NATIONAL
CENTER *for* ANALYSIS *of* LONGITUDINAL DATA *in* EDUCATION RESEARCH

TRACKING EVERY STUDENT'S LEARNING EVERY YEAR

A program of research by the American Institutes for Research with Duke University, Northwestern University, Stanford University, University of Missouri-Columbia, University of Texas at Dallas, and University of Washington

AIR

Accountability Pressure and Non-Achievement Student Behaviors

JOHN B. HOLBEIN
HELEN F. LADD

WORKING PAPER 122 • FEBRUARY 2015

Accountability Pressure and Non-Achievement Student Behaviors

John B. Holbein
Duke University

Helen F. Ladd
Duke University

Contents

Acknowledgements

We wish to thank the National Science Foundation (#SES-1416816) and the Center for the Analysis of Longitudinal Data in Education Research (#R305C120008) funded by the U.S. Department of Education for providing funding that supported this work. We also would like to give special thanks to Tom Ahn, Nicholas Carnes, Andrew Heiss, Sunshine Hillygus, Jeffrey Traczynski and Jacob Vigdor for their thoughtful comments and assistance throughout this project.

CALDER • American Institutes for Research
1000 Thomas Jefferson Street N.W., Washington, D.C. 20007
202-403-5796 • www.caldercenter.org

Accountability Pressure and Non-Achievement Student Behaviors
John B. Holbein and Helen F. Ladd
Working Paper 122
February 2015

Abstract

In this paper we examine how failing to make adequate yearly progress under No Child Left Behind (NCLB), and the accountability pressure that ensues, affects various non-achievement student behaviors. Using administrative data from North Carolina and leveraging a discontinuity in the determination of school failure, we examine the causal impact of accountability pressure both on student behaviors that are incentivized by NCLB and on those that are not. We find evidence that, as NCLB intends, pressure encourages students to show up at school and to do so on time. Accountability pressure also has the unintended effect, however, of increasing the number of student misbehaviors such as suspensions, fights, and offenses reportable to law enforcement. Further, this negative response is most pronounced among minorities and low performing students, who are the most likely to be left behind.

"We are transforming our schools ... We are insisting on accountability ... We will leave no child behind." –President George W. Bush[1]

"Stress, stress, stress. [High-stakes testing] … leads to student[s] becoming so emotional and anxious that they don't even have the ability to function properly." – Julia Skinner-Grant[2]

In recent years, education policy has moved distinctively towards a system of performance-based accountability as a primary means of improving student outcomes. This approach—which places pressure on schools by measuring, publishing, and incentivizing their performance—has been integral to both federal and state-level policies. The many empirical studies that have evaluated performance-based reforms have focused almost exclusively on student test scores or the behavior of teachers or school administrators. Much less work has paid attention to how accountability affects the non-achievement behaviors of students. In this paper we begin to fill this gap.

To do so we use administrative data from North Carolina to examine the extent to which the accountability pressure generated under the federal No Child Left Behind Act (NCLB) affects student behaviors of two types: first, whether students show up to school when they are supposed to and second, whether students behave while in school. To identify the causal impact of accountability pressure on these outcomes, we estimate regression discontinuity models that leverage exogenous variation in accountability pressure at the arbitrary AYP cutoff.[3] Such an approach isolates the causal effect of accountability pressure apart from other distinct observable and non-observable characteristics of the students, schools, or surrounding communities.

[1] Quote taken from the 2004 Republican Convention, Sep 2, 2004.
[2] Julia Skinner-Grant was a fifth grader at Chevy Chase Elementary School in Montgomery County, MD in 2011 that wrote an op-ed on the impact of NCLB. See "5th grader's essay: High-stakes tests lead to stress, not learning", Valerie Strauss, Washington Post, June 3, 2011.
[3] Adequate yearly progress (AYP) is the criterion used to categorize schools based on student performance on standardized tests and other academic indicators. Schools that do not make adequate yearly progress are labeled "failing" and face sanctions if they fail multiple times consecutively.

We find that accountability pressure produces mixed results for the behaviors we examine. On the one hand, when schools face accountability pressure, students respond, as intended, by showing up to school and doing so on time. On the other, we find that pressure has unintended and undesirable effects. Our models indicated that accountability pressure appears to adversely affect how students behave when they are in school, where our measures of misbehavior include externalizing behavior in the form of suspensions, sexual offenses and offenses that are required to be reported to law enforcement. In addition to these overall results, we find substantively important heterogeneities across school and student characteristics. Student responses also vary depending on the non-achievement measures that NCLB requires schools to report. In addition, we find that increases in externalizing behaviors occur most among minority and low performing students, those who already exhibit higher levels of these anti-social behaviors at baseline. In sum, while performance-based school accountability produces some desired behaviors, it comes at the expense of harming non-incentivized behaviors.

Our analysis makes three main contributions. First, it speaks directly to the lively policy debate surrounding performance-based accountability. Despite more than a decade of experience with the federal No Child Left Behind program, the debate involving standards and accountability continues.[4] Our results provide policymakers with causal evidence of how accountability pressure affects students beyond their performance on standardized tests. Second, this paper reconciles the differing effects of accountability pressure on "showing up to school" vs. "behaving in school" by appealing to a multitasking principal-agent framework. In so doing, our paper extends this model to include situations where agents (school officials) must in turn delegate responsibilities to second-level agents (students). Our results show that such a framework is valuable in understanding why performance-based accountability applied to schools affects student behaviors both positively and negatively. Finally, our

[4] For example, a recent spate of federal waivers from No Child Left Behind's key provisions has allowed states to rethink how they will measure, report, and incentivize student performance (see for examples, http://topics.nytimes.com/top/reference/timestopics/subjects/n/no_child_left_behind_act/index.html).

work informs the growing body of research involving non-cognitive skills. An expanding literature has shown that student outcomes not fully captured by standardized test scores are important for performance in school and beyond (e.g., Heckman 2000; Jacob 2002; Gilman et al. 2006; Carneiro et al. 2007; Jackson 2012). Despite this literature, much less is known regarding the targeted policies that can help nurture, or alternatively, harm these skills. Our analysis suggests that education policies primarily targeted towards the development of cognitive skills—like NCLB—may also noticeably affect non-achievement metrics of student success. But, accountability-induced changes do not always occur in expected or desired directions. Instead of leaving no child behind, performance-based accountability policies may harm and perpetuate inequalities in the attributes shown to be important in school and beyond.

