibility waivers. Rather than being simply a short-term, stopgap measure, the waiver era provides a preview of the types of differentiated accountability systems that may develop under the newest iteration of the federal government's education policy. Ex ante, we would have predicted that Louisiana's Focus School treatment, if anything, would have led to stronger positive responses from low-performing schools, compared to other states, due to its inclusion of letter grades. Based on descriptions of the interventions the state designed and accounts of the system's implementation, we hypothesize that our findings are in part driven by differences in the federal vision for Focus School policy and the state's own vision and goals for their accountability system. Our findings serve as a cautionary reminder that the success of top-down accountability reforms can vary widely across states and that the ability of such reforms to lead to the intended results depends on the alignment of federal goals with the goals and implementation at the ground level.

The paper is organized as follows: Section 2 will describe the policy background context and literature relevant to our research questions. Section 3 will discuss details of the accountability system and waiver reforms in Louisiana. Section 4 discusses our data and identification strategy. Section 5 presents our results and Section 6 describes robustness checks. We conclude with a discussion of our results and how they relate to findings from other states.

2. PRIOR LITERATURE AND THEORETICAL CONSIDERATIONS

The reforms we study in Louisiana are situated in both the literature on school accountability and the literature on whole school reform. The reforms as they were implemented

in Louisiana combined a system of public accountability (i.e., explicit school letter grades and public identification as a Focus School) with supports in the form of technical assistance intended to be personalized to the unique school needs. Unlike many other waiver states, the Focus School reforms in Louisiana also had a whole-school character. That is, the state classified schools based on an overall low level of performance rather than because of the low performance of any particular subgroup within schools. As a practical matter, this approach could still reduce overall achievement gaps by race and income because of the high levels of segregation across Louisiana's public schools. For example, the mean school-level percentage of black students in the first Focus School cohort was 87% and the median was 93%; the mean percentage of FRPL students was 93% and the median was 95%.

There are several broad theoretical motivations for school accountability policies. For example, it may be that well-intentioned district and school staff lack full information on the true character of their school's performance relative to public expectations and that accountability policies convey this information emphatically. Alternatively, the performance of school and district staff in chronically underperforming schools may suffer from coordination problems that are attenuated by accountability incentives. Another possibility is that school and district staff may not undertake the desired behaviors or pursue key goals because their objectives differ from those of the public (i.e., a type of moral hazard problem). In this scenario, the incentives created by accountability reforms may also improve school effectiveness. A variety of studies have empirically examined the effectiveness of accountability reforms. In 2011, the National Research Council released a comprehensive research survey on the impact of incentives and test-based accountability in which the panel concluded that test-based incentive programs tend to result in positive, though not necessarily transformational, changes in achievement, especially in 4th grade math (NRC 2011). In a separate survey, Figlio and Loeb (2011) similarly cite the evidence from studies of NCLB (Dee and Jacob, 2011; Wong, Steiner, and Cook, 2015) and studies of pre-NCLB accountability policies (Hanushek and Raymond, 2005) indicating that accountability policies that articulate consequences for low-performing schools improve their performance. Interestingly, pre-NCLB reforms that publicized information on school performance but did not articulate any labels or sanctions (i.e., "report card" accountability) appeared to be ineffective (Hanushek and Raymond, 2005).

This element of public sanction is particularly relevant in the case of Louisiana, where a critical component of the school accountability system—which we discuss in more detail in the following section—is the merging of the Focus School label with a well-publicized and intuitive "F" letter grade. Although we are not aware of any school accountability system that completely mirrors Louisiana's, previous research on Florida's A+ Plan accountability system is relevant as context for our own findings. Starting in 1999, the A+ Plan called for Florida to issue school letter grades based on student achievement on annual curriculum-based tests administered to students grades 3-10. Schools with high grades earned rewards while low-performing schools (those with an F) received both assistance and sanctions. Research on Florida's A+ Plan has generally shown positive impacts of its accountability system, from early studies done not long after the plan was implemented to more recent re-examinations of its effects (Greene, 2001; Figlio and Rouse, 2006; West and Peterson, 2006; Chiang, 2009; Rouse et al, 2013). Additionally, New York City's school accountability system, although not statewide, is similar to Louisiana's in that the district's Department of Education issues annual letter grades to every school. Rockoff and Turner (2008) found that receiving a low letter grade led to significant improvements in both math and reading test scores, as soon as one year following the policy's implementation.

The potential impact of a system of public accountability is, of course, the result not just of external pressure, but also of the support and interventions that are made available to districts and schools. Under NCLB, schools failing to meet AYP were required to be treated with one or more of a set of increasingly strict and prescriptive interventions outlined by the legislation. However, NCLB waivers provided states with the flexibility to bring any relevant evidence-based reforms to its identified Focus Schools. Whether and when this more flexible approach to accountability is successful is an important and open empirical question.

Louisiana's Focus School reforms identified whole schools rather than targeted subgroups for intervention. Therefore, this reform effort should also be situated in the broad body of literature exists on comprehensive school reform (CSR). This prior approach to school improvement, in contrast to targeted interventions such as pull-out programs or other piecemeal Title I-funded programs, gained momentum in the late 1990s when Congress legislated millions of dollars to support evidence-based, comprehensive school reform. The U.S. Department of Education identified 11 necessary elements to qualify as CSR. Unsurprisingly, a large number of

CSR models were developed, ranging from home-grown models to those developed by universities or research centers and packaged for national distribution. Some of the most well known models, such as Direct Instruction and Success for All, have been the subjects of dozens of studies (e.g., Brent and DiObilda, 1993; O'Brien and Ware, 2002; Madden et al, 1993; Slavin and Madden, 2000). A 2003 meta-analysis of the CSR research done by Borman et al. (2003) found that the research base was limiting but that overall the effects of CSR appeared to be promising, and found that the Direct Instruction and Success for all, as well as the School Development Program, had the strongest evidence of effectiveness with respect to student achievement. The authors also concluded that the heterogeneity in the effectiveness of CSR models was likely due to considerable challenges of conducting a consistently high-fidelity implementation of reforms across several schools.

The case of Louisiana's Focus School reforms sits at the intersection of the policy research on school accountability and comprehensive school reform, both of which have been thoroughly examined and been shown to have the potential for impact. However, the reforms we study here also represent a new type of approach to achieving school improvement. The accountability system developed in Louisiana is characterized by a built-in flexibility, coupled with high stakes accountability. Rather than relying on a prescribed set of interventions, or even a packaged model of whole-school reform, the state identifies the explicit steps for catalyzing school improvement. Whether this new type of reform model is effective in leading to improved achievement for low-performing schools is an empirical question that we turn to now in this paper.

Our research into this question is particularly relevant given the recent reauthorization of the Elementary and Secondary Education Act, now known as the Every Student Succeeds Act (ESSA), which formally ended the No Child Left Behind era. President Obama signed the new legislation into law on December 10, 2015 and, although it also officially ended the era of waivers, key elements of the waivers still remain. In particular, the new law requires states to identify a set of schools needing "comprehensive support" that, among other things, make up the lowest-performing 5 percent of Title I schools, comparable to Priority Schools, as well as a set of schools needing "targeted support" that have consistently underperforming subgroups of students, comparable to Focus Schools (National Association of Secondary School Principals, n.d.). With regard to Focus Schools, states are required to tailor interventions based on unique

school needs and to couple these interventions with support and oversight but are otherwise given flexibility. Examining the impact of the reform models that developed in different states under waivers offer us an important opportunity to inform the in-progress implementation of ESSA.

3. WAIVERS AND ACCOUNTABILITY IN LOUISIANA

In this section, we provide an overview of NLCB flexibility waivers as well as Louisiana's waiver in the context of the state's existing accountability system, and discuss what this means for the particular treatment contrast examined in this paper.

3.1 NCLB Flexibility and Differentiated Accountability

In September 2011, the U.S. Department of Education announced an official application process through which states could apply for flexibility waivers from the toughest requirements of NCLB. In particular, states would be released from the requirement that they set performance standards for schools and districts to reach the goal of 100% student proficiency on math and reading standards by 2014, which by then had been widely criticized as unrealistic. Additionally, districts were released from the requirement that they respond in a number of specified and increasingly consequential ways toward Title I schools that failed to meet their determined adequate yearly progress (AYP) for two years in a row, such as requiring them to provide supplemental educational services (SES).

In exchange for this and other forms of flexibility, states had to make plans to adopt a number of new educational policies. Chief among these was the adoption of "college- and career-ready" content standards in English and Math (which the majority of applying states fulfilled through the adoption of the Common Core), and the development of a system of differentiated recognition, accountability and support. Under this new differentiated accountability system, states would be required to identify two categories of low-performing schools and implement particular interventions in them. The lowest-performing group, those identified as "Priority Schools," would be made up of at least five percent of the state's Title I schools and would receive multifaceted and prescriptive interventions consistent with federal turnaround principles. The second group, "Focus Schools," would be Title I schools that were contributing to the state's achievement gaps, or Title I high schools with a graduation rate below 60 percent for multiple years. The Focus Schools had to constitute at least ten percent of the Title

I schools in the state. The interventions required for this second group were not particularly prescriptive as states were required to implement evidence-based interventions in Focus Schools that were based on assessments of the particular needs of each school and its students (US Department of Education, ESEA Policy Document, 2012). Importantly, unlike some federal policies that dictated interventions for struggling schools, such as School Improvement Grants (Dee 2012), the accountability requirements introduced under the waivers were not attached to any additional funding. So, while systems of assessment and interventions were required for the Priority and Focus Schools, states had to find ways to use their existing Title I funds to pay for these systems.

3.2 The Context of Accountability in Louisiana

Louisiana has been on a trajectory of increased school accountability prior to the enactment of NCLB. In the late 1990s, public criticism over the performance of Louisiana's schools pushed the state to revamp its testing and accountability structures. The state's standardized tests, known as the Louisiana Educational Assessment Program (LEAP) had been in place since 1989, but the tests set a low bar for passing and poor performance was not attached to any consequences for schools. So the state developed new LEAP tests with math and English rolling out in 1998 and science and social studies in 1999. These tests were in place until the 2014-15 school year, when they were replaced by new exams aligned with the Common Core State Standards.[1]

The new LEAP tests were accompanied by a new policy of publicly issuing School Performance Scores (SPS) based primarily on students' test results. School Performance Scores—the calculation of which we discuss in section 4.1—fell on a scale of 0-200 and were intended to bring increased accountability and transparency. When the SPS system first began, schools that earned fewer than 30 points were labeled as Academically Unacceptable Schools (AUS); this threshold for AUS status gradually increased over time. With each additional year that a school was labeled Academically Unacceptable, it was required to implement certain strategies meant to spur improvement, such as required reporting, limited school choice, and

[1] In 2015, Louisiana lawmakers decided that they would not continue using the complete Common Core aligned exams developed by the Partnership for Assessment of Readiness for College and Careers (PARCC) the following year. Instead, the assessments for 2015-16, referred to as updated LEAP tests, were made up of approximately 49 percent PARCC items and 51 percent items specifically for Louisiana. In 2014-15, the test scores from the PARCC test were entered into the calculation of School Performance Scores in the same way that the LEAP scores had previously been included.

Supplemental Education Services (SES), in accordance with NCLB requirements. Sanctions on AUS also included the threat of takeover by the state-run Recovery School District, which was created in 2003 and originally was intended as a last resort to turn around failing schools.[2]

In the late 2000s, in an attempt to attach more meaning to the School Performance Scores, the state created performance labels corresponding to the numbers: in addition to being labeled Academically Unacceptable if they received fewer than 60 points, schools received stars—from one star up to five—if they earned 60 or higher, with scores 140 and above reserved for five stars. If schools were chronically labeled Academically Unacceptable, they risked takeover by the state. Despite this attempt to make the School Performance Scores meaningful to the public, the state continued to face complaints that the system was unintuitive, difficult to understand, and a poor representation of schools' actual performance. The state legislature responded by passing Act 718 in 2010, which mandated that the state assign annual letter grades to schools based on their SPS and publicly announce them every fall. The state announced the first set of A through F school letter grades in October 2011. In this first year, the cutoff for receiving an F was an SPS less than 65 (though this cutoff would change in subsequent years). 115 schools received an F letter grade based on their 2010-11 data.

3.3 Louisiana's Introduction of NCLB Waivers

Soon thereafter (i.e., February 2012), the Louisiana State Department of Education (LDOE) submitted its NCLB waiver application. To fulfill the waiver's requirement to establish a system of differentiated accountability, Louisiana leveraged the school accountability system it was already in the midst of strengthening and simply aligned the two. Instead of using the Focus and Priority school definitions that the U.S. Department of Education had prescribed, Louisiana created its own definitions: under the waiver, Priority schools would be all schools in Louisiana's state-run Recovery School District, and Focus Schools would be all remaining schools that either a) received F letter grades, or b) were high schools with graduation rates below 60%. With these state-specific definitions, Louisiana's Focus Schools were similar to what most states would deem their Priority schools—they were the lowest performing schools,

[2] The state-run Recovery School District (RSD) was created in 2003 with the stated intent of taking over and turning around chronically underperforming schools. Initially the criterion for RSD takeover was four consecutive years of academically unacceptable status; however, following Hurricane Katrina, the state expanded the district's eligibility criteria to include any school with a below-average performance score or schools in an "academic crisis" district. The vast majority of the district is now made up of schools in New Orleans (68 of 80 in 2012-13).

based on their SPS, without particular consideration for achievement gaps or Title I status in terms of their selection.

The U.S. Department of Education approved Louisiana's waiver in May 2012. In October 2012, the state also released its first official list of Priority and Focus Schools and letter grades for each school along with the baseline School Performance Scores on which these designations were based. According to the LDOE's timeline, the implementation of waiver reforms began at this time (i.e., at the start of the 2012-13 school year). The first cohort of Priority schools simply consisted of the 80 schools in the Recovery School District. The first cohort of Focus Schools consisted of 135 schools.[3]

The one-year lag between the timing of the first school-letter grades (i.e., October 2011) and the first Focus School designations (i.e., October 2012) merits further commentary. In particular, it should be noted that our RD design relies on the baseline SPS used to determine the first cohort of Focus Schools but the *second* cohort of F-rated schools. As a practical matter, there is considerable overlap between the first and second cohorts of F-rated schools. Of the 115 schools that received an F in the first year of the letter-grade system, only 5 advanced to D grades in the following year while 68 retained F status in both years. Of the remaining 42 schools, 38 closed and 4 were taken over by charter organizations and labeled as "Transition" schools under the grading system. Therefore, these schools are not part of our "intent to treat" (ITT) population. The closure (or transition) of some low-performing schools prior to the Focus School determination provides a modest external validity caveat to our findings. However, it should also be noted that the threshold used to determine the first cohort of Focus Schools (and, correspondingly, the second cohort of F-rated schools) was increased from 65 to 75. This increase implies that our ITT population also includes a number of schools that had a D rating in the prior year. We discuss the treatment contrast created by the assignment rule and the corresponding Focus reforms and ratings below.

3.4 The Focus School Treatment Contrast

The aim of this paper is to identify the causal impact of the federally mandated Focus School designation on schools' student outcomes in Louisiana. But what exactly does it mean for

[3] Nearly all of these Focus Schools (n=129) were eligible because their 2011-12 SPS were below 75 and, so, they also received an F grade. An additional 6 schools received Focus status (but not an F rating) based on having a high-school graduation rate below 60. As we describe below, we exclude all high schools with graduation rates below this threshold and focus on the SPS eligibility margin.

a school to be designated a Focus School in this particular state? Because Louisiana built its waiver reform policies on its existing school accountability system, the answer is twofold. Being a Focus School meant receiving specific types of attention under the state's waiver reforms, as well as being labeled a failing school via well-publicized letter grades, both of which are determined by earning a low SPS.

We consider the former component of the treatment contrast first: interventions established under Louisiana's existing accountability system, which they aligned with their NCLB waiver. In its waiver application, the LDOE outlined the supports offered to Louisiana schools. These reforms had two distinctive features. One was a comprehensive data review and needs assessment to help schools diagnose problems and to determine necessary programs/interventions. The second was a coordinated system of support through the LDOE's technical assistance network, which serves all schools but prioritizes the needs of Focus Schools and focuses primarily on the implementation of Common Core and teacher evaluation systems. The state also indicated that students in Focus Schools would be given the option of transferring to another school with a higher grade, with costs covered by the district. In our analyses, we examine whether such mobility occurred and whether it constitutes an internal validity threat.

Because Focus School interventions were intended to address schools' specific needs, the interventions were only broadly defined and relied heavily on effective processes for identifying needed supports. However, the available implementation evidence suggests that the state may not in fact have had the systems in place to adequately assess the needed supports and implement effective solutions. In August 2013, the U.S. Department of Education conducted a monitoring review and found Louisiana's implementation of supports for Focus Schools to be "not meeting expectations." The evaluation report noted that no evidence had been provided to show that targeted interventions were being implemented in the Focus Schools (U.S. Department of Education, 2014). In December 2014, Louisiana was granted a waiver extension from the federal government, but the letter granting the extension warned that the waiver's renewal was contingent on improving its plans for implementing and monitoring school improvement interventions for Focus Schools.

This fairly limited view into the state of Louisiana's waiver implementation indicates that the Focus School interventions as stated in their application may not have been implemented with fidelity, at least during their first two years. Anecdotally, our review of state documents

suggests that the set of treatments and support offered to Focus Schools resemble those offered to all other schools, with the exception of particularly harsh sanctions, suggesting that the special interventions described in the waiver application were not clearly distinctive nor well implemented. If this is indeed the case, then the primary treatment contrast between Focus and non-Focus Schools that we are identifying may have been the impact of being publicly labeled an F school rather than the impact of receiving a standard set of "Focus interventions." While there was little press coverage in Louisiana of the Focus and Priority school designations once the waiver reforms were put in place, the annual release of School Performance Scores is regularly covered in the news and the introduction of their newly associated letter grades was widely discussed. The letter grade designation is itself a mechanism for accountability and perhaps a source of motivation for schools and their districts.

4. DATA AND SPECIFICATIONS

In this section, we discuss details on our data and analytical sample as well as the basic econometric specifications we use in this paper.

4.1 Analytical Sample and Variables

For our analysis we use publicly available school-level data provided annually by the LDOE on School Performance Scores, the underlying state standardized test scores, and Focus School assignments from 2012 to 2015. We supplement this with data on Louisiana schools from the NCES Common Core of Data, which provides information on schools' Title I eligibility status, the demographic composition of students, the share of students eligible for free and reduced-price lunches, and student-teacher ratios.

Our analytical sample consists of traditional primary and secondary public schools. According to LDOE data, there were 1,303 primary and secondary schools in Fall 2012 with valid baseline School Performance Scores (i.e., the assignment variable in our RD design). To maintain a focus on traditional public schools that were not already implementing federally prescribed reforms, we drop special education schools (n = 4), vocational schools (n=3), alternative schools (n = 20), schools with prior School Improvement Grants (n = 12), and charter schools (n=84). These edits collectively reduce our sample to 1,182 schools. We retain magnet

schools in the sample.[4] Additionally, because all schools in the Recovery School District are designated as Priority schools and are ineligible for Focus School assignment, we drop all remaining Recovery School District schools (n = 10), bringing our sample to 1,172 schools.

In this paper, we leverage a regression discontinuity design (discussed in more detail in the following subsection) to identify causal impacts of Focus School assignment. Schools were assigned to Focus School status based on two rules: whether they a) fell below the SPS threshold for an F letter grade, or b) were a high school with a graduation rate below 60 percent. We also eliminate the small number of remaining high schools that had a graduation rate of 60 percent or below in 2011-12 (n=14).[5] This allows us to isolate the treatment effect of assignment across our primary "frontier" of interest, the SPS threshold. We privilege this assignment variable because the vast majority of Focus Schools were identified as such due to their SPS rather than their graduation rate. Our final analytical sample is made up of 1,158 schools, of which 94 are Focus Schools. These 1,158 schools include 681 elementary schools, 217 middle schools, and 260 high schools.[6] The 94 Focus Schools are all school-wide Title I schools.

Our three main dependent variables of interest are School Performance Scores (SPS) in each of the three years following the implementation of the Focus School reforms (i.e., their 2012-13, 2013-14, and 2014-15 SPS). The SPS measures vary by the grades a school serves and, during this period, has a maximum value of 150. For schools serving grades from pre-kindergarten through 6th grade, the SPS is an index based on student performance on state standardized tests in each of the four core academic subjects (LEAP and iLEAP).[7] For schools serving grades from pre-kindergarten through 8th grade, 95% of the SPS consists of the test score index while the remaining 5% is based on the success of the school in supporting its students' transition to high school (i.e., based on dropout behavior and credit accumulation). For schools

[4] To identify alternative schools, we went through those schools marked as alternative in the Common Core of Data—which included both alternative and magnet schools—and dropped those schools with the words "alternative" or "center" in their names, those that had an unusually small student body (student n<10), and those described as alternative schools in online searches.

[5] To check whether limiting our sample in this way affects our estimates, we also estimate our main results for an analytical sample with all high schools included (using 2SLS instrumental variable regression, as this becomes a fuzzy regression discontinuity) and with no high schools included. These alternative specifications do not substantively affect our results.

[6] For more detail on how the different school levels are defined, see Appendix I.

[7] The Louisiana Educational Assessment Program (LEAP) is a criterion-referenced test taken by 4th and 8th grade students in each of the four core subjects. The integrated LEAP (iLEAP) is a norm and criterion-referenced test in each core subject taken by students in grades 3, 5, 6, and 7. In our last study year, the test component of the SPS tracked the state's switch to a test aligned with the Common Core.

serving grades 9 through 12, the SPS reflects equal 25% weights on ACT scores, end-of-course exams, a diploma index based on Advanced Placement and International Baccalaureate exams, and a cohort graduation rate. The SPS for schools serving a combination of grades is a weighted average of the grade-appropriate SPS. There is a modest amount of missingness in the SPS over the first three years of Focus School reforms (i.e., 1 school in 2013, 17 in 2014, and 27 in 2015). These missing data are due almost exclusively to school closures. However, auxiliary RD estimates indicate that this missingness is balanced across the Focus School threshold and, therefore, does not appear to constitute an internal-validity threat.

Our focus on the SPS as an outcome is sensible in that it is the performance measure that both determined Focus status and whether a school would be seen as improving. However, in order to focus specifically on student learning (and possibly heterogeneous effects by subject), we also use as a dependent variable school proficiency rates on LEAP/iLEAP exams by subject in grades 3-8 for the 2012-13 and 2013-14 school years. Unfortunately, neither the SPS nor the available test data provide information on the cognitive performance of student subgroups. However, we strongly suspect that a subgroup analysis would not yield significantly different results from the school-level analysis given the racial and ethnic makeup of schools near the threshold. Around the cutoff, the average school share of black students in 2012-13 was about 0.8 and the average share of free and reduced price-eligible students was around 0.9. The relatively homogeneous makeup of these student bodies suggests to us that drawing different conclusions from subgroup analyses is unlikely.

Finally, the covariates we include in our analysis are a schools' Title I eligibility, percentages of black and Hispanic students, the percent of students eligible for free and reduced price lunch, student-teacher ratio, and fixed effects for grade level (primary, middle, or high). These data come from the 2012-13 NCES Common Core of Data (CCD) to reflect the year the implementation of the Focus School policies began. In cases where schools are missing 2012-13 data, we use their previous year's data, which we obtain from the 2011-12 CCD. In examining covariate balance, we relied on these data as well as school characteristics for the 2011-12 year (i.e., the year before treatment status was announced).

4.2 Regression Discontinuity Design and Assignment to Treatment

In this study, we use a "sharp" regression discontinuity (RD) specification to estimate the causal impact of Focus School designation on the performance of Louisiana schools. This estimation strategy leverages the fact that Focus School assignment is determined by whether a continuous rating variable— here, the baseline SPS— is above or below an arbitrary threshold value. As mentioned previously, we eliminate all high schools with a graduation rate of 60% or below from our sample, in order to conduct a "frontier RD" analysis that focuses specifically on schools for which the SPS assignment rule applies (i.e., the vast majority of Focus Schools). The resulting assignment scenario creates a sharp and plausibly exogenous assignment between treatment and control groups among those schools with School Performance Scores very close to the threshold. This straightforward RD design thus allows us to determine the effect of the combined Focus School and accountability labeling for those schools local to the "failing" threshold in their 2011-12 School Performance Scores.

We use the following general model to identify the estimated treatment effect α of Focus School assignment on school outcome Y_i:

$$Y_i = \alpha I(S_i \leq 0) + h(S_i) + \Gamma X_i + \varepsilon_i \quad (1)$$

Here, α signifies the discrete jump that occurs at the cut score for Focus School assignment, $h(S_i)$ is a function of the centered School Performance Score, and X_i is a vector of school-level covariates. Determining the correct form of the function $h(S_i)$ is an important consideration so we consider several forms of relevant evidence. For example, we consider unrestrictive graphical presentations that allow us to examine the underlying functional form of $h(S_i)$. In our estimates of equation (1), we also allow the relationship between the rating variable and the outcome variable to vary above and below the threshold, as well as model this relationship both linearly and with a quadratic function. We also consider alternative estimates that use subsets of the data within increasingly tight bandwidths around the threshold as another check for the robustness of our results (i.e., non-parametric local linear regressions).

Before estimating treatment effects, there are multiple key assumptions that are necessary to examine. One assumption is that there is no manipulation of the underlying continuous forcing variable, or in other words, that SPS scores are independent of the cut point and, for those schools near the threshold, falling on one side versus the other of the cut score is as good as random. While it is reasonable to think that schools would have a strong motivation to score just

above the cutoff point for an F letter grade, there is little evidence to suggest that they would be able to systematically manipulate their SPS to do so. School Performance Scores are based primarily on standardized test scores, which would be difficult for schools to manipulate outside of outright cheating (e.g., changing student answers). Additionally, 2011-12 was only the second year that the state assigned letter grades based on SPS. Although the components of each school's SPS were clearly stated, it was likely not clear to schools how their day-to-day actions would translate into a score that was just high enough to keep them from avoiding F status. Such anticipatory responses on the schools' part seems even more implausible given that in 2011-12, the threshold for F grades was raised to 75 or below, up from 65 or below the previous year, giving schools little experience of what "just good enough" educational practices would look like on the ground.

Figure 1 shows the density of the 2011-12 School Performance Scores, centered at the cut score of 75, which throughout the paper we will refer to as the forcing variable. The SPS are calculated to the 0.1 decimal place. The density test introduced by McCrary (2008) examines the null hypothesis that the distribution of observations is smooth at the threshold. A rejection of this hypothesis would indicate that observations cluster on one side of the threshold (i.e., possibly due to manipulation). Figure 1 shows that there are no significant jumps in density at the cut score. Figure 2 shows histograms of the forcing variable, to give us a better perspective on whether there is any abnormal heaping of the forcing variable that may not be apparent in an estimated density. While the full histogram suggests potential evidence of heaping on the right hand side of the cut score, a zoomed-in version with a smaller bin width (Panel B) shows little evidence of abnormal heaping.[8] Overall, this graphical evidence collectively suggests that there is no evidence of manipulation of the forcing variable.

Another key consideration for an RD estimation is determining whether the regression discontinuity in question is sharp or fuzzy. In other words, to what extent do schools in our

[8] To ensure that heaping is not an issue, we test it in two ways. First, we drop all observations for which the SPS's have a value frequency of 5 or greater, which eliminates 174 observations from the sample. Doing so does not change the estimated main results in terms of either sign or significance, and changes in magnitude are only minor. Secondly, we drop all observations on the right-hand side of the cut score that are both within five points of the cut score and have a frequency of 4 or greater. This eliminates groupings of observations that we might worry are unnaturally clustered just above the cut score. This reduces the sample by 27 observations. Doing so also does not affect the sign or significance of the main results, with the exception that the full sample, no controls specification for 2014 SPS effects becomes -4.236 with a standard error of 1.962, statistically significant at $p<0.05$. However, with controls, the estimate for this specification is still statistically insignificant (-1.592, standard error of 2.003).

sample fully comply with their Focus School treatment assignment determined by their school performance scores? Figure 3 shows the probability of a school being a Focus School based on their 2012 school performance score, centered at the cut score of 75. We can see that at the cut score, the probability jumps sharply from 0 to 100. Indeed, once we limit our sample to the "frontier" sample by eliminating all high schools with graduation rates of 60% or below, the regression discontinuity is completely sharp. However, it should be noted that, over our 3-year study window, a small number of schools (n=25) entered Focus status. This would introduce some fuzziness in our first-stage relationship if one chose to define treatment as *ever* being in Focus status rather than as being in the first large Focus cohort. Our analysis emphasizes the reduced-form effects of the intent to treat (ITT) but the implications of this later fuzziness with regard to defining a "treatment on the treated" (TOT) estimate should be noted.

Finally, we have the assumption that schools in a close neighborhood to the cut score are not systematically different depending on which side of the threshold they are on. To test the validity of this assumption, we consider whether there is any evidence that these schools differ based on observables by analyzing whether the observable covariates included in our analysis—school type (elementary, middle or high), student-teacher ratio and percentages black, Hispanic, and free and reduced price lunch eligible in the 2012-13 school year—are continuous across the threshold. Our findings suggest that the schools are not observably different on either side of the threshold, which supports our assumption that they are no different on un-observables either.[9] We also test for imbalance in missing outcome data across the threshold out of concern that school closures or other causes of missing data might be more likely for failing schools, but we find no evidence of imbalance. As mentioned in Section 3.4, a considerable number of schools that received F's in 2011 – 38 in total—closed prior to the start of the 2012-13 school year. It is possible that in the first year of the letter grade system, an F grade induced the worst schools to close and effectively weeded out the very bottom, so that by the following year when Focus

[9] In analyzing our covariate balance, we found that the distribution of the covariates across schools led to unusual functional forms, in large part because of the high degree of racial and income-based segregation in Louisiana schools. This led us to believe the full sample, linear specification of the RD estimate for these covariates was a poor fit for the data. However, we also found an occasional significant estimate of an imbalance in a covariate variable across the threshold. To more deeply explore whether these occasional results were evidence of a covariate imbalance, we also estimated a model where the dependent variable was a regression-weighted index of the covariates, each weighted by their estimated effects on the outcomes. The results from this additional model suggest that there were not significant imbalances in the covariates. See Appendix II for more details and for the results of all models.

reforms began, an F grade did not disproportionately induce schools to close. See Appendix II for more details on covariate balance.