2. Background

2.1 Performance-Based Accountability and No Child Left Behind

In recent years, policymakers have implemented performance accountability systems widely across health, agriculture, law-enforcement, nonprofit, environment, foreign policy and education sectors.[5] These systems differ in their form and substance, but generally have three components: namely, measurement of performance, publication of results, and incentives to meet targets. Under the first component, policymakers set performance standards, measurement criteria, and determine how performance is reported. Under the second, individual actors' performance results are published. Finally, if the relevant actors fail to meet set standards, they face consequences. Prominent among these performance-based reforms is the federal No Child Left Behind Act of 2001 (NCLB).

NCLB is considered by many to be the "most far-reaching education policy ... over the last four decades", with the law substantially altering the education system by implementing a universal performance accountability system (Dee and Jacob 2011, 149). Under this system, student performance

[5] For an overview of performance-based accountability across these sectors, see Stecher et al. (2010).

is evaluated primarily using student test scores. Schools whose students fail to meet arbitrary performance thresholds are labeled "failing". Additionally, a less publicized provision requires schools to measure and report other academic indicators, which in many states include attendance (for elementary and middle schools) or graduation rates (for high schools).[6] If schools fail twice consecutively they enter a system of increasingly punitive sanctions. In the first sanction period, schools must allow transfers out of the school. In the second, schools must offer supplementary services (i.e. tutoring). In later periods, schools must alter their leadership structure (by removing administrators or implementing school-takeover).[7] The stigma that comes with failing and the anticipation and realization of these sanctions combine to place a significant amount of *accountability pressure* on schools that fail.

In contrast to policies that mandate exit exams, specifically incentivize student behaviors, or impose promotion requirements, NCLB applies pressure on schools, not students. As a result, accountability pressure is likely to affect school administrators and teachers most directly. Nonetheless, pressure may be likely to spill over to students: who, in turn, may react in positive or negative ways. Previous qualitative work supports this notion. Wheelcock et al (2000) note that accountability pressure spills over to students, who respond to the introduction of high-stakes tests with increased levels of anxiety, anger, pessimism, boredom and loss of motivation. Similarly, Hoffman, Assaf, and Paris (2001) argue that tested students under accountability systems often exhibit stomachaches and headaches that are indicative of increased levels of anxiety. Further, the conclusions of these qualitative studies are supported by teacher-level surveys, in which teachers report that students respond to testing with increased levels of anxiety and lower levels of confidence and "love of learning" (Jones et al. 2007).

[6]The other academic indicators we mention are those in place in the state we analyze: North Carolina. More generally, NCLB requires that schools report at least one other academic indicator. However, the law gives states leeway in deciding which measure to use. For example, states may use "state or locally administered assessments, decreases in grade-to-grade retention rates, attendance rates, and changes in the percentages of students completing gifted and talented, advanced placement, and college preparatory courses" (NCLB, Part A, Subpart 1, Section 111, 2CVii).

[7] Schools can exit these sanctions by passing two years consecutively.

While these accounts are illustrative of the potential impacts of accountability pressure on students, the findings may not be generalizable and the causal role that accountability pressure plays in moving these outcomes remains unclear.

2.2 Previous Research on Accountability Pressure

Previous empirical treatments of NCLB—and the similar performance reforms that preceded it—have focused primarily on its effect on student test scores or the behavior of school administrators and teachers. While a complete review of the studies that examine these outcomes is beyond the scope of this paper, they provide important context for this study of how accountability pressure affects non-achievement student outcomes.

Scholars studying the test score impacts of accountability pressure have used a variety of panel and quasi-experimental techniques. In one of the first studies to examine NCLB-like accountability pressure's effects on test scores, Ladd found that pre-NCLB accountability reforms in Dallas led to test score gains for Hispanic and Caucasian middle school students (1999). Similarly, Jacob (2005) found that accountability pressure increased test scores in Chicago, under that city's pre-NCLB accountability policy. Test score gains emerged, however, only for high-stakes exams suggesting that strategic behavior—or teaching to the test—occurred as a response to accountability pressure. In another study, Rouse et al. (2007) found that schools facing NCLB pressure saw some gains on student test scores. Using a national sample, Reback et al. (2011) also found modest positive gains in reading, math, and science tests in response to accountability pressure. Similarly, Neal and Schanzenback (2010) showed some evidence of test score gains in response to accountability, but primarily from those students close to the proficiency cutoff, with students well above and well below seeing few improvements. Finally, Ahn and Vigdor (2014a) used regression discontinuity with data from North Carolina to examine the effect of pressure on test scores. They found positive effects in schools first entering the NCLB sanction

regime and for schools entering higher sanction levels, but no effect of pressure in intermediate years. They also found some important differences across student subgroups.

Related to the research on test scores is a body of work that examines the specific responses of school administrators and teachers to accountability pressure. Examining the behavior of teachers, Clotfelter et al. (2004) found that accountability pressure made it more difficult for low performing schools to retain teachers. Feng et al. (2010) found similar results in Florida. Reback, et al. (2011) also found that accountability pressure lowered teacher perceptions of job security and increased the number of hours untenured teachers in high-stakes grades worked: perhaps explaining other findings regarding teacher exit. In their 2014 paper, Ahn and Vigdor found that failure—especially when restructuring sanctions are at stake—causes higher levels of administrative turnover (2014a). Additionally, Chiang (2009) found that schools respond to accountability pressure by increasing school spending for instructional technology, curricular development, and teacher training. Some studies also find evidence that principals respond to pressure by strategically determining the composition of students who take tests (Cullen & Reback 2006). For example, administrators in Florida's early performance-based system tended to "reclassify students as disabled and therefore ineligible to contribute to the school's aggregate test scores" (Figlio and Getzler 2006, 1). Such behavior was also observed in Chicago public schools (Jacob 2005). Finally, Figlio, and Winicki found that school administrators under accountability pressure were willing to go to extreme lengths, by "increas[ing] the caloric content of their lunches on testing days in an apparent attempt to boost short-term student cognitive performance" (2005, 381).[8]