5. RESULTS

We first illustrate our findings graphically by showing the subsequent school-performance measures as a function of the baseline SPS that determined a school's Focus status (and whether it received an F label). First, panel A of Figure 4 shows the SPS for our full analytical sample in 2012-13 (i.e., the first treatment year). Panel B shows the same data but only within a bandwidth of 20 points relative to the threshold value that determined treatment status. Figures 5 and 6 similarly illustrate SPS for 2014 and 2015, respectively. These figures suggest that the relationship between current and baseline SPS scores exhibits mild curvature over the full range of data but is more clearly linear over tighter bandwidths of the data. More critically, this visual evidence does not show any notable jumps in school performance at the threshold for any year with the possible exception of a modest *decrease* after 3 years (i.e., the 2015 SPS in Figure 6).

In Table 2, we present the key regression results for versions of equation (1) that correspond to these figures. The estimated parameter of interest, α, identifies the jump in future SPS measures at the threshold that defines Focus School (and F) status. The first two columns for each outcome measure show results for a linear spline model, where the slope can change on either side of the threshold. The second two columns for each outcome measure show results for a quadratic spline model, where both the slope and the curvature can change on either side of the threshold. School-level controls are included in the even-numbered columns in the table.

The results for the most part suggest that being part of the first cohort of Focus Schools did not have a significant impact on the performance of those schools. There are a few exceptions, the first of which is in column (1), where we see a statistically significant (and *negative*) effect when the outcome variable is 2013 SPS. However, this specification includes no school controls and when controls are added, the point estimate is cut in half and the statistical significance goes away. Additionally, we see significant negative point estimates in columns (9) and (10) for the linear specifications of the 2015 SPS outcomes. The significance remains even when controls are added. This suggests that in the third year following the start of the schools'

Focus treatment, this first cohort of schools was actually performing significantly *worse* than other low-performing schools.

We do not take this finding at face value, however. Referring back to the first panel of Figures 4, 5, and 6, we consistently see that the distribution of performance outcomes when ranked by the centered rating variable has considerable variance in the tails, particularly the upper tail. It is straightforward to see that a linear spline model using the full sample of data does not fit the data particularly well. To assess this observation more formally, we also calculated the Akaike's information criteria (AIC) for these full sample models and include them in Table 2. A smaller valued AIC indicates a better fit of the model to the data and, as the table shows, the information criteria indicates that, when using the full sample, a quadratic spline model is a more appropriate specification for all three years of SPS outcomes. Given this, our estimates suggest that there was no significant impact—positive or negative—of Focus School assignment on the first cohort of schools.

To test the robustness of our results, we also estimate our regression model described above using alternative bandwidths, as discussed in Section 4. Table 3 shows our results for our estimated α coefficient using the full sample as well as results for local linear regressions using subsamples defined by bandwidths of 30 points down to 8 points. For context, the suggested bandwidth that is calculated using the Calonico, Cattaneo, and Titiunik (2014) procedure ranges from 7.4 for the models with 2015 SPS as the outcome to 9.3 for models with 2014 SPS as the outcome. The optimal bandwidth according to the Imbens and Kalyanaraman (2012) algorithm ranges from 15.0 when 2013 SPS is the outcome to 26.7 when 2014 SPS is the outcome. And the bandwidth suggested from a cross-validation procedure that aims to minimize mean squared error ranges is estimated to be 25.6, regardless of the outcome year. In addition to varying the bandwidths, we also estimate the α coefficient using a triangular kernel-weighted subset of data, where we weight data points by decreasing amounts the further they are from the threshold. With the exception of the spurious full-sample evidence for negative effects noted above, none of the alternative specifications yield a significant estimate for the treatment effect for 2013, 2014, or 2015 results.[10]

[10] Linear probability models that used an indicator variable for having an F grade in 2013, 2014, or 2015 as the dependent variable yielded similar results. The estimated jump coefficients were insignificant across all bandwidth restrictions with the exception of the full sample.

We also examine whether stacking our three years of outcome data offers us any additional insights by increasing the power of our estimates. To do so, we run regressions for both full-sample linear and quadratic splines, as well as local linear regressions with restricted bandwidths, for models where the unit of observation is school by year and our controls include year fixed effects. Using these specifications, we similarly find no evidence that these reforms led to statistically significant increases in school performance. We also examined stacked-year models that allow for separate jump parameters for each of the three outcome years. For this analysis, we find that the 2015 α coefficient is negative and statistically significant for the full-sample linear specification and the full-sample quadratic. The full-sample quadratic has a lower Akaike's information criteria than the linear, suggesting it is a better fit to the data. However, an F test of the equivalence of the three treatment effects in both the linear and quadratic models indicate that we cannot reject the null that the treatment effect is the same across years, so we again conclude that the evidence around the 2015 outcomes is only suggestive.

It is possible that all of these aggregate models may be masking heterogeneous effects of Focus School assignment that differ by type of school. More specifically, it is reasonable to think that the types of interventions and responsiveness to interventions may differ depending on whether a school is a primary, middle, or high school. Additionally, because the components of School Performance Scores differ by school level, it is possible that some components are more responsive to the treatment than others. To examine these possibilities, we estimate separate regressions for primary, middle and high schools. The results for full sample, linear-spline specifications are shown in Table 4.[11] These models do not yield significant treatment effects for primary, middle, or high schools. The one exception is the estimate for the effect on 2014 SPS outcomes for high schools. However, when school controls are added the point estimate changes significantly and the statistical significance goes away. The large change in estimates with and without controls is likely because there are only five Focus high schools, making the estimates for the impact on their performance fairly imprecise. Although not included in Table 4, the

[11] For the heterogeneous effects models, we ran both linear and quadratic spline models and found that for full sample specifications, the Akaike information criteria was smaller (signifying a better fit to the data) for the linear model than for the quadratic for seven out of the nine models. Of the two models for which a quadratic was a better fit, middle school 2014 and 2015 SPS models, the quadratic model yielded a significant estimate for only one. The middle school 2015 SPS estimated effect with school controls is -10.29 with a standard error of 4.898; however, this significance is not robust to linear models estimated with narrower bandwidths of data.

results for heterogeneous effects (elementary, middle, and high) are robust to alternative models that limit the data to narrower bandwidths around the threshold.

We also consider the possibility that using the composite School Performance Score as an outcome measure may obscure effects on test scores in particular subjects (i.e., LEAP and iLEAP). We therefore run our regression models using the available 2013 and 2014 grade-level data from each school on proficiency rates in Math, ELA, Science, and Social Studies tests as our outcome variables, stacking the data and including grade fixed effects. Because the unit of observation is school-grade, the sample size in each subject is larger than in our main school-level models, with 3,178 to 3,179 observations in the full sample for 2013, and 3,137 to 3,139 in the full sample for 2014 (depending on subject). The missingness for LEAP/iLEAP scores for each subject in each year is balanced across the threshold. Following Louisiana's own definition of "proficient," we calculate the percent proficiency as the percent of students scoring Advanced, Mastery, and/or Basic on the state tests, across all grades in the school that take the LEAP or iLEAP in the subject of interest.[12] Our results using both the full sample and a subsample limited to a bandwidth of 20 points are shown in Table 5. Consistent with our other models, we find no evidence of significant impacts on test performance. Our results do not change when we examine effects on scores for LEAP and iLEAP separately.

Additionally, in any study that examines the effect of a treatment on school-level outcomes, there is the concern that any impacts detected are not the result of a true treatment effect on the school, but rather are the result of shifts in the student population. Because we are not working with student-level data, we cannot define the intent to treat at the student level. We examine the number of students enrolled, the percentages of students who are free and reduced price lunch eligible, black, Hispanic, and white, as well as the student-teacher ratio at schools in the years following Focus School Assignment (2013-14 and 2014-15).[13] Using these measures, we do not find evidence that Focus Schools' student enrollment changed within the first two to three years after their assignment, relative to any changes that occurred at other schools close to the SPS threshold. These findings are consistent with those from research on the impacts of the

[12] Louisiana schools earned points toward their SPS for students scoring at Basic or above, and students below this level were considered non-proficient. This definition of proficiency does not align, however, with NAEP definitions, for which "Basic" is considered below "Proficient."

[13] See Appendix III for details. At the time of writing, the 2014-15 school year only has data on total enrollment and percent FRPL available. We are therefore unable to test for discontinuities in the 2014-15 percentages of black, Hispanic, or white students or in student-teacher ratios.

school grade accountability systems in Florida (Chakrabarti, 2007; Figlio and Rouse, 2006; Chiang, 2009), New York City (Winters and Cowen, 2012), and Michigan (Hemelt and Jacob, 2017), all of which do not find evidence that changes in student composition were drivers of the patterns of school performance they found.

6. DISCUSSION

This paper provides evidence on the effect of a new differentiated and flexible accountability system implemented in Louisiana in the era of ESEA waiver reforms; specifically, the effect that the system had on low-performing schools identified as Focus Schools. The impact of the accountability systems states established under waivers is particularly interesting given the recent passage of the newly authorized ESEA legislation, titled the Every Student Succeeds Act, which offers guidance around school accountability for low-performing schools that closely mirrors the requirements outlined for Priority and Focus Schools under such waivers. Our estimates indicate that low-performing schools identified as Focus Schools that were close to the identification threshold did not significantly improve relative to comparable low-performing schools not put in the Focus group after the first, second, or third year of the reform. Our results suggest that, at least with regards to Focus Schools, Louisiana's accountability system has not been effective in implementing interventions for its most challenged schools that successfully improve their performance and their students' outcomes.

One potential concern readers may have with our estimation is that the relatively small sample size of first cohort Focus Schools (n=94) may give us insufficient power to detect results. To consider this, we can identify the upper bound of the estimated effect size of the Focus School/F-letter treatment. For example, using the estimates from column (4) in Table 2, the 95% upper-confidence limit for the impact estimate is approximately 3.5 points on the SPS scale, an effect size of nearly 0.20 SD *at the school level*. The implied effect size in terms of a *student-level* standard deviation is substantially smaller, implying that these estimates are fairly precise and that, at best, these reforms had a modest impact on the state's most challenged schools.

We do not have rich implementation data that would allow us to explicate in detail the seeming failure of Louisiana's Focus School reforms to improve student outcomes. However, several forms of descriptive evidence – the interventions described in the waiver, federal monitoring reports, news accounts, and conversations with state officials – point to both weak

treatment design and weak implementation. First of all, the interventions described in Louisiana's waiver application are somewhat vague. The application states that the treatments to be implemented in Focus Schools were instead to be determined by needs assessments that would determine the challenges schools faced and supports they needed.

The Louisiana Department of Education (LDOE) also noted that its capacity is "extremely limited" and that the effectiveness of turning around the performance of the schools will rely heavily on building capacity in the *districts* to take on the effort. Additionally, the state's descriptions of the technical assistance provided to districts and supports by the LDOE suggest that they are not targeted to the needs of Focus Schools. The primary mechanism for offering state technical assistance to districts and schools is the network of support teams (District Network Teams) employing education specialists that focus on six areas of school improvement.[14] Although Focus Schools are described as "a high priority" for these supports, the purpose of the teams is to support all schools and there are, by and large, no resources dedicated solely to the Focus Schools. One of the few examples we could find of Focus Schools getting particularly targeted attention was in 2013-14 when, according to the state's updated waiver application, all Focus Schools worked with District Network Teams to analyze student-level data and set goals for the upcoming school year, followed by planning meetings to create strategy toward those goals. Throughout the school year, the schools and teams continued to meet to monitor and trouble shoot progress. However, as noted above, monitoring reports written by the U.S. Department of Education and the contingencies outlined in Louisiana's waiver renewal state multiple times that the state's plan for monitoring the process of supporting Focus Schools needed improvements, and that there was limited to no evidence that interventions for Focus Schools were taking place.

A possible explanation for our null findings is that Louisiana's state system of technical assistance directed broad attention to all low-performing schools and perhaps improved the performance not just of F schools, but also D and C schools. The available information on Louisiana's waiver implementation suggests that this is highly unlikely. However, we can also speak indirectly to this possibility by noting the broad changes in SPS results over this period. Between 2012-2013 and 2014-15 school years, the mean SPS does not increase for the set of F

[14] The six areas that the District Network Teams focus on are: school leader and teacher learning targets, assessment and curriculum, school and teacher collaboration, Compass observation and feedback (their teacher/principal evaluation tool), pathway to college and career, and aligned resources.

schools, for the set of F and D schools, or, for the set of F and D and C schools. This suggests that an improvement across all low performing schools does not explain our findings unless the relevant counterfactual was one of broad declines in school performance. The overall mean of the SPS across all schools in our sample has similarly not been increasing, suggesting that if there is broad improvement across a larger set of low-performing schools, it is not being reflected in the key metric of the state's accountability system.

These null findings resemble those based on Focus and Priority School reforms in Michigan (Hemelt and Jacob, 2017) but differ with respect to the Focus School reforms in Kentucky (Bonilla and Dee, 2017). For Kentucky, the authors found positive impacts as a result of Focus School assignment in both math and reading for the targeted "gap group" students. Although the key drivers of these impacts are unknown, the authors found suggestive evidence that comprehensive school planning and higher quality professional development were key mediators of these effects. Nonetheless, even if the Focus School treatment were weak, our findings of null effects on this group of schools are fairly surprising given the evidence that exists around the effectiveness of consequential accountability in the form of publicized school letter grades. In both Florida and New York City, researchers have found that receiving a low letter grade led to school performance improvement both in the short-term and sustained over time (Greene, 2001; Figlio and Rouse, 2006; West and Peterson, 2006; Chiang, 2009; Rouse et al, 2013; Rockoff and Turner, 2008; Winters and Cowen, 2012). Coverage in the local news and the reactions of local policymakers to the letter grade system provides some perspective on their usefulness as a mechanism for improvement. In 2011, the Louisiana Federation of Teachers President Steve Monaghan publicly voiced criticism of the letter grade system, calling the letter grades "simple labels for complex problems" that are driven by politics and do not take into account differences in student populations. Monaghan also criticized significant school funding cuts made by the Jindal administration (The Jena Times, 2011). Additionally, the grading system received criticism from policymakers. In 2013-14 the state began modifying the school letter grades out of concern that they would go down as a result of the challenges of implementing the Common Core State Standards. Thus, the cutoffs were adjusted so that the distribution of letter grades that year mirrored the distribution the previous year. The same was done in 2014-15 and the state requested that this "curve policy" be extended to 2015-16. In 2015, gubernatorial candidates stated that, until the methodology behind the grades was established, the policy

should be put on pause (Sentell, 2015). This sentiment was echoed in 2016 in a document put together by the Transition Team for the newly elected Governor John Bel Edwards, which advised that stakeholders were skeptical of the grades, particularly with regard to the lack of stability in the calculations across years. The document went on to recommend that accountability ratings be put on hold until new curricula, assessments, and accountability measures could be aligned (Richard and Smith, 2016).

According to the research on accountability, and examples of effective school letter grade systems, the effectiveness of such policies hinge on public buy-in of their meaningfulness as well as available supports for struggling schools to improve their performance. On the surface, it seems that the lack of buy-in and the contextual churn of the letter grades may be part of the explanation for why they were less effective in driving improvement. Moreover, outside of the technical assistance through District Network Teams that the LDOE describes in its waiver application, it is unclear what if any special interventions or supports are available for F schools. If the state relied on the public pressure of letter grades to motivate districts and schools to seek out extra assistance and support, it appears that the lack of buy-in regarding the meaningfulness of such grades may have hurt the state's ability to drive improvement in its lowest-achieving schools.

In conclusion, our analysis suggests the uncontroversial but sometimes underappreciated fact that, when it comes to policy efforts to improve our nation's underperforming schools, local implementation and contextual details are highly relevant. This may be particularly true when the reform impetus begins with the federal government rather than through efforts initiated within state and local communities. As the United States moves towards a period that is likely to see increased flexibility for states in responding to federal education policy, this insight has implications for the role of federal oversight and the possibly heterogeneous effects of national initiatives across different states.

REFERENCES

Bohrnstedt, G., Kitmitto, S., Ogut, B., Sherman, D., and Chan, D. (2015). *School Composition and the Black–White Achievement Gap* (NCES 2015-018). U.S. Department of Education, Washington, DC: National Center for Education Statistics. Retrieved 3/6/2016 from http://nces.ed.gov/pubsearch.

Bonilla, S., Dee, T.S. (2017). The Effects of School Reform under NCLB Waivers: Evidence from Focus Schools in Kentucky. NBER Working Paper.

Borman, G.D., Hewes, G.M., Overman, L.T., and Brown, S. (2003). Comprehensive School Reform and Achievement: A Meta-Analysis. *Review of Educational Research, 73(2)*, 125-230.

Brent, G., & DiObilda, N. (1993). Effects of curriculum alignment versus direct instruction on urban children. *Journal of Educational Research, 86*, 333–338.

Calonico, S., Cattaneo, M. D. and Titiunik, R. (2014), Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs. *Econometrica*, 82: 2295–2326.

Chakrabarti, R. (2007). Vouchers, Public School Response, and the Role of Incentives: Evidence from Florida. *Federal Reserve Bank of New York Staff Reports,* no. 306.

Chiang, H. (2009). How accountability pressure on failing schools affects student achievement. *Journal of Public Economics, 93,* 1045-1057.

Dee, T. S. and Jacob, B. (2011). The impact of no Child Left Behind on student achievement. *Journal of Policy Analysis and Management*, 30: 418–446.

Dee, T.S. (2012). School Turnarounds: Evidence from the 2009 stimulus. National Bureau of Economic Research, NBER Working Paper No. 17990.

Figlio, D., Kenny, L. (2009). Public sector performance measurement and stakeholder support. *Journal of Public Economics*. 93 (9–10), 1069–1077.

Figlio, D. and Loeb, S. (2011). School Accountability. In Eric A. Hanushek, Stephen Machin, and Ludger Woessmann, editor: *Handbooks in Economics*, Vol. 3, The Netherlands: North-Holland, pp. 383-421.

Figlio, D.N. and Rouse, C.E. (2006). Do accountability and voucher threats improve low-performing schools? *Journal of Public Economics 90*, 239-255.

Greene, J.P. (2001). An evaluation of the Florida A-Plus Accountability and School Choice Program. Center for Civic Innovation at the Manhattan Institute; Program on Education Policy and Governance, Harvard University.

Hanushek, E.A., Raymond, M. (2005). Does school accountability lead to improved school performance? *Journal of Policy Analysis and Management*. 24 (2), 297–329.

Hemelt, S. and Jacob, B. (2017). Differentiated Accountability and Education Production: Evidence from NCLB Waivers, NBER Working Paper.

Imbens, G., Kalyanaraman, K. (2012). Optimal Bandwidth Choice for the Regression Discontinuity Estimator. *The Review of Economic Studies*, *79(3*), 933-959.

The Jena Times. "LFT: School letter grades an exercise in intellectual dishonesty." (Oct. 12, 2011).

Madden, N. A., Slavin, R. E., Karweit, N. L., Dolan, L. J., & Wasik, B. A. (1993). Success for All: Longitudinal effects of a restructuring program for inner-city elementary schools. *American Educational Research Journal, 30,* 123–148.

McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics, 142*(2), 698-714.

National Association of Secondary School Principals. (n.d.). Summary of the Every Student

Succeeds Act. Retrieved from https://www.nassp.org/advocacy/learn-the-issues/elementary-and-secondary-education-act-(esea)-reauthorization/summary-of-the-every-student-succeeds-act.

National Research Council. (2011). *Incentives and Test-Based Accountability in Education*. Committee on Incentives and Test-Based Accountability in Public Education, M. Hout and S.W. Elliott, *Editors*. Board on Testing and Assessment, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

O'Brien, D., & Ware, A. (2002). Implementing research-based reading programs in the Fort Worth Independent School District. *Journal of Education for Students Placed At Risk, 7*, 167–197.

Reardon, S. F. (2011). The widening academic achievement gap between rich and poor: New evidence and possible explanations. In G. J. Duncan & R. J. Murnane (Eds.), *Whither opportunity? Rising inequality, schools, and children's life chances* (pp. 91–115). New York: Russell Sage Foundation.

Richard, S.M. and Smith, A. (2016). Onward Louisiana: Committee on K-12 Education Transition Advisory Team. Retrieved from http://www.lsba.com/Images/Interior/k_12_transition_report_2-4-16.pdf.

Rockoff, J.E. and Turner, L.J. (2008). Short run impacts of accountability on school quality. National Bureau of Economic Research, NBER Working Paper No. 14564.

Rouse, C.E., Hannaway, J., Goldhaber, D., and Figlio, D. (2013). Feeling the Florida heat? How low-performing schools respond to voucher and accountability pressure. *American Economic Journal: Economic Policy, 5*(2), 251-281.

Sentell, W. (Feb. 28 2015). "What's the future of Louisiana rating public schools with letter grades? Many think it's senseless." The Acadiana Advocate. Retrieved on 3/6/2016 from http://theadvocate.com/news/acadiana/11705389-123/public-school-letter-grades-face.

Slavin, R. E., & Madden, N. A. (2000). Roots & Wings: Effects of whole-school reform on student achievement. *Journal of Education for Students Placed at Risk, 5,* 109–136.

US Department of Education. (2012). *ESEA Flexibility Policy Document, Updated June 7 2012.* Retrieved from https://www2.ed.gov/policy/eseaflex/approved-requests/flexrequest.doc.

US Department of Education. (2014). *Elementary and Secondary Education Act of 1965, as Amended Flexibility Part B Monitoring Report.* Retrieved from https://www2.ed.gov/admins/lead/account/monitoring/reports13/lapartbrpt2014.pdf.

West, M.R. and Peterson, P.E. (2006). The efficacy of choice threats within school accountability systems: Results from legislatively induced experiments. *The Economic Journal*, *116*, C46-C62.

Winters, M.A. and Cowen, J.M. (2012). Grading New York: Accountability and student proficiency in America's largest school district. *Educational Evaluation and Policy Analysis, 34*(3), 313-327.

Wong, M., Cook, T. D., & Steiner, P. M. (2015). Adding Design Elements to Improve Time Series Designs: No Child Left Behind as an Example of Causal Pattern-Matching. *Journal of Research on Educational Effectiveness*, *8*(2), 245-279.

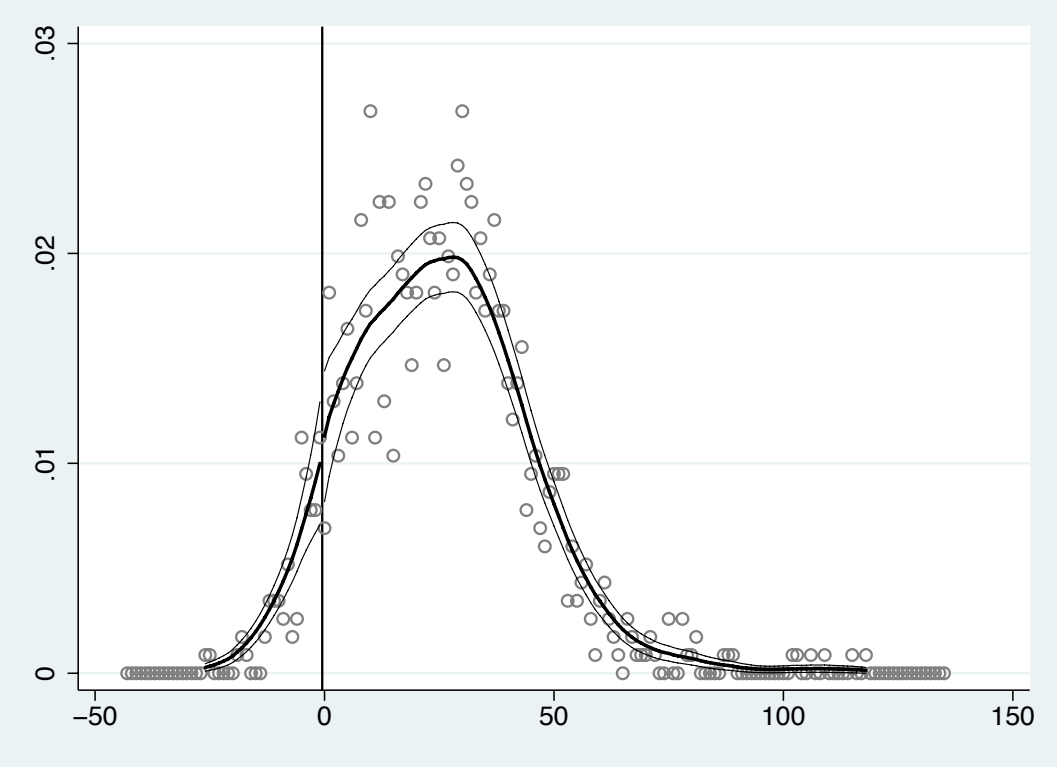
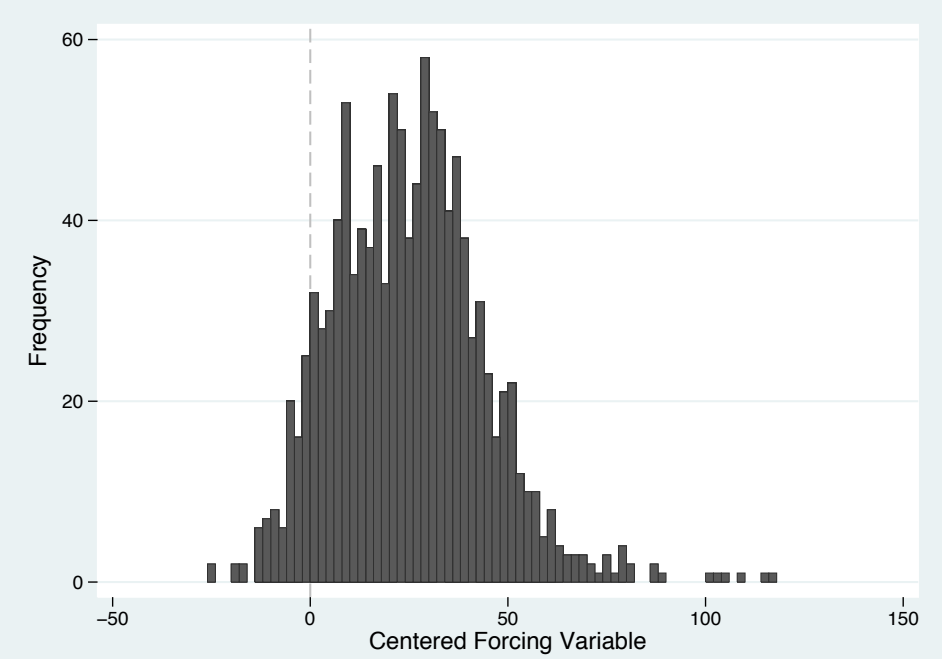


FIGURE 1 - DENSITY OF THE FORCING VARIABLE

Notes: The discontinuity estimate (log difference in height) is .0403 with standard error of .2203 (McCrary 2008). The *z* score is .183, the p-value is .86.

Panel A. Full analytic sample



Panel B. Restricted Bandwidth $S_i \leq 40$

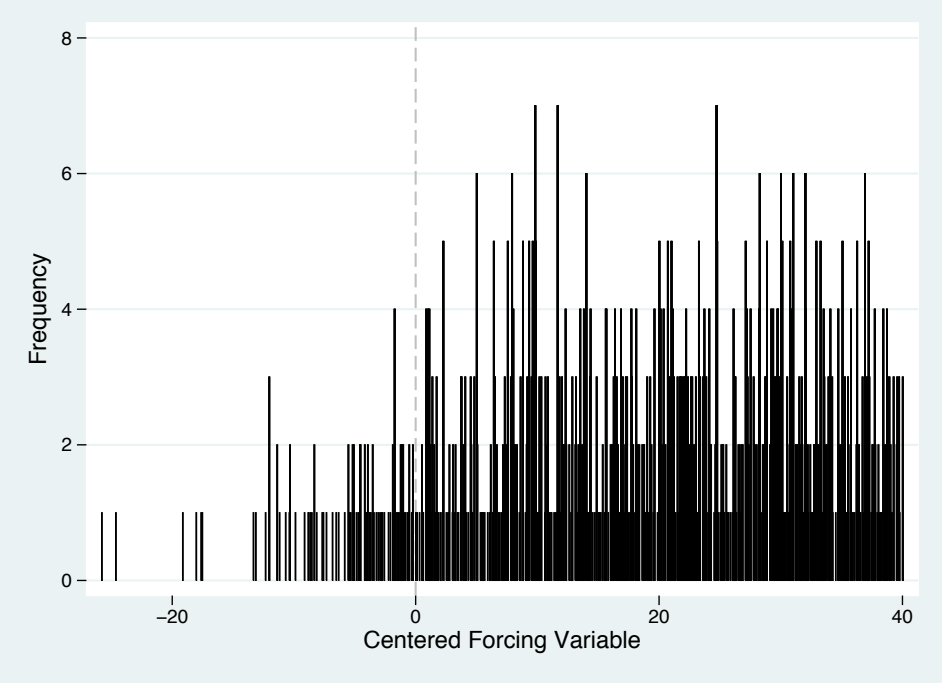


FIGURE 2 – HISTOGRAMS OF THE FORCING VARIABLE

Notes: For Panel A, bin width is 4. For Panel B, bin width is 0.1. School Performance Scores are calculated to the 0.1 decimal place. Panel B reflects the analytic sample, restricted to observations within +/- 40 points of the intent to treat cut score.