In short, previous relevant work has shown that students react to school-level accountability pressure in their performance on standardized tests. Still, gains in test scores may be relatively small and may not be consistently realized across different tests of student cognitive ability, student

[8] For another example of shifts in administrator behavior in response to accountability see Ladd and Zelli (2002).

subgroups, or school characteristics. Moreover, the evidence shows that school administrators are sensitive to, and react to, the accountability pressure that originates when their school does not perform satisfactorily, with their actions often leading to unintended and undesirable outcomes. In total, this literature should provide some hope, but also pause for proponents of performance-accountability.

2.2.1 Accountability and Student Behaviors Beyond Test Scores

While the existing literature shows that students and school administrators are sensitive to school-level accountability pressure, only a few studies to date have examined how that pressure affects the non-test score behaviors of students. For example, in the Ladd study already mentioned, the author concluded that Dallas' accountability system reduced the dropout rate of high school students (Ladd 1999). Additionally, Chiang (2009) examined the impact of Florida's accountability system on absences and disciplinary incidents.[9] Although he found no overall impact of accountability pressure on these outcomes, he did find that student absences increased somewhat when schools failed repeatedly.

Additionally, some studies have used student responses to surveys about their experiences in school as outcomes. For example, as part of their analysis, Reback, Rockoff and Schwartz (2011) examined student reports of enjoyment of learning and anxiety towards testing. They showed that accountability pressure had little impact on enjoyment of learning and it decreased students' anxiety towards testing.[10] Similarly, Whitney's 2013 working paper borrowed Dee and Jacob's (2012) comparative interrupted time series approach to examine the effect of NCLB implementation on various survey-based measures of student attitudes, such as their "love for learning." She found no statistically significant effect of performance-based accountability on these survey measures.

[9] See Chiang's online appendix (table B.4).

[10] Reback et al. (2011) speculate that a decrease in student anxiety to accountability pressure occurs because failing schools prepare students for exams by using practice exams.

Aside from these few studies, the empirical literature on performance-based accountability includes little examination of how it affects non-test score measures of student behavior. This situation is unfortunate given the growing literature that documents the significant contribution to student success of what many researchers refer to as "non-cognitive skills."

2.3 Non-Cognitive Skills

The various, so called "non-cognitive skills" that children develop (or fail to develop) while they are in school capture the learned attitudes, behaviors, and strategies that help children assimilate in society. These attributes have a number of titles, including: soft skills, non-cognitive abilities, character, emotional intelligence, meta-cognitive learning skills, and socio-emotional skills. Examples of non-cognitive skills include: grit (Duckworth et al. 2007), prosocial behavior (Ladd G. 2005), and emotional regulation (Tomer 2003). Non-cognitive skills share the common characteristic of being distinct from measures of cognitive proficiency and personality traits largely inherited at birth (Heckman 2000; Heckman and Kautz 2013).[11] A growing body of research suggests that these skills are central to performance in school and beyond (e.g., Heckman and Kautz 2013).

As with all constructs, measurement of non-cognitive skills comes with error, regardless of the approach used. In this paper, the outcomes we examine come from administrative reports of student behaviors. This approach follows the observed-behaviors technique to understanding students' skills (e.g., Jacob 2002; Carneiro et al. 2003; Heckman et al. 2006; Heine et al. 2008; Heckman, Pinto, and Savelyev 2013).[12] Scholars who use this approach argue that observed human behaviors are informative

[11] We choose to use the term "non-cognitive skills" to show that our examination is separate from the analyses of performance-based accountability's impact on measures of cognitive proficiency. In so doing, it is not our intention to address whether non-cognitive skills rely on cognitive processes.

[12] Importantly, the observed-behaviors approach benchmarks well with survey-based methods of measuring non-cognitive skills. For example, Pratt and Cullen show that survey and behavioral measures of self-control appear to measure a similar underlying construct, with these measures being similarly predictive of crime in adulthood (2000; see also Benda 2005). Furthermore, the observed-behaviors approach has other virtues. It avoids problems of survey-based measures such as reference bias and survey item non-response (Heckman and Kautz 2013).

of the underlying set of skills regulating those behaviors (Heckman and Kautz 2013, 13-21). For example, Heckman et al. (2011) draw inferences about the impact of the education system on non-cognitive skills from measures of observed misbehavior. Similarly—and most comparable to the approach we employ in this paper—Jackson (2012) draws inferences about the impacts of teacher quality on students' non-cognitive skills by using observed absences and suspensions found in the same North Carolina school administrative files we use.

Following this work, we use three types of observed behaviors that relate to children's abilities to "be in school when they are supposed to" and to "behave while they are in school"—namely, absences, tardies, and student misbehaviors. The first two—absences and tardies—are likely strongly related to the underlying skills associated with showing up on time. These skills are likely to be important in school and later in life in the workforce, as both of these settings require that individuals know how to adhere to a set schedule (Gottfried 2009). The misbehaviors we examine include: suspensions, fights, possession offenses, violent offenses, sexual offenses, weapons-related offenses, disruptive offenses, falsification-related offenses, and reportable offenses. Avoiding these externalizing behaviors is also an important skill in school and beyond. Our examination of behaviors relevant to showing up to school and these externalizing behaviors provides a broader picture of NCLB's impacts than has been documented in previous work.

3. Conceptual Framework

To frame our examination of the impact of accountability pressure on whether students "show up when they are supposed to" and "behave while they are in school", we appeal to a multitasking

Moreover, as administrative data has grown, the observed-behaviors approach allows for explorations in broader contexts.

framework. [13] Under this framework, school personnel have multiple tasks. These include improving student achievement, getting students to come to class, and encouraging them to behave once they are there.[14] To accomplish these tasks requires resources—including time, money, and personnel—that are in limited supply. Hence, school administrators face tradeoffs in how to allocate resources to these tasks.