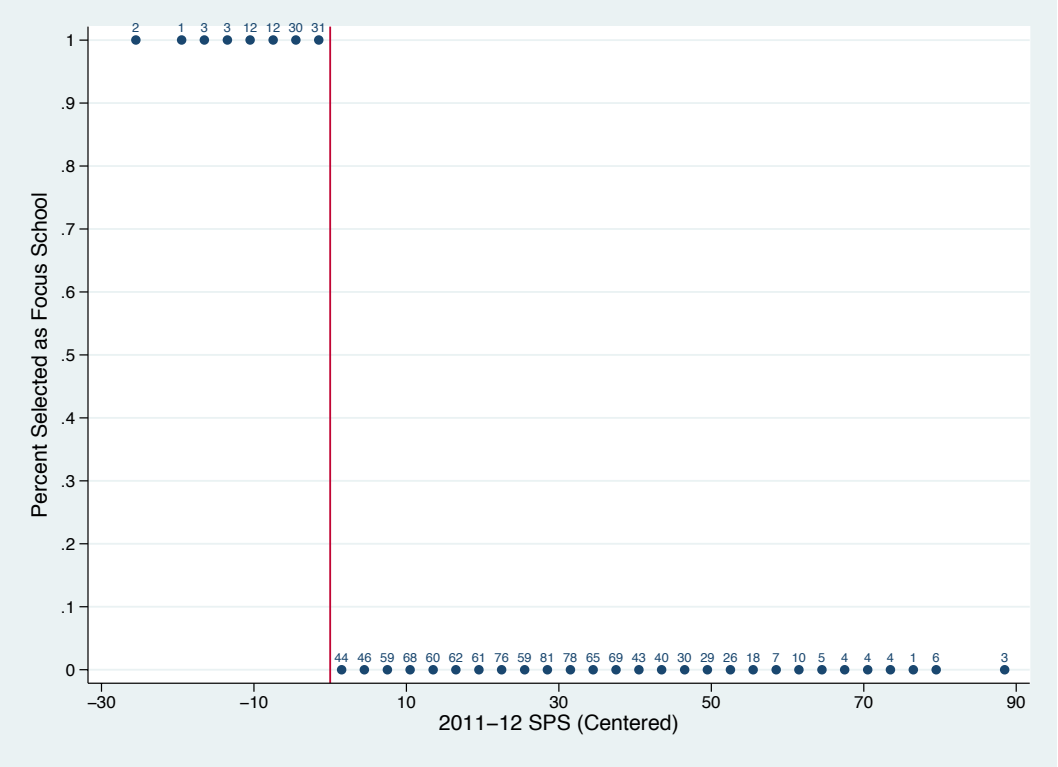
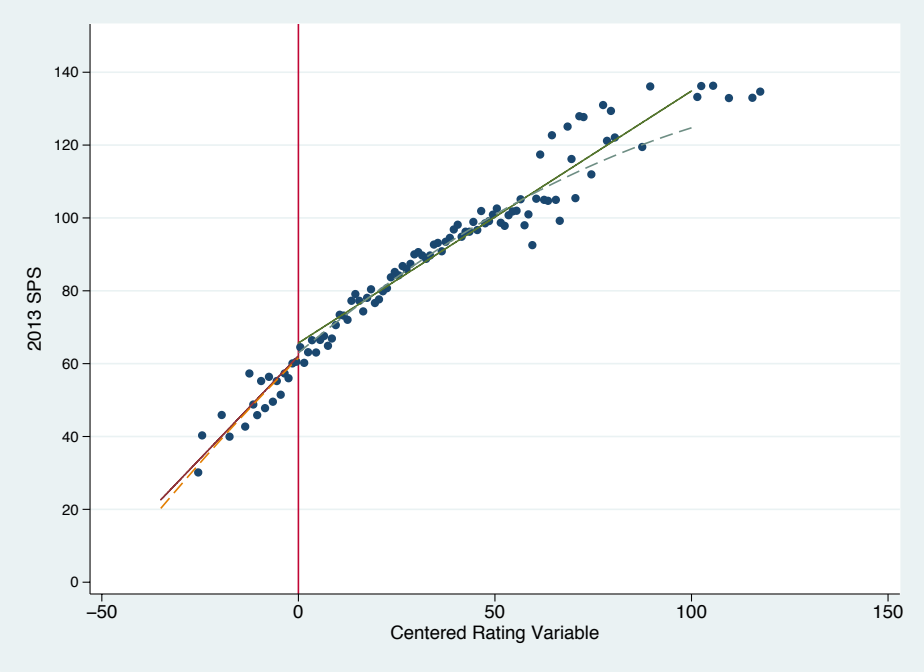


FIGURE 3 - FOCUS SCHOOL ASSIGNMENT

Notes: Graph reflects analytic sample, restricted to within -30 to +90 points of the intent to treat cut score. Bins are of size 3. Numbers above markers indicate the number of schools in the bin.

Panel A. 2013 School Performance Scores, full sample



Panel B. 2013 School Performance Scores, restricted bandwidth $S_i \leq 20$

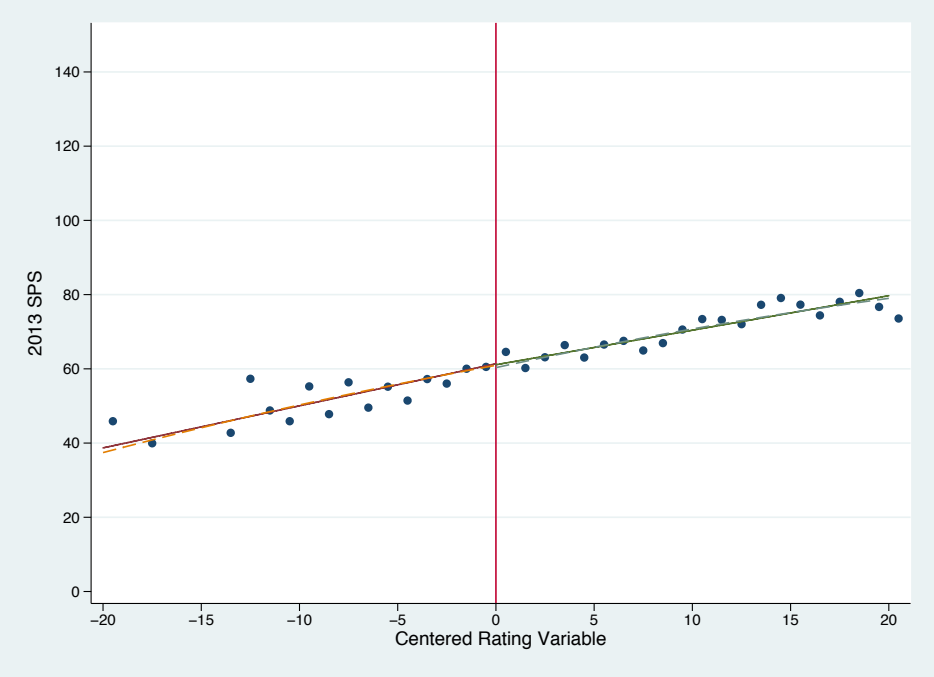
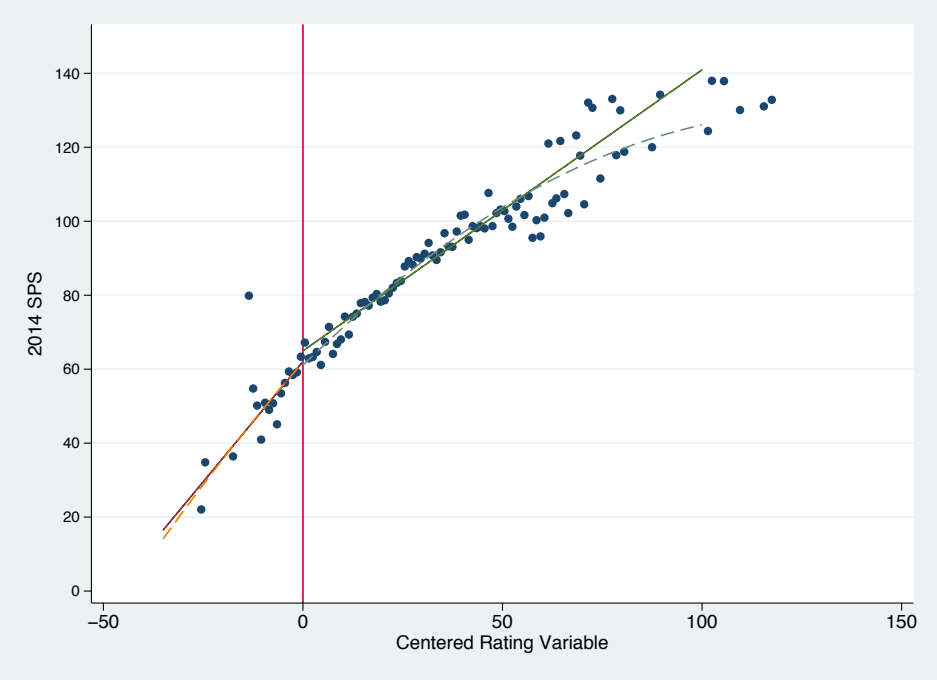


FIGURE 4 - 2013 SCHOOL PERFORMANCE SCORES
BY CENTERED RATING VARIABLE

Notes: In both panels, the solid line shows the fitted model with a linear spline; the dashed line shows the fitted model with a quadratic spline. Fitted models do not include school controls. Bins are of size 1.

Panel A. 2014 School Performance Scores, full sample



Panel B. 2014 School Performance Scores, restricted bandwidth $S_i \leq 20$

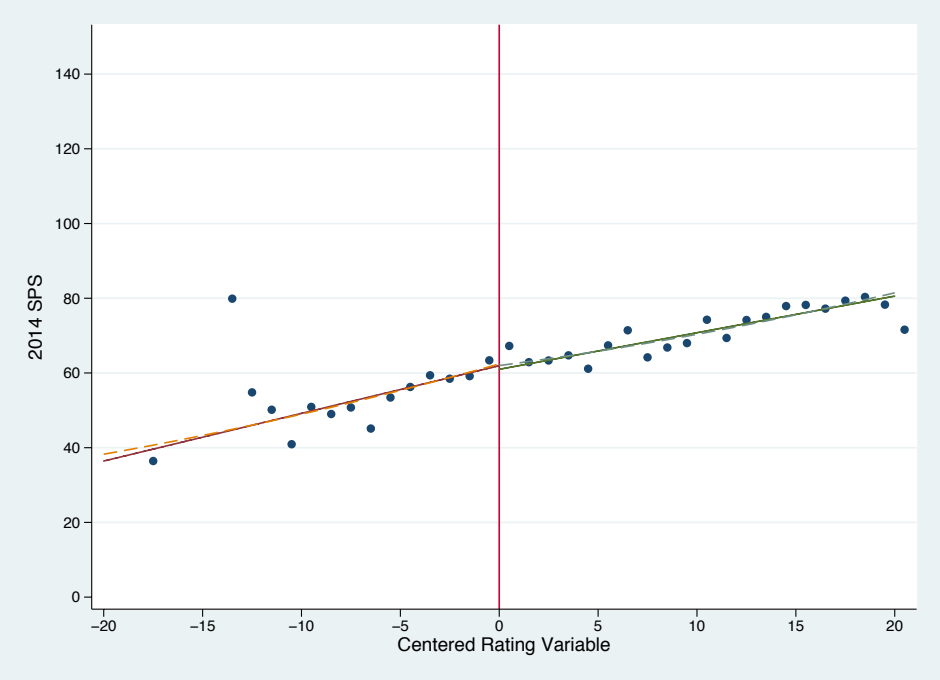
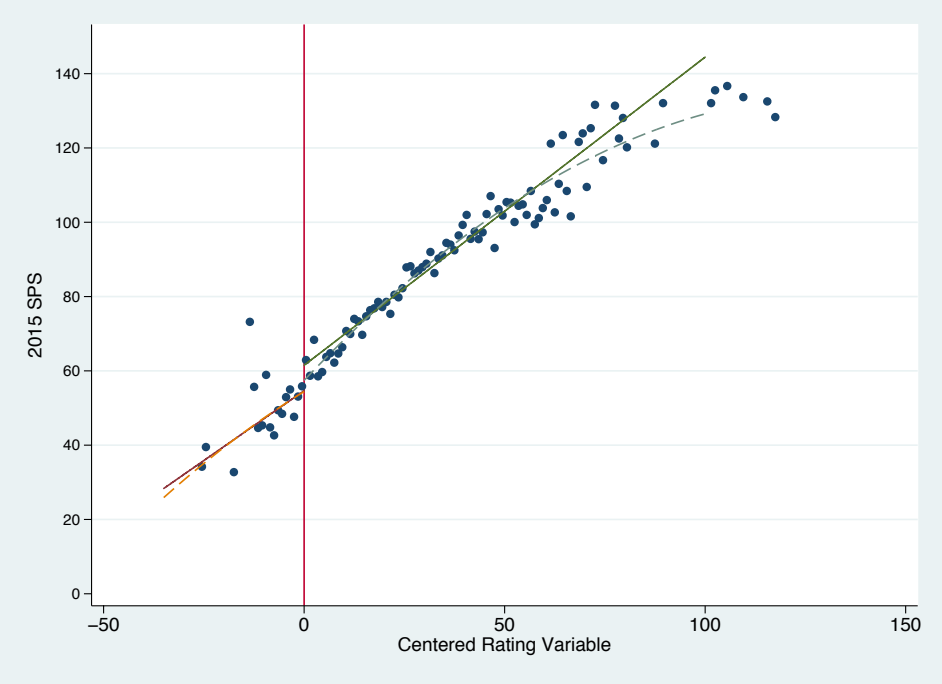


FIGURE 5 - 2014 SCHOOL PERFORMANCE SCORES
BY CENTERED RATING VARIABLE

Notes: In both panels, the solid line shows the fitted model with a linear spline; the dashed line shows the fitted model with a quadratic spline. Fitted models do not include school controls. Bins are of size 1.

Panel A. 2015 School Performance Scores, full sample



Panel B. 2015 School Performance Scores, restricted bandwidth $S_i \leq 20$

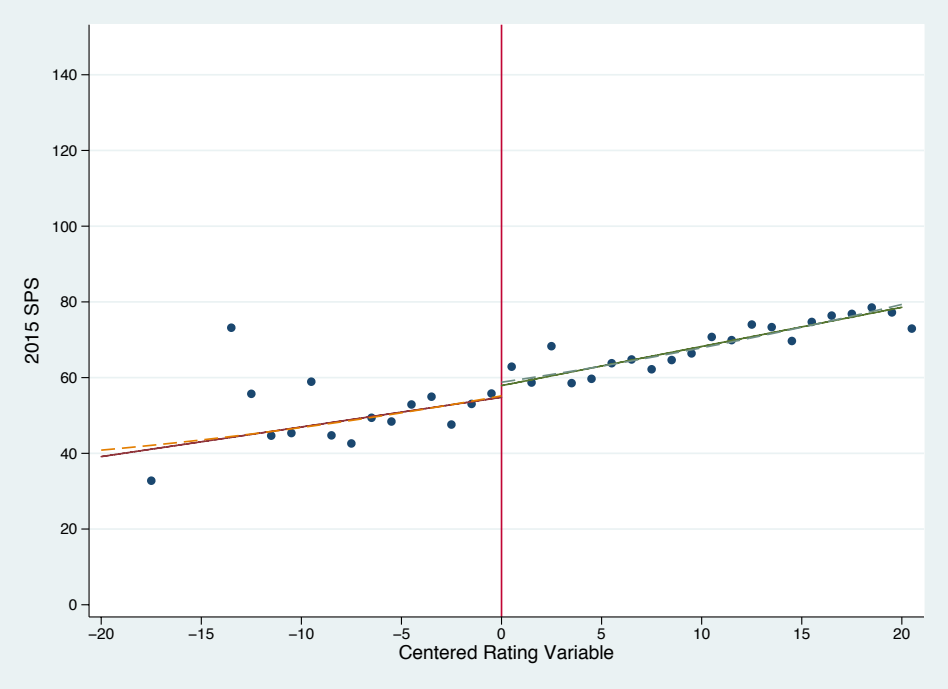


FIGURE 6 - 2015 SCHOOL PERFORMANCE SCORES
BY CENTERED RATING VARIABLE

Notes: In both panels, the solid line shows the fitted model with a linear spline; the dashed line shows the fitted model with a quadratic spline. Fitted models do not include school controls. Bins are of size 1.

TABLE 1 - DESCRIPTIVE STATISTICS FOR ANALYTICAL SAMPLE

	Mean	Std Dev	Min	Max	Count
Outcomes					
2012-13 SPS	82.45	17.75	28.5	136.3	1157
2013-14 SPS	83.60	19.02	21.2	138	1141
2014-15 SPS	82.01	20.03	21.7	136.7	1131
Baseline Characteristics					
2011-12 SPS (centered)	25.00	19.13	-25.8	117.7	1158
Focus School/F letter grade	0.08	0.27	0	1.00	1158
Title I-eligible	0.91	0.29	0	1.00	1158
Percent Black	0.43	0.31	0	1.00	1158
Percent Hispanic	0.04	0.06	0	0.53	1158
Percent Free/Reduced-Price Lunch	0.69	0.21	0.03	1.00	1158
Student-Teacher Ratio	15.27	2.69	4.16	36.50	1158
Elementary Schools	0.59	0.49	0	1.00	1158
Middle Schools	0.19	0.39	0	1.00	1158
High Schools	0.22	0.42	0	1.00	1158

Source: Louisiana State Department of Education data files (School performance scores and letter grades) and NCES Common Core of Data.

Notes: Baseline enrollment characteristics (Title I status, percent Black, percent Hispanic, percent FRPL, student-teacher ratio and school type) are from 2012-13 SY. The analytical sample is made up of traditional public schools; alternative and charter schools are excluded. Missingness of outcome test scores is balanced across the Focus threshold. One school is missing data in 2012-13 but is assigned SPS's again starting the following year. See text for more details on the analytical sample.

TABLE 2 - RD ESTIMATES OF TREATMENT EFFECT ON SCHOOL PERFORMANCE SCORES

	2013 School Performance Scores				2014 School Performance Scores				2015 School Performance Scores			
Independent Variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
I $(S_i \leq 0)$	-4.336*	-2.409	-0.737	-0.927	-2.883	-0.426	1.058	0.707	-6.836***	-3.918*	-2.864	-3.495
	(1.718)	(1.614)	(2.443)	(2.209)	(2.034)	(2.005)	(2.973)	(2.803)	(2.039)	(1.847)	(2.981)	(2.570)
School Controls	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Linear Spline	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Quadratic Spline	No	No	Yes	Yes	No	No	Yes	Yes	No	No	Yes	Yes
AIC	8466.1	8112.9	8437.5	8111.3	8737.7	8326.8	8698.3	8322.3	8656.2	8396.8	8614.3	8394.2
Observations	1157	1157	1157	1157	1141	1141	1141	1141	1131	1131	1131	1131

Robust standard errors in parentheses

* p<0.05 ** p<0.01 *** p<0.001

Notes: The analytical sample is made up of traditional public schools; alternative and charter schools are excluded. School controls are the student-teacher ratio, the percentages of Black, Hispanic, and FRPL students in the school in the 2012-13 school year, and the school type (Elementary/Middle/High). The 2013 School Performance Scores are based on 2012-13 school outcomes and were reported in October of 2013. The 2014 SPSs are based on the 2013-14 school outcomes and were reported in October 2014. AIC refers to Akaike's Information Criteria.

TABLE 3 - RD ESTIMATES USING ALTERNATIVE BANDWIDTHS

	2013 School Performance Scores			2014 School Performance Scores			2015 School Performance Scores		
	(1)	(2)	*n*	(3)	(4)	*n*	(5)	(6)	*n*
Full-sample	-4.336*	-2.409	1157	-2.883	-0.426	1141	-6.836***	-3.918*	1131
	(1.718)	(1.614)		(2.034)	(2.005)		(2.039)	(1.847)	
$\|S_i\| \leq 30$	-0.916	-1.039	715	1.307	1.049	700	-3.205	-2.747	692
	(1.811)	(1.667)		(2.156)	(2.060)		(2.177)	(1.920)	
$\|S_i\| \leq 20$	-0.513	-0.272	469	0.971	1.072	457	-3.090	-2.623	449
	(2.059)	(1.819)		(2.553)	(2.324)		(2.708)	(2.401)	
$\|S_i\| \leq 10$	-0.227	1.273	259	0.0499	1.443	255	-5.049	-3.537	250
	(2.884)	(2.484)		(3.398)	(3.191)		(3.528)	(3.082)	
$\|S_i\| \leq 8$	-0.667	0.0933	201	0.128	0.480	198	-5.228	-4.793	194
	(3.063)	(2.623)		(3.645)	(3.372)		(3.963)	(3.531)	
Kernel-weighted	0.357	0.980	452	0.744	1.285	441	-3.964	-3.287	433
	(2.366)	(2.004)		(2.889)	(2.618)		(3.009)	(2.564)	
School controls	No	Yes		No	Yes		No	Yes	

Robust standard errors in parentheses

* p<0.05 ** p<0.01 *** p<0.001

Notes: All models use a linear spline specification. Full sample is the analytical sample, made up of traditional public schools; alternative and charter schools are excluded.. School controls are the student-teacher ratio, the percentages of Black, Hispanic, and FRPL students in the school in the 2012-13 school year, and the school type (Elementary/Middle/High). Kernel-weighted estimates use a triangular kernel weighting. The 2013 School Performance Scores are based on 2012-13 school outcomes and were reported in October of 2013.The 2014 SPSs are based on the 2013-14 school outcomes and were reported in October 2014.

TABLE 4 - RD ESTIMATES FOR HETEROGENEOUS TREATMENT EFFECTS BY SCHOOL LEVEL

	2013 School Performance Scores			2014 School Performance Scores			2015 School Performance Scores		
	(1)	(2)	*n*	(3)	(4)	*n*	(5)	(6)	*n*
Primary Schools	-2.543	-1.178	681	-0.238	1.760	668	-3.978	-1.381	660
	(2.178)	(1.987)		(2.595)	(2.521)		(2.539)	(2.239)	
Middle Schools	-1.370	1.377	217	-0.864	1.450	214	-5.102	-2.117	212
	(2.414)	(2.243)		(3.113)	(2.973)		(4.590)	(4.534)	
High Schools	-1.206	1.709	259	-4.509**	-0.292	259	-7.162	-5.005	259
	(2.851)	(3.011)		(1.386)	(1.744)		(4.829)	(4.735)	
School controls	No	Yes		No	Yes		No	Yes	

Robust standard errors in parentheses

* p<0.05 ** p<0.01 *** p<0.001

Notes: All models use a linear spline specification. Relative to other full sample specifications (quadratic and without controls), this specification yielded the lowest Akaike's information criteria for seven out of the nine models run. School controls are the student-teacher ratio and the percentages of Black, Hispanic, and FRPL students in the school in the 2012-13 school year. The 2013 School Performance Scores are based on 2012-13 school outcomes and were reported in October of 2013. The 2014 SPSs are based on the 2013-14 school outcomes and were reported in October 2014.

TABLE 5 - RD ESTIMATES OF EFFECT ON 2013 AND 2014 SUBJECT-LEVEL LEAP/ILEAP TEST SCORE OUTCOMES, GRADES 3-8

	Panel A: 2013 Scores					
	Full Sample			$\|S_i\| \leq 20$		
	Estimate	Mean (Std Dev)	*n*	Estimate	Mean (Std Dev)	*n*
Math	-2.583	69.73	3179	-1.890	58.96	1360
	(2.022)	(17.71)		(2.316)	(16.29)	
ELA	-1.858	71.73	3179	-0.684	60.66	1360
	(1.419)	(16.46)		(1.616)	(14.34)	
Science	-2.084	65.36	3178	0.418	51.03	1360
	(1.371)	(19.05)		(1.617)	(15.67)	
Social Studies	-0.681	66.59	3178	1.715	53.42	1360
	(1.700)	(18.91)		(1.988)	(16.18)	
	Panel B: 2014 Scores					
	Full Sample			$\|S_i\| \leq 20$		
	Estimate	Mean (Std Dev)	*n*	Estimate	Mean (Std Dev)	*n*
Math	-0.228	71.55	3139	-0.0404	61.21	1311
	(1.925)	(17.70)		(2.322)	(16.05)	
ELA	-0.777	70.37	3139	0.300	59.62	1311
	(1.458)	(16.54)		(1.720)	(14.18)	
Science	-0.875	65.80	3138	0.766	51.83	1311
	(1.410)	(19.14)		(1.792)	(15.35)	
Social Studies	0.755	67.42	3137	1.459	55.12	1311
	(1.580)	(18.52)		(1.917)	(15.71)	

Standard errors in parentheses are clustered at the school level
* p<0.05 ** p<0.01 *** p<0.001

Notes: All models use a linear spline specification and include grade fixed effects and school controls (student-teacher ratios, percentages of Black, Hispanic, and FRPL students in the 2012-13 school year, and school type). Dependent variables are the subject-level percent proficient on the 2013 or 2014 Louisiana Educational Assessment Program (LEAP) tests, grades 4 and 8, and the 2013 or 2014 Integrated Louisiana Educational Assessment Program (iLEAP) tests, grades 3,5,6,7.

APPENDIX I - SCHOOL GRADE LEVEL CATEGORIES

Louisiana has a large number of schools with a range of grade levels that do not fit into typical definitions of school levels (elementary, middle and high schools). Additionally, the school levels defined in the Louisiana Department of Education (LDOE) data are not always consistent with the levels in the NCES Common Core of Data (CCD). Interestingly enough, even within each dataset, the level categorizations have inconsistences; for example, in both the LDOE data and the CCD data, some grade 8-12 schools are labeled high schools, while others are labeled Combination Schools. To maintain consistency, we use the CCD categorizations to determine whether a school was characterized as elementary, middle, or high—with some key exceptions.

Elementary and Middle Schools are categorized based on CCD data. We manually adjust the following:

- 7 schools with grades 4-5 were marked in CCD as middle; we changed to elementary.
- 1 school with grades 4-6, titled Breaux Bridge Elementary School, was marked in CCD as middle; we changed to elementary. All other grade 4-6 schools were kept as middle, consistent with CCD.

To determine distinctions between high school and combination schools, we examined the most common unusual grade ranges in Louisiana schools. A large number of high schools start as early as 8th grade. We keep all these grade 8-12 schools categorized as high schools. However, to be consistent with the typical definitions of elementary/middle/high, we categorize any school that starts in 7th grade or below and goes through 12th grade as a combination school. This includes 36 grade 7-12 schools that were originally marked in CCD as high schools. Schools with grades 8-9 or 9 only are considered high schools.

With these rules, there are 113 combination schools (all these schools are either a conventional school or a magnet school):[15]

- Grades 7-12: 36
- Grades 6-12: 19
- Grades 5-12: 2
- Grades 4-12: 4
- Grades KG-12: 2
- Grades PK-12: 50

Finally, for simplicity, we incorporate the combination category into the more traditional elementary/middle/high categories. Because the majority of combination schools actually have the term “high school” in their name (including 44 of the 50 PK-12 schools and both of the KG-12 schools), we re-categorize all combination schools as high schools.

Count with combination schools:

Level	Frequency
Elementary	681
Middle	217
High	147
Combination	113

Count without combination schools:

[15] We manually checked through online searches all schools listed as “magnet/alternative” in the CCD to confirm that they were magnet schools. Alternative schools are not included in our analytic sample.

Level	Frequency
Elementary	681
Middle	217
High	260

Throughout the analyses we report, we use the school level variable without combination schools as a control variable in our models. We also use this variable for our regression models examining the heterogeneous effects of Focus School assignment across school type (Table 4). When we estimate our models for our main results using instead the school level variable that defines combination school as a separate category, our results do not change. We do not estimate the heterogeneous effects model with the school level variable that lists combination schools separately because the number of Focus high schools (n=1) and Focus combination schools (n=4) are too small for separate RD analyses to be meaningful.

APPENDIX II - TESTING FOR COVARIATE BALANCE

To test for covariate balance, we estimate our regression discontinuity model, making the dependent variable one of the set of continuous baseline covariates we include in our main estimations, namely the percent of economically disadvantaged students (free and reduced-price lunch eligible), the percent black, the percent Hispanic, and the student-teacher ratio. We do this for variables both from the 2012-13 year—the year that the Focus School assignments were announced—and the 2011-12 year—the year that determined the Focus School assignments. Our results are shown in Table A1.

For perfectly balanced covariates, we would hope that the estimated jump coefficient would be statistically insignificant for all specifications. However, we do find estimated coefficients from some models and specifications that are statistically significant, namely the full sample linear model for percent black and Hispanic (both 2011-12 and 2012-13), the full sample quadratic for FRPL percent (both 2011-12 and 2012-13), the 30-point bandwidth specification for Hispanic percent (2011-12), and the 10-point bandwidth specification for FRPL percent (2012-13).

We look to the graphical representations of the covariate data to gauge the fit of the various models to the data and assess whether or not the regression models are picking up true imbalances in the observable school characteristics. In large part because of the high degree of racial and income-based segregation in Louisiana schools, the distribution of the covariates across schools lead to unusual functional forms. As such, a simple visual analysis of the data indicates that the full sample, linear specification of the RD estimate is a poor fit for the data. Figures A1 and A2 show the distribution of percent black and percent FRPL students by the centered rating variable and help illustrate this point (Hispanic percent is a similarly non-linear functional form, though we direct less attention to it because the average percent of Hispanic students in our schools is very low at 4 percent).

Even after ruling out the estimates from the full sample linear models, the remaining significant coefficients may be cause for concern or may be a result of a multiple-comparisons problem. To further interrogate whether these results reflect true imbalances that would affect the appropriateness of a regression discontinuity model for our data, we estimate a composite variable, made up of a weighted average of all the covariates in our model where the weights indicate the extent to which the covariate predicts the outcome. In practice, we create this composite variable, which we call the "achievement index," by regressing the outcome variable

(SPS) on the baseline covariates and then predict the outcome. We thus are able to calculate achievement indices for the 2013 SPS, 2014 SPS, and 2015 SPS. Table A2 shows the RD estimates for these achievement indices across multiple specifications. The achievement indices are shown graphically in Figures A3, A4, and A5.

As Table A2 shows, the estimated jump coefficients for the achievement index is insignificant for all specifications except the full sample linear model. Similar to the individual covariates, a visual inspection of the data suggests that the odd functional form of the achievement index makes a full sample linear model a poor fit to the data. This is confirmed by the Akaike's information criteria (shown in Table A2), which is smaller for the full sample quadratic models than the linear models, indicating that the quadratic is a better fit to the data. Using a quadratic model or limiting the data to a smaller bandwidth results in insignificant coefficients. The results for the achievement indices, in combination with the individual covariate models, leads us to conclude that there are not significant imbalances in the observable covariates and that the regression discontinuity model is appropriate for our data.

APPENDIX III - TESTING FOR CHANGES IN POST-TREATMENT STUDENT ENROLLMENT

To test for evidence that the student population changed in Focus Schools as a result of being identified as a Focus School, or more likely, as a result of receiving an F grade, we estimate regression discontinuity models with post-treatment school characteristics as the dependent variable. The question of changing student populations is particularly salient given that part of the treatment for failing schools is the offering of school choice to families attending those schools. Testing for discontinuities in post-treatment school characteristics provides empirical evidence on whether policy-endogenous student mobility constitutes an internal-validity threat. We run regression discontinuity models for 2013-14 student enrollment, percent free and reduced-price eligible, percent black, percent Hispanic, percent white, and student-teacher ratio to see whether there were significant changes in these school-level characteristics the year after the set of Focus Schools was announced. We also run models for 2014-15 student enrollment and percent economically disadvantaged, the data for which comes from the LDOE website. At the time of writing, the data for racial percentages and student teacher ratio, which come from the Common Core of Data, are not available for the 2014-15 school year. The results are shown in Table A3.