Such tradeoffs are amplified by the incentives provided by NCLB. According to a standard principal-agent model, if the principal actors (in this case policymakers) incentivize only some tasks, the agents (in this case principals and teachers) will devote more attention to the incented tasks and less to the others (Holmstrom and Milgrom 1991; Gibbons 1998; Laffont and Martimort 2009; Fryer and Holden 2012). NCLB incentivizes school officials to place attention on raising scores on standardized tests and other academic indicators, while not incentivizing other behaviors, such as how students behave when they are in school.

With respect to the first task of improving standardized test scores, educators have some clear tools at their disposal for inducing desired student outcomes. Some of these tools—such as improvements to curriculum or teaching practices—may generate long-term positive gains in learning. Other tools for raising achievement, however, may lead to more limited, short-term gains. For example, if administrators lack the capacity to assure that their students realize specified achievement goals, they may game the system in various ways.

Educators also have some methods for pursuing the second task of making sure that students come to class. They can send out reminders, report previous attendance, and threaten various punitive

[13] Throughout the paper we reference two constructs. First, we use the terms "behaving in school", "anti-social behaviors", and "externalizing behaviors" interchangeably to specifically describe student misbehaviors. Second, we use the terms "showing up", "being where they are supposed to be when they are supposed to be", and "attendance-related behaviors" to reference the specific measures of absences and tardies. It is not our intention to confuse, but rather to avoid repetitious use of the same term.

[14] These are, in fact, a smaller subset of the tasks school officials face. Other tasks include: instilling democratic values, bestowing a love of learning, teaching practical skills for the workforce, etc.

measures or legal actions for those who do not show up. In short, school officials can promote both higher student test scores and better attendance at school by transferring accountability pressure to students.

School officials also have some levers to further the final task of ensuring that students behave in school. Teachers can devote classroom time to teach the non-cognitive skills associated with regulating behavior in social settings. Additionally, administrators can put various measures in place to encourage students to conform to a set of behavioral standards. For example, they can implement various components of "no-excuse" reforms, by placing requirements on students regarding their behavior (e.g., Angrist et al. 2012). Students who do not meet these rigorous requirements may face short-term punishments such as limits on extracurricular activities or long-term punishments such as removal from the school. In contrast to the other tasks, however, accountability pressure typically provides no direct incentives for educators to use these tools.

3.1 Hypotheses

This conceptual framework leads to a set of testable hypotheses about the student behaviors that are the focus of our empirical work. NCLB specifically incentivizes schools to limit the absentee rate, at least in some states in certain grades. Given that absences are easily measured and are included in the determination of adequate yearly progress in North Carolina among elementary and middle schools, we think it reasonable that schools facing accountability pressure find ways to decrease absenteeism. Still, this increased emphasis by school officials may or may not come to fruition because students may be resistant or even adversarial to these changes.

Our expectation for the second metric—student tardies—is ambiguous because NCLB does not directly incentivize schools in North Carolina to reduce the rate of tardies. Still, students might alter their behavior in response to the related changes NCLB encourages. The direction of this shift, however, is less clear. On the one hand, we might expect that students' decision to show up on time (i.e. to avoid

tardies) shares a common construct with the broader decision to show up at all (i.e. to avoid absences). Under this scenario, the resources dedicated to decreasing absences might spill over into tardies, causing a decline in their frequency. On the other hand, stress from accountability pressure could make students behave in undesirable ways such as arriving at class late while they are at school. Hence, the direction of the effect of accountability pressure on the number of tardies is ambiguous.

Our third sets of measures—student misbehaviors—are not directly incentivized by NCLB. As a result, it is difficult to predict how accountability pressure will affect them. On the one hand, changes in curricular or teaching strategies designed to raise test scores and attendance may spill over into student behaviors in a positive manner, leading to a decline in misbehaving. On the other, accountability pressure could lead students to misbehave more often. If students themselves have limited capacity to respond in positive ways to increased pressure to do better on tests, for example, they may respond by acting out. Given a multitasking framework in which pressure on schools and educators is transferred to pressure on students, accountability pressure may lead to higher levels of student misbehavior unless the schools take explicit actions to counter that behavior as part of their efforts to raise student achievement.

In addition to these overall expectations, we hypothesize that accountability pressure will generate different effects on different groups of students. The largest negative effects may occur for the students who are least able to meet the requirements that administrators place on them. These students, stressed by increased levels of pressure and lacking the means for reaching proficiency, may react more negatively than their more-able counterparts. Alternatively accountability pressure may place needed attention on students being left behind. Indeed, this was part of the law's original focus: to "leave no child behind" (Bush 2004).

Whether accountability pressure has these expected effects, or any effects at all, on our outcomes of interest is an empirical question.

4. Data

4.1 School Accountability Pressure

Our independent variable of interest is the accountability pressure (P_{st}) schools (s) face under NCLB in a given year (t). This pressure is a function of two factors: first, whether a school fails to make adequate yearly progress in the previous year (F_{st-1}) and second, whether the school is already subject to sanctions in the current period (I_{st}) because of its previous performance.

$$P_{st} = f\,(F_{st-1}, I_{st}) \quad [1]$$

$$I_{st} = f\,(F_{s,\,t-1}, F_{s,\,t-2} \ldots F_{s,t-n}) \quad [2]$$

Schools that fail to make adequate yearly progress (F_{st-1}) in a given year face a discrete jump in accountability pressure during the next school year. Schools face this pressure for three reasons: they face the negative stigma of being labeled a "failing" school, they anticipate future sanctions that will come if they fail in the future, and they have failed previously and are currently under sanction (I_{st}).

It is important to note, here, that we use the term "accountability pressure" throughout this article as shorthand for the pressure that originates from school failure. It is important to acknowledge that all schools, regardless of their performance, face some accountability pressure because they are all subject to NCLB's requirements. Schools that marginally meet NCLB's standards, for example, feel pressure because of the potential for failure in future years. Although some researchers have argued that the difference in accountability pressure between passing and failing schools near the adequate yearly progress cutoff may not be very large (Ahn & Vigdor 2014a, 2014b; Dee & Jacob 2011), failing schools undoubtedly face an added dose of accountability pressure because of negative stigma that comes when schools fail and their greater likelihood of facing sanctions. This difference is likely to be meaningful even if it does not fully capture the sum total of pressure imposed by the implementation of accountability systems.

Failure under NCLB is determined by a complex formula. At its most basic level, the students in a school are required to perform up to certain levels of proficiency on standardized tests. Roughly

speaking, schools with a sufficiently high percentage of students scoring at or above proficiency on the state's tests pass, while schools that do not, fail. No Child Left Behind complicates this simple determination by including a number of sub-provisions. Among these, NCLB mandates that performance be assessed for ten subgroups, which include: all students, American Indian, Asian, Black, Hispanic, multi-race, Caucasian, economically disadvantaged, limited English proficiency, and students with disabilities. Each of these subgroups must meet a set of performance thresholds in both reading and math[15] and all sub-groups must pass for the school to avoid failing.[16] A second sub-provision requires that schools can meet each of the subgroup thresholds through one of three channels, which include: passing by simply having enough students at proficiency (termed "passing with level"), improving significantly from one year to the next ("passing with growth") or being arbitrarily close to passing ("passing with confidence interval").[17] Finally, schools must report other academic indicators such as absentee rates (in elementary and middle schools) or dropout rates (in high schools) for each subgroup. These conditions combine to determine whether the school makes or fails to make AYP.

When schools fail twice consecutively, they enter a graduated system of sanctions. In the first year of sanctions, schools must allow students to transfer within their district. In the second year of sanctions, schools must offer supplementary education services—such as tutoring or other after-school programs. In the third year, schools must take corrective action. In the fourth year of improvement, schools must formulate and implement a restructuring plan, which entrails altering of school leadership or altering the school's categorization (e.g., to a charter). Schools can exit this system by making AYP in two consecutive years.

[15] Other tests such as science are taken in NC schools. These, however, do not go towards determining AYP.
[16] Reporting exemptions are given to schools when they have small number of students in a given subgroup.
[17] Some notes about the second and third channels as they are applied in North Carolina are in order. First, the thresholds for passing with level were arbitrarily set each year. These were constant across schools, but rose over time. Second, students must grow at differential levels in their performance on the test scores (10%), graduation (2%), and attendance (10%) to use the growth channel. Growth is only available for subgroups performing at a low level (up to 80% graduating and 90% attendance). Finally, passing with confidence interval only applies in one direction—there is no such thing as failing with confidence interval.

In this article, information on the accountability pressure individual schools face is based on North Carolina's public school performance data from 2006-2011. These data publicly report school failure status the summer after a given school year (around July 30th). In any given year in North Carolina, the number of schools failing varies widely, from 20-80% of schools. The most commonly failed subgroup categories come, perhaps unsurprisingly, from the performance of low-SES and minority subgroups: that is, groups with historically low performance.

4.2 Outcomes: Reported Student Behaviors

Our dependent variables include measures of student absences, tardies, and misbehaviors. These come from the North Carolina Education Research Data Center (NCERDC) data files on student offenses and demographic data.[18] To explore the effect of accountability pressure on these outcomes, we constructed a panel with many student characteristics for all public school students in the state over a six-year period.

With these data, we match school performance in a given year (t) to our student-level outcomes in the following year (t+1). We use a leading dependent variable to guarantee that our outcome measures occur after the "treatment" of school failure.[19] In practice, this approach means that we use matched school performance data from 2006-2011 to student behaviors from 2007-2012.[20] Our full estimation sample consists of about five million student-year observations nested in about 11,000 school-year observations.

[18] As with all administrative data sets, there are some likely data entry errors in the NCERDC offenses file. To mitigate this, we impute extreme values for schools, replacing them with the top non-outlier observation. We do this for all our dependent variables.

[19] If administrators anticipate that their school will fail in year T, this could affect student performance in year T. Anticipating marginal failure is difficult, however, because marginal failure in one year does not guarantee marginal failure in the next. Because many measures go into determining AYP, it is difficult to correctly anticipate failure and the accountability pressure that follows.

[20] Data for our outcomes are not available before 2007.

In our analyses we examine absences and tardies separately.[21] This separate analysis provides an informal check that the measures are indicative of student behavior, rather than strategic under- or over-reporting by school administrators. School administrators in North Carolina have little incentive to alter how they report tardies because, in contrast to absences, they are not included as an academic indicators under NCLB. As a result, a finding that absences and tardies move in the same direction in response to accountability pressure, gives us added confidence we are measuring a change in student behavior and not an artificial change in reported levels.[22]

For student misbehaviors, North Carolina documents approximately 70 reportable offenses, each of which can occur at multiple points for an individual student in a given year (see table A.1 in the appendix for a full list). Using each of these as separate outcomes is unpalatable due to the problems, and limited solutions, associated with multiple hypothesis testing. Following previous work examining student misbehaviors (e.g., Flay et al. 2004), we group offenses together into a set of logically coherent categories, aggregating the 70 measures into seven separate additive scales.[23] These scales include the number of drug-related behaviors (termed "possession" in tables below), violence-related behaviors, risky sexual behaviors, weapons-related behaviors, disruptive-related behaviors, acts involving some form of deception (termed "falsification" below), and offenses that are reportable to law enforcement agencies.

In addition, we examine three student misbehaviors individually because of their frequency and severity. These measures are fights, in-school suspensions, and out-of-school suspensions. Fighting, for

[21] At the individual level, estimates for tardies are conditional on attendance.

[22] We are somewhat skeptical of school's ability to fake absence numbers. Doing so would require coordination of teachers, principals, and other school administrators to count students as present when they are not. Moreover, North Carolina has several checks in place to make sure that these numbers are reported accurately. Changing these numbers (i.e. underreporting absences) is not impossible, but also not simple.

[23] Our measures follow the general pattern found in surveys of youth behaviors, such as the Youth Risk Behavior Surveillance System. Scholars who use these tend to group misbehaviors into sexual, drug, violence, delinquent, and health categories (Grunbaum et al. 2004; Eaton et al. 2011). Our grouping follows a similar logical hierarchy for the measures that fit into these categories. Given North Carolina' rich classification of offenses we are also able to construct additional scales that do not fit into these traditional categories (weapons, disruptions, falsification, and reportable). In the appendix Table A1, we show the individual offenses used to construct these additive scales.

example, is the most common single observed misbehavior; and suspensions, whether in school or out of school, reflect more serious types of student misbehavior.[24]

Our measures of student misbehavior reflect both the underlying student behavior in which we are most interested and possible official reactions to that behavior. Schools might react to accountability pressure, for example, by getting tougher and reporting misbehaviors more actively than they otherwise would. Although schools have no direct incentive under NCLB to change their reporting or punishment of student misbehavior, we cannot entirely rule out this response. Nonetheless, even if were the case, that some of the changes in reported misbehaviors reflected a change in how administrators respond to student behaviors, there would still be reason for concern. When students are suspended at a higher rate, for example, they miss valuable instructional time.

5. Methods

We leverage a discontinuity in the determination of AYP and use regression discontinuity (RD) models to isolate the causal effect of accountability pressure on our outcomes of interest. As has been well established, regression discontinuity allows scholars to draw causal inferences in analyses that use observational data (e.g., Imbens and Lemieux 2008; Lemieux and Milligan 2008; Lee and Lemieux 2010). Under this method, observations close to an arbitrary cutoff are separated by exogenous shocks (Butler and Butler 2006, 443-444). Applied to our NCLB case, schools very close to failing could have easily fallen on either side of the arbitrary cutoff. These schools are separated by small, quasi-random events that push them to one side of the arbitrary cutoff or the other. RD models take advantage of this exogenous variation, using data on either side of the cutoff to estimate the change in an outcome. While it generates good internal validity, this technique may come at the cost of limiting inferences to units around the cut point.

[24] Some have observed that administrators act strategically in their suspension decisions (Figlio 2006). In North Carolina, administrators are somewhat limited in their ability to suspend strategically by NCLB, which requires that schools report the number of tested students.

The starting point for any RD model is the identification of the treatment and the running variable. In our NCLB application, identifying treatment status is relatively straightforward: treatment consists of a school failing to make adequate yearly progress (AYP) and our control consists of schools making AYP. As school failure status is publically available, it is easy to identify.

The running variable—in this case the variable that determines how close schools are to failing—is more difficult to identify because the basic calculation for determining school failure is complicated by the two sub-provisions we referred to earlier. First, because NCLB requires all subgroups to pass, if one subgroup in one subject fails, the school fails. Second, because schools can achieve the cutoff through three channels—by simply meeting the cutoff requirement, by being close to meeting the requirement, or by improving sufficiently from one year to the next—the school passes if any one of these channels puts a school over the arbitrary cutoff. Hence, to approximate how close schools are to failing we have to capture both subgroup scores and channels of passing. With multiple subgroup categories and three channels of passing in each subgroup, identifying the running variable is no small task.

To do so, we use the approach proposed by Ahn and Vigdor (2014a) for their study of how NCLB pressure affects student test scores.[25] This procedure mirrors the codified rules in NCLB and chooses one channel of passing per subgroup, and then one subgroup per school to represent the running variable. For each subgroup we first choose one channel of passing. The decision rule for choosing the channel of passing is:

> [**D1**] For each subgroup in a school, choose the channel that gives the subgroup the highest score.

The intuition behind [**D1**] is that under NCLB if any one channel places the subgroup above the AYP threshold, that subgroup is marked passing. Thus, the channel that indicates the highest school

[25] See Jacob and Lefgren (2004); Matsudaira (2008); Balcolod, DiNardo, and Jacobson (2009); Imbens and Zajonc (2011); Reardon and Robinson (2012); Wong et al. (2013); and Holbein (2014) for applied examples of similar approaches.

performance identifies how far a school's performance would have to deteriorate to not pass through at least one channel. Conversely, if all channels are below the AYP threshold, the maximum channel chooses the threshold closest to passing AYP.

Once a channel of passing is decided for each subgroup, we choose one subgroup score as a measure of the running variable. The decision rule we use for choosing the subgroup score is:

[**D2**] For each school, choose the minimum subgroup score.

The intuition behind [**D2**] is that under NCLB if any subgroup score falls below the cutoff, the school fails. If schools are failing, passing only occurs once all subgroup categories are brought above the threshold. Thus, the lowest subgroup score approximates how far a failing schools has to improve to pass. Conversely, passing schools are most likely to fail if their lowest subgroup score slips below the threshold.[26]

Although it is not the only possible approach, we use the Ahn and Vigdor approach because it conceptually mirrors the process of determining failing/passing under NCLB.[27] As a result, this approach is relatively accurate in sorting schools into the correct pass/fail groups based on their running variable score and correctly identifies 80% of schools.[28] Moreover, this approach benchmarks well with slightly different methods of specifying of the running variable in the NCLB context (Traczynski and Fruehwirth

[26] This logic makes several assumptions, including: that all channels improve (or deteriorate) in an order-preserving fashion that accountability pressure comes from failing overall, not the number of conditions failed, and that performance in years previous to the most recent years does not influence proximity to failure in the most current year. We argue that these first two are justified given their alignment with NCLB's AYP determination. We relax this last assumption later in the paper by allowing for failure in previous years to influence outcomes in the current year.

[27] Some work has begun to grapple with the issue of specifying the running variable with multiple inputs. This work deals with multiple measures determining multiple treatments (Papay et al. 2011, 204).

[28] Misidentification of school failure status does sometime occur: the proximity variable sometimes indicates that a school failed, when we know from the public data that the school actually passed, and vice-versa. This misidentification comes primarily because of ambiguity in the interval channel (the interval used is not made public) and the other academic indicators.

N.d.).[29] Additionally, our approach bests a naive averaging over the subgroups, which correctly identifies only about 50% of schools. Given the presence of some error in our proximity measure, we have to use a fuzzy regression discontinuity approach (Matsudaira 2008).

With both treatment and the running variable specified, we can estimate our fuzzy regression discontinuity model as follows:

$$F_{st-1} = \gamma_0 + \gamma_1 P_{st-1} + g(R_{st-1}) + \varepsilon_{st} \quad [3]$$

$$O_{st} = \beta_0 + \beta_1 \hat{F}_{st-1} + g(R_{st-1}) + \beta X_{st-1} + \varepsilon_{st} \quad [4]$$

Equation [3] models observed actual failure (F_{st-1}) as a function of the failure status of the school as predicted by the running variable (P_{st-1}) and the running variable (R_{st-1}) [30] In equation [4], we estimate each outcome variable (O_{st}) as a function of the instrumented failure variable , ($\widehat{F}_{st-1}$), proximity to failure (R_{st-1})—which we model with a flexible non-parametric form, denoted by g(and a set of controls (X_{st-1}).[31] For the running variable, we use the nonparametric approach proposed by Hahn, Todd, and Vander-Klaauw (2001). This approach uses flexible local linear regression to fit smoothed functions of our running variable that are flexible on either side of the failure cutoff. To increase precision and to ensure that any slight covariate imbalances at the failure discontinuity do not confound our results, many of our models also include statistical controls for variables that show some sign of being imbalanced at the failure cutoff. Because the behavioral outcome measures we use are count measures that are skewed towards zero for individuals, and toward low numbers at the school level, we transform them into logarithmic form.[32]

[29] Traczynski and Fruehwirth (N.d.) find that the results for test scores are similar regardless of whether they use the minimum test passes or minimum subgroup score. The minimum test passes metric does have the virtue of producing more precise estimates.

[30] P_{st} = 1 if the running variable indicates that a school has failed, regardless of actual school failure status. P_{st} = 0 if the running variable indicates that a school has passed.

[31] We use a triangle kernel that places greater weight on points around the cutoff (Nichols 2012).

[32] Some of our measures, particularly the misbehavior measures, are relatively rare. As such we do have some instances where schools reported no offenses in a given year. Rather than throw these schools out of our sample, we impute these with an arbitrarily low number (1) in our models here. Our results tend to remain similar if we hold these schools out of our estimation sample. However, including them increases our precision. At the

Below we show that our models are robust to a variety of necessary modeling choices. One choice relates to the size of the bandwidth. Our preferred results are based on the optimal bandwidth recommended by Imbens &Kalyanaraman (2012), but we also report results for other bandwidths. The narrower is the bandwidth the more confident we are that the schools being compared are essentially random, but that comes at the cost of a smaller sample size and less precision. In addition, we provide estimates from models with and without controls for the variables that show any sign of imbalance at the cutoff. Overall, our results are robust to variations in these modeling decisions.

We estimate our models from data collapsed to the school-year level, and weighted by the number of students in a given school in a given year. This approach follows that of other similar work involving NCLB (e.g., Ahn and Vigdor 2014a; Traczynski and Fruehwirth N.d.) and allows us to preserve the virtues of our rich individual-level, account for the clustered nature of our school-level treatment, and model our outcomes in a way that is readily interpretable and justified by the distribution of our outcomes. As such, the coefficients for failure can be interpreted as the weighted average percentage change in outcomes in schools as a result of accountability pressure. These represent causal estimates provided the "as-good-as random" assignment of schools holds at the failure cutoff.

6. Specification Checks

If the assumptions of regression discontinuity hold, the estimate for Failing School (β_1) will be unbiased by confounders or simultaneity because schools fail as-good-as randomly within a narrow bandwidth (Lee 2008; Lemieux and Milligan 2008). Thus, the estimates are similar to those in a randomized-control experiment. Exploring whether our discontinuity satisfies the implications of local randomization, then, is of the upmost importance.

---

individual level, our outcomes are heavily skewed towards zero—many students have few absences, tardies, and offenses, especially. Modeling alternatives that account for this distribution—such as Poisson, tobit, or negative binomial models—converge at a sufficiently slow rate to make such approaches unpalatable. This is due to the complex inversion required when combining our two-staged procedure, a clustering adjustment, and our large sample size. When our measures are collapsed to the school-level, they become roughly continuous and normally distributed (once logged).

As with traditional experiments, a check of the identifying assumptions surrounding regression discontinuity involves examining covariate balance across treatment and control groups at the failure cutoff. For this check, we estimate the regression discontinuity models presented in the previous section on a set of potential confounders. Specifically, we run models [3] and [4] using lagged versions of our outcome variables and other school characteristics as the dependent variables. If these models generate significant estimates at the cut point for failing, one might question whether schools are indeed randomly distributed across the cut point.

Table 1 shows this specification check for models that do not control for prior failure status (columns 1 and 2) and those that do (columns 3 and 4). We account for previous failure to allow for the possibility that lagged school performance influences our lagged outcomes. The entries in columns 1 and 3 (labeled T-C) are the estimated coefficients of the differences between treatment (i.e. failing) schools and control (i.e. passing) schools at the cut point from separate models for each of the variables listed in column 1. The entries in columns 3 and 5 are the probabilities that the estimated coefficients are zero, given the truth of the null hypothesis. If schools were to fail in an as-good-as random fashion at the school failure cutoff, we would expect most of these tests to not be statistically significant at standard levels, such as 0.05.

Table 1: Balance at the Discontinuity

Variable	Without Previous Failure (1) T − C	(2) $P(T=C\|H_0)$	With Previous Failure (3) T − C	(4) $P(T=C\|H_0)$
A. Lagged Dependent Variables				
Log(in-school Suspensions)	-0.05	0.50	-0.06	0.47
Log(out-school Suspensions)	0.05	0.42	0.05	0.36
Log(Fights)	0.03	0.71	0.07	0.37
Log(Possession-Related Offenses)	0.05	0.40	0.06	0.29
Log(Violence-Related Offenses)	0.00	0.99	0.04	0.62
Log(Sexual-Related Offenses)	0.07	0.26	0.09	0.20
Log(Weapons-Related Offenses)	-0.01	0.62	-0.01	0.47
Log(Disruption-Related Offenses)	0.01	0.94	0.06	0.57
Log(Falsification-Related Offenses)	0.01	0.86	0.02	0.71
Log(Reportable Offenses)	0.03	0.71	0.02	0.84
Log(Absences)	-0.18	0.25	-0.24	0.15
Log(Tardies)	0.14	0.32	0.12	0.38
B. Other Controls				
Log(Truancy)	0.09	0.32	0.13	0.20
# of Students	7.22	0.69	26.48	0.21
% Title I School	-5.01%	0.12	-3.81%	0.26
Pupil/Teacher	0.55	0.05	0.61	0.07
% Hispanic	-0.32%	0.64	-0.22%	0.74
% Limited English	-0.07%	0.88	0.01%	0.98
% Reading Homework (HS)	0.55%	0.53	0.59%	0.53
% Female	0.73%	0.13	0.06%	0.91
% Migrant	-0.08%	0.05	-0.04%	0.37
% Free/Reduced Lunch	0.02%	0.99	-0.03%	0.98
% Gifted in Reading	-0.70%	0.38	-1.00%	0.21
% Gifted in Math	-0.14%	0.86	-0.43%	0.60
% African American	1.26%	0.26	1.19%	0.33
# Exempt Subgroups	0.14	0.46	0.27	0.26
Failed in a Previous Year	6.88%	0.00	.	.

Note: Check for covariate balance at the pass/fail margin (i.e. check for local randomization). The second and fourth columns show the difference between treatment (failing) and control (passing) groups on these baseline covariates. Differences are the coefficients on failure from a regression discontinuity model measuring the effect of school failure on these lagged outcomes. The differences reported in columns (3) - (4) control for school failure status in the previous 4 years, to account for the possibility of imbalances being driven by failure previous to when our outcomes are available. All models use the optimal bandwidth and a local-linear specification of the running variable.

For the most part, we find balance of potentially important variables at the cutoff across these two methods. Importantly, the lagged versions of our 12 dependent variables are balanced at the cutoff (panel A).[33] This placebo test provides powerful evidence that our discontinuity is sorting schools in an as-good-as random fashion.

Among the 14 other variables we examine in panel B, 11 are balanced and the remaining 3 show minimal imbalances. Marginally failing schools are similar to marginally passing schools with respect to

[33] We determine a variable to be balanced if both columns 3 and 5 do not reach significance. This is a conservative approach.

school size, the percent female, migrant, free/reduced lunch, gifted in math and reading, Hispanic, LEP, Disability, and the percentage of students who indicate that they spend time after school on reading homework.[34] On some metrics, however, there is evidence of imbalance. Marginally failing schools have slightly higher pupil/teacher ratios and a higher probability of having failed previously.

In total, then, 23 out of 26 (roughly 90 %) of our baseline covariates are balanced. These results are consistent with a discontinuity that sorts schools to either failure or passing status in an as-good-as random manner. Still, to be conservative we explicitly control for the three variables that show some sign of imbalance in our model. These controls should increase precision, while providing us with a check of the robustness of our results to potential omitted characteristics.

A second specification check involves ruling out precise sorting—or widespread manipulation of the running variable by schools near the cutoff. The test for covariate balance provided in Table 1 is one way to check whether precise sorting occurs (Lee & Lemieux 2010). Another check for this violation is the McCrary density test (McCrary 2008). This test looks for whether an abnormally large number of schools are located near –perhaps just on the passing side –of the failure discontinuity. If this occurred, we might be worried that schools were able to precisely manipulate their running variable score close to the failure cutoff (Lee & Lemieux 2010; McCrary 2008). While we think this pattern is unlikely given the large number of subgroups scores that go into our running variable, we tested for this possibility and concluded that there precise sorting was not a problem (see appendix).

Taken together, these specification checks assure us that the NCLB failure cutoff sorts schools in an as-good-as random manner, which allows us to proceed with our analysis of how accountability pressure affects student behavior.

[34] This survey item is only available for high school students.

7. Results

7.1 The Effect of Failure on Being in School

We start with the impact of school failure on absences and tardies, that is, whether students are "in school when they are supposed to be." Table 2 shows that, consistent with our expectations, accountability pressure causes a reduction in the number of student absences and the effect size is noticeable. Our preferred estimates are in the first row, which is based on the optimal bandwidth. We see that failure causes absences to decline by about 60%, on average, during the following school year, and the effect is statistically different from 0 at the 5% level.[35] While there is some variation in the size of this estimate across model bandwidths as shown in the next three rows and in the corresponding column with controls, the direction of the effect does not vary. Our estimate represents about 0.22 of a standard deviation in logged student absences. This large estimate is not simply of a low base rate of absences—converting the logged measure to a count measure of absences, we estimate that failure causes about 280 fewer absences in a school, on average. To further put this effect size into perspective, a back of the envelope calculation reveals that this school-level reduction represents a reduction of about 0.5 fewer absences per student, on average.

Table 2 also shows the effect of accountability pressure on tardies. As we mentioned earlier, the predicted effect of accountability pressure on tardies is ambiguous. Our results suggest that even though NCLB does not directly incentivize tardies, the fact that more school resources are devoted to encouraging students to show up in school dominates any negative response to the pressure that students may exercise in the form of being late. When schools receive an added dose of accountability pressure from failing, tardies decline by about 24%, on average. This effect is both distinct from zero at

[35] That this effect is not significant in very narrow bandwidths is likely a matter of sample size rather than a lack of an effect. This estimate is statistically indistinguishable from estimates in wider bandwidths that allow for more precision.

the 5% level and substantively meaningful. It represents about 0.10 of a standard deviation, which equates to about 80 fewer tardies at the school level or about 0.20 tardies