The results largely suggest that in 2013-14, the student population did not significantly change as a result of Focus School assignment. We find significant point estimates in models for percent FRPL, percent black, percent Hispanic, and percent white, but only for full sample models. However, the covariate distributions tend to be closer to a quadratic functional form and, in the full sample, we only find that the intent to treat appeared to influence percent FRPL and percent Hispanic. Moreover, the sign on these point estimates are inconsistent with the mobility patterns that might be a source of concern. In particular, they suggest that a Focus School and F designation *reduced* the share of economically disadvantaged children in a school. These point estimates also become smaller and statistically insignificant in local linear regressions based on tighter bandwidths of data. For the available 2015 school characteristics, the results are largely similar. There are several statistically significant estimates for percent FRPL. However, the signs of the estimated change in student population are in opposite directions, indicating that any estimated effect is not robust to changes in the model specification. A visual analysis of the 2015 percent FRPL data also does not indicate any consistent or clear discontinuity at the threshold.

To further interrogate any post-treatment student population imbalances, we create "achievement indices" much like we do for pre-treatment covariate imbalances by regressing 2014 SPS and 2015 SPS on the 2013-14 school characteristic variables and predicting the outcome. For the 2015 SPS models, we also include available 2014-15 enrollment variables. The estimated jump coefficient on these modified achievement indices is significant only for the full sample linear model, which is the worst fit to the data according to Akaike's information criteria, and insignificant for all other models (quadratic and restricted bandwidth), supporting our conclusion that policy-endogenous student mobility is not a threat to internal validity.

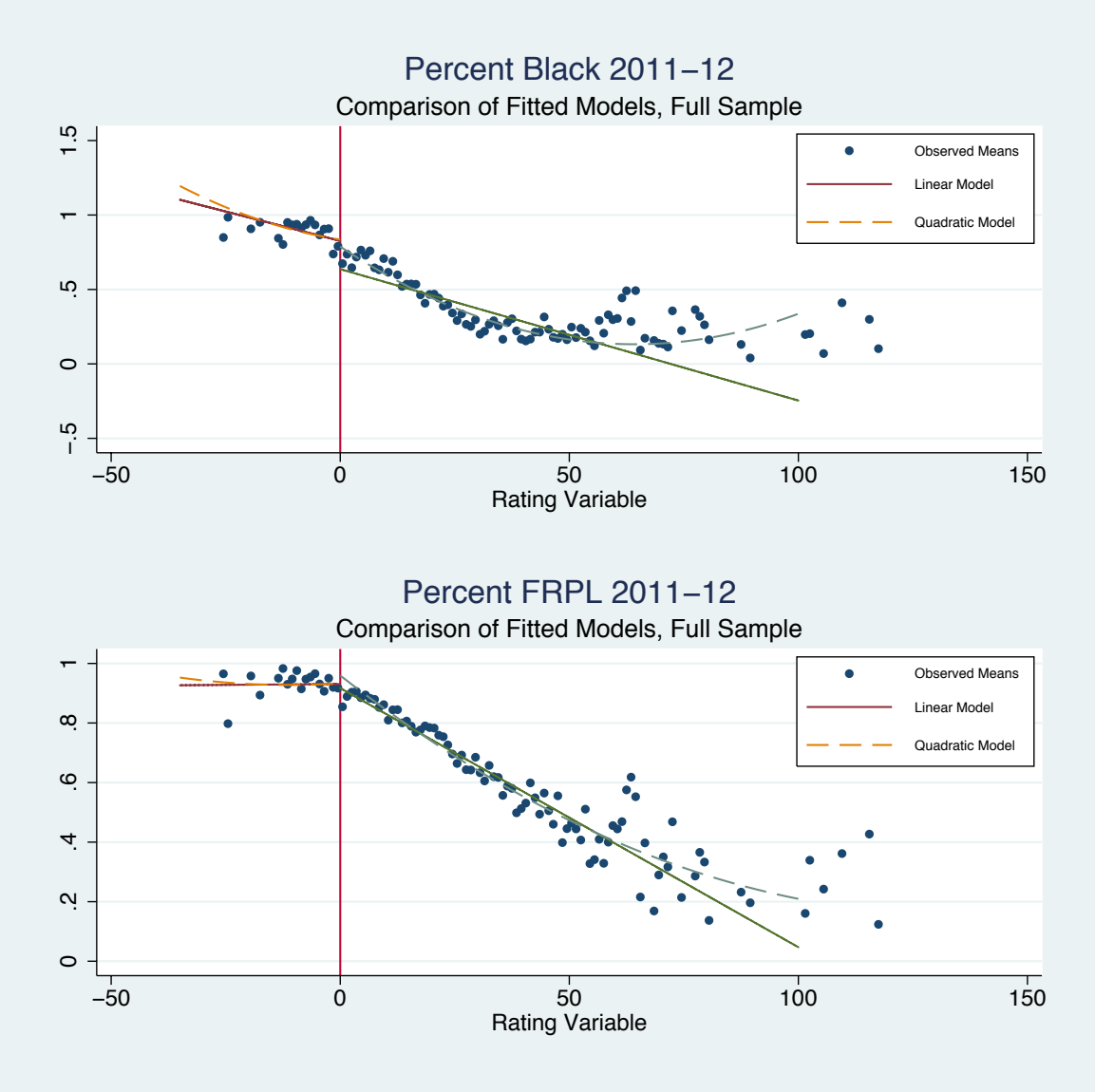


FIGURE A1. 2011-12 BASELINE SCHOOL CHARACTERISTICS
BY CENTERED RATING VARIABLE

Notes: The solid line shows the fitted model with a linear spline; the dashed line shows the fitted model with a quadratic spline. Bins are of size 1.

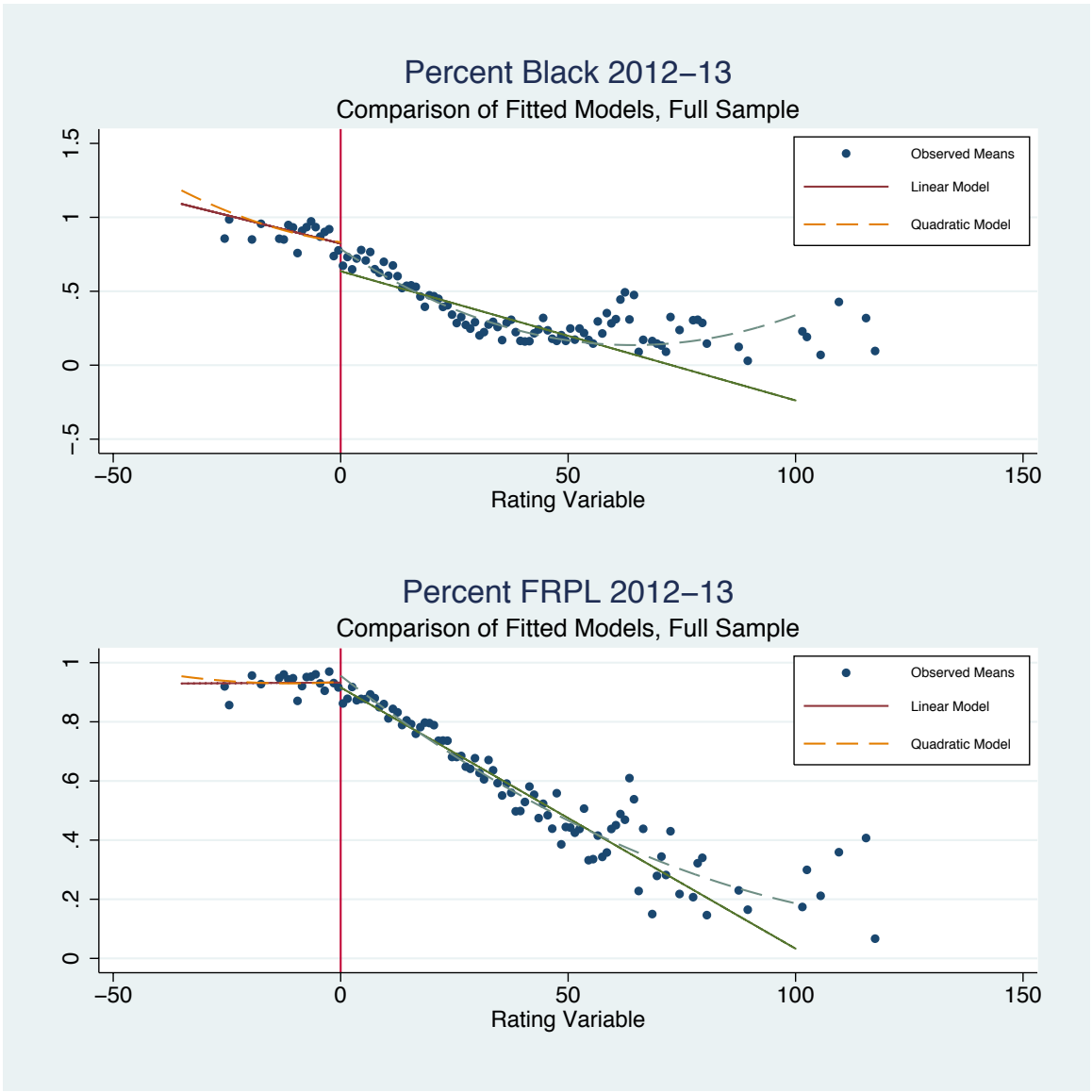
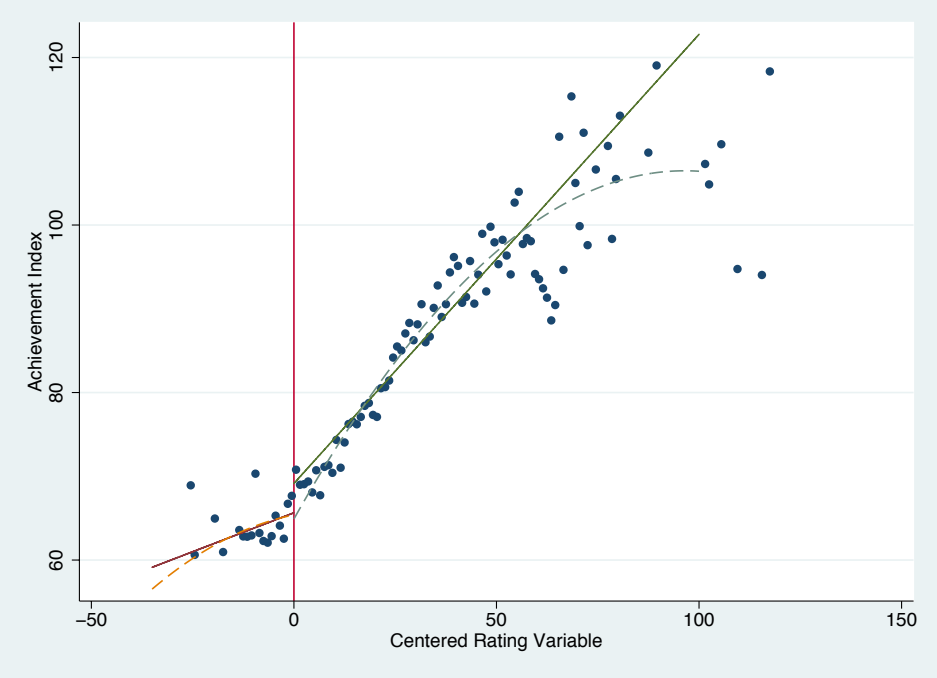


FIGURE A2. 2012-13 BASELINE SCHOOL CHARACTERISTICS
BY CENTERED RATING VARIABLE

Notes: The solid line shows the fitted model with a linear spline; the dashed line shows the fitted model with a quadratic spline. Bins are of size 1.

Panel A. Achievement index for 2013 School Performance Scores, full sample



Panel B. Achievement index for 2013 School Performance Scores, restricted bandwidth $S_i \leq 20$

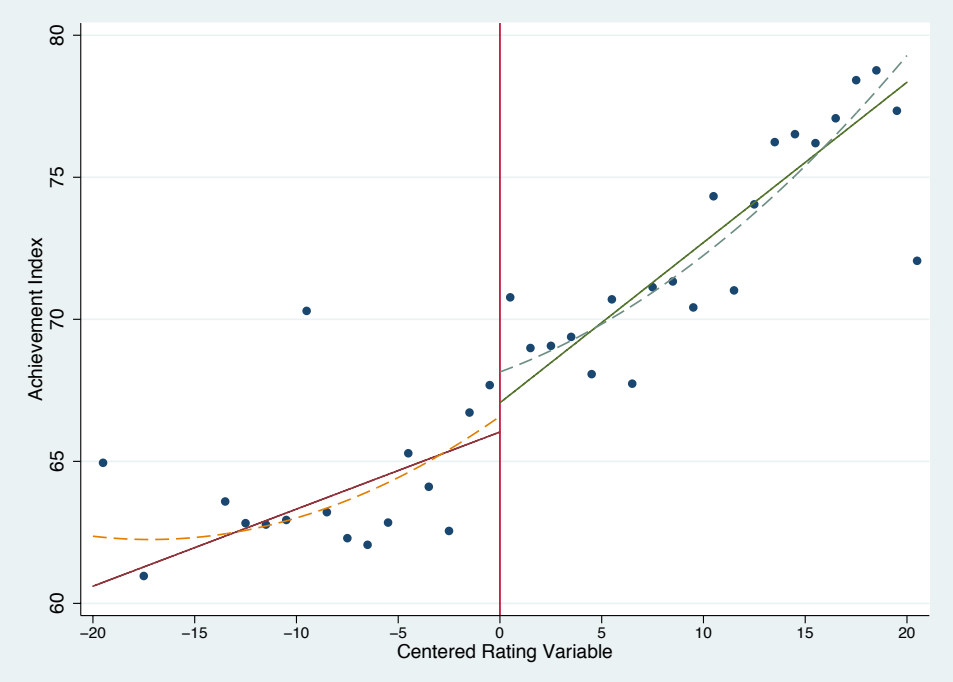
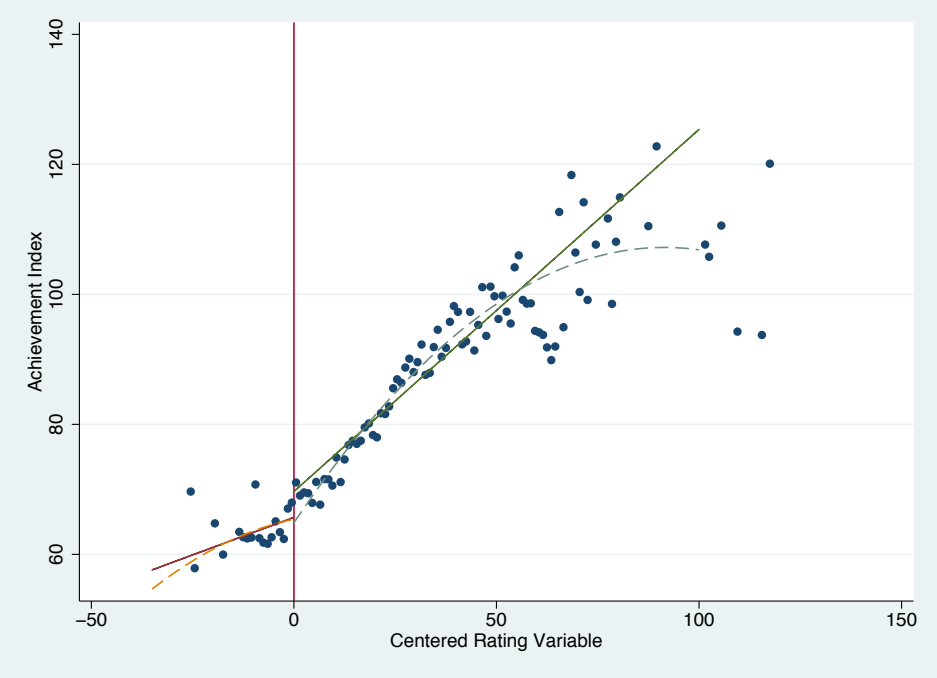


FIGURE A3. ACHIEVEMENT INDEX FOR 2013 SCHOOL PERFORMANCE SCORES BY CENTERED RATING VARIABLE

Notes: Bins are of size 1.

Panel A. Achievement index for 2014 School Performance Scores, full sample



Panel B. Achievement index for 2014 School Performance Scores, restricted bandwidth $S_i \leq 20$

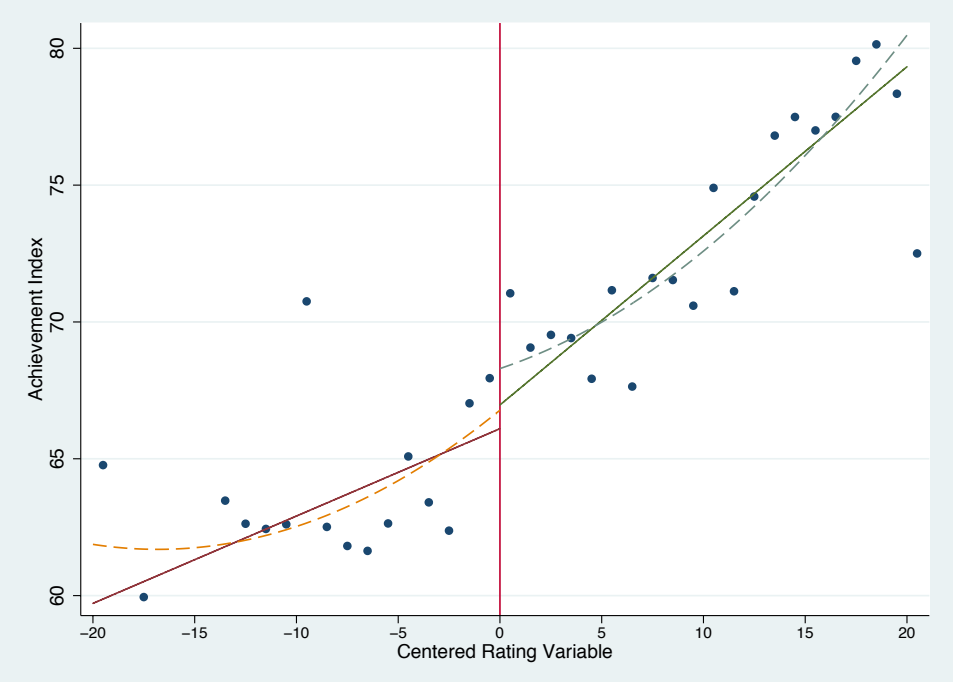
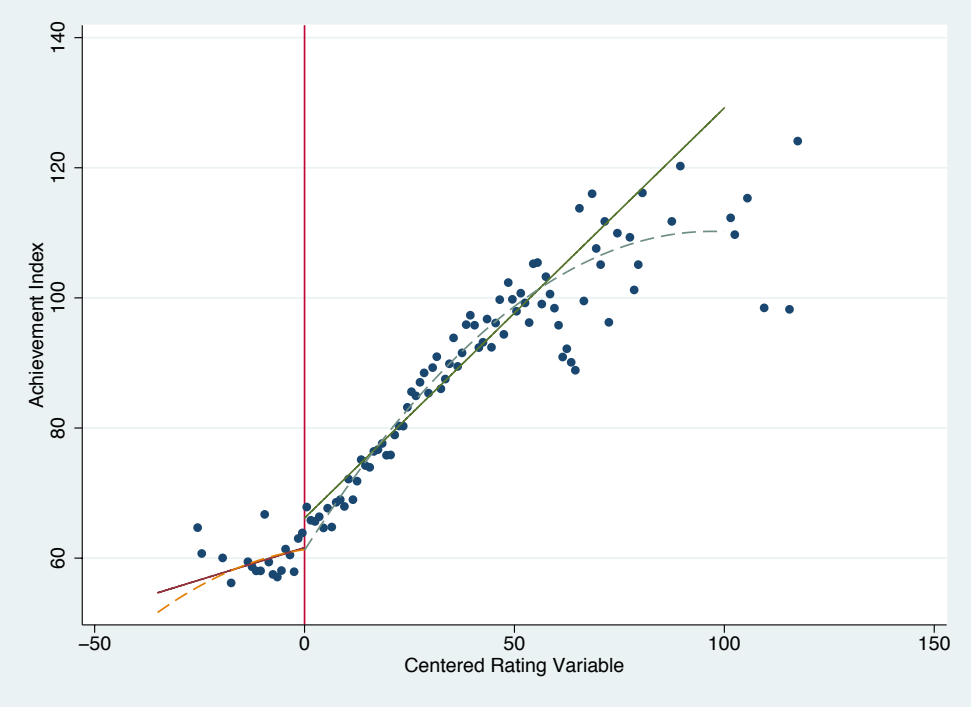


FIGURE A4. ACHIEVEMENT INDEX FOR 2014 SCHOOL PERFORMANCE SCORES BY CENTERED RATING VARIABLE

Notes: Bins are of size 1.

Panel A. Achievement index for 2015 School Performance Scores, full sample



Panel B. Achievement index for 2015 School Performance Scores, restricted bandwidth $S_i \leq 20$

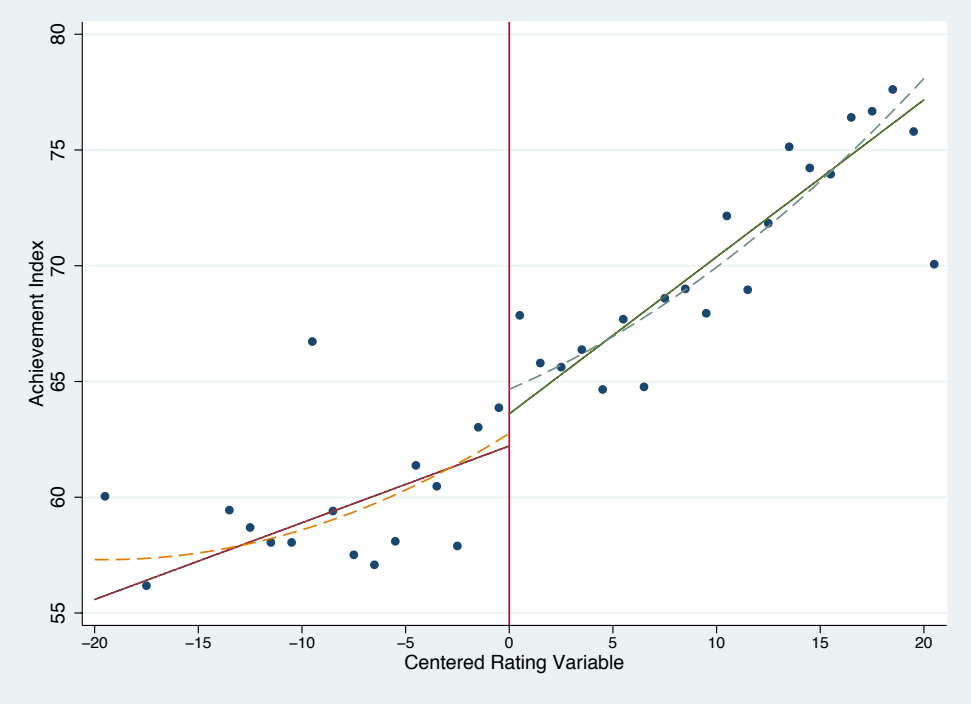


FIGURE A5. ACHIEVEMENT INDEX FOR 2015 SCHOOL PERFORMANCE SCORES BY CENTERED RATING VARIABLE

Notes: Bins are of size 1.

TABLE A1 - AUXILIARY RD ESTIMATE: COVARIATE BALANCE CHECK

	Panel A: 2011-12 Baseline Covariates				
	FRPL percent	Black percent	Hispanic percent	Student-teacher ratio	*n*
Linear Spline	0.0136	0.190***	-0.0200*	0.197	1158
	(0.0122)	(0.0326)	(0.00802)	(0.431)	
Quadratic spline	-0.0428**	-0.0100	-0.0203	0.0918	1158
	(0.0145)	(0.0450)	(0.0108)	(0.515)	
$\|S_i\| \leq 30$, linear spline	-0.000920	0.0234	-0.0187*	0.319	716
	(0.0126)	(0.0341)	(0.00904)	(0.448)	
$\|S_i\| \leq 20$, linear spline	0.0129	0.0339	-0.0146	-0.0944	469
	(0.0137)	(0.0395)	(0.0105)	(0.434)	
$\|S_i\| \leq 10$, linear spline	0.0200	0.0535	-0.0162	0.207	259
	(0.0186)	(0.0533)	(0.0143)	(0.540)	
$\|S_i\| \leq 8$, linear spline	0.0221	0.0302	-0.0133	0.600	201
	(0.0193)	(0.0612)	(0.0170)	(0.585)	

	Panel B: 2012-13 Baseline Covariates				
	FRPL percent	Black percent	Hispanic percent	Student-teacher ratio	*n*
Linear Spline	0.0167	0.187***	-0.0207*	-0.677	1158
	(0.0116)	(0.0331)	(0.0103)	(0.583)	
Quadratic spline	-0.0322*	-0.00871	-0.0150	-0.548	1158
	(0.0146)	(0.0459)	(0.0143)	(0.663)	
$\|S_i\| \leq 30$, linear spline	0.00595	0.0205	-0.0182	-0.614	716
	(0.0125)	(0.0347)	(0.0113)	(0.601)	
$\|S_i\| \leq 20$, linear spline	0.0238	0.0320	-0.0146	-1.181	469
	(0.0144)	(0.0402)	(0.0130)	(0.719)	
$\|S_i\| \leq 10$, linear spline	0.0404*	0.0586	-0.0149	-0.537	259
	(0.0200)	(0.0551)	(0.0178)	(0.666)	
$\|S_i\| \leq 8$, linear spline	0.0387	0.0243	-0.00894	-0.574	201
	(0.0205)	(0.0620)	(0.0212)	(0.764)	

Robust standard errors in parentheses
* p<0.05 ** p<0.01 *** p<0.001

Notes: Dependent variables are continuous baseline covariate variables included in our models, namely the student-teacher ratio, percent Black, percent Hispanic, and percent FRPL from the 2011-12 SY and 2012-13 SY). 2011-12 SY charactertistics correspond to the year the rating variable was determined. 2012-13 SY characteristics correspond to the year Focus school assignment was announced and interventions began.

TABLE A2 - AUXILIARY RD ESTIMATE: ACHIEVEMENT INDEX BALANCE CHECK

	2013 Estimate	*n*	2014 Estimate	*n*	2015 Estimate	*n*
Linear Spline	-3.467***	1158	-3.963***	1158	-4.544***	1158
	(1.024)		(1.158)		(1.188)	
Quadratic spline	1.996	1158	2.172	1158	2.135	1158
	(1.332)		(1.505)		(1.546)	
$\|S_i\| \leq 30$, linear spline	-0.0667	716	0.131	716	-0.640	716
	(1.076)		(1.214)		(1.249)	
$\|S_i\| \leq 20$, linear spline	-1.018	469	-0.904	469	-1.387	469
	(1.207)		(1.353)		(1.413)	
$\|S_i\| \leq 10$, linear spline	-2.295	259	-2.254	259	-2.717	259
	(1.689)		(1.887)		(1.995)	
$\|S_i\| \leq 8$, linear spline	-1.483	201	-1.296	201	-1.692	201
	(1.822)		(2.044)		(2.157)	
Linear model AIC	8611		8393		8621	
Quadratic model AIC	8531		8317		8534	

Robust standard errors in parentheses
* p<0.05 ** p<0.01 *** p<0.001

Notes: Dependent variables are the predicted values from the regressions of the 2013, 2014 or 2015 School Performance Scores on the baseline covariates included in models (student-teacher ratio, percent Black, percent Hispanic, percent FRPL, and school type from the 2012-13 SY). In other words, dependent variables are the average of the covariates, weighted by their association with the main outcomes of interest. AIC refers to Akaike's information criteria, for which a smaller value indicates a better fit of the model to the data when using the full sample of observations.

TABLE A3: AUXILIARY RD ESTIMATE: POST-TREATMENT ENROLLMENT AND SCHOOL CHARACTERISTICS

	Panel A: 2014 School Characteristics						
	Enrollment	% FRPL	% Black	% Hispanic	% White	Student-teacher ratio	*n*
Linear spline	33.22	0.00808	0.179***	-0.0599*	-0.184***	0.199	1137
	(42.26)	(0.0119)	(0.0332)	(0.0272)	(0.0367)	(0.576)	
Quadratic spline	41.39	-0.0455**	-0.0104	-0.0611*	0.000638	-0.129	1137
	(51.24)	(0.0146)	(0.0459)	(0.0282)	(0.0441)	(0.653)	
$\|S_i\| \leq 30$, linear spline	0.383	-0.00373	0.0194	-0.0523	-0.0282	0.179	697
	(42.84)	(0.0124)	(0.0349)	(0.0293)	(0.0379)	(0.605)	
$\|S_i\| \leq 20$, linear spline	13.00	0.0139	0.0364	-0.0565	-0.0593	-0.114	454
	(41.44)	(0.0138)	(0.0402)	(0.0378)	(0.0478)	(0.681)	
$\|S_i\| \leq 10$, linear spline	-48.18	0.0269	0.0696	-0.0199	-0.0501	-0.246	254
	(55.98)	(0.0190)	(0.0550)	(0.0291)	(0.0504)	(0.694)	
$\|S_i\| \leq 8$, linear spline	-30.36	0.0190	0.0345	-0.0258	-0.0396	-0.0716	197
	(63.73)	(0.0200)	(0.0619)	(0.0343)	(0.0571)	(0.818)	

	Panel B: 2015 School Characteristics						
	Enrollment	% FRPL					*n*
Linear spline	0.919	0.0263**	n/a	n/a	n/a	n/a	1127
	(37.80)	(0.00961)					
Quadratic spline	-7.777	-0.0258*	n/a	n/a	n/a	n/a	1127
	(48.33)	(0.0116)					
$\|S_i\| \leq 30$, linear spline	-32.56	0.00184	n/a	n/a	n/a	n/a	689
	(38.31)	(0.00992)					
$\|S_i\| \leq 20$, linear spline	-37.02	0.0180	n/a	n/a	n/a	n/a	446
	(41.72)	(0.0112)					
$\|S_i\| \leq 10$, linear spline	-54.68	0.0313*	n/a	n/a	n/a	n/a	249
	(56.27)	(0.0152)					
$\|S_i\| \leq 8$, linear spline	-54.54	0.0296	n/a	n/a	n/a	n/a	193
	(64.56)	(0.0159)					

Robust standard errors in parentheses

* p<0.05 ** p<0.01 *** p<0.001

Notes: Quadratic spline model uses the full analytic sample. Enrollment numbers include all public primary and secondary students at each school site. No controls are included. 2014 data comes from the Common Core of Data. 2015 data comes from LDOE enrollment data. 2015 racial percentages and student teacher ratio data are not available at the time of writing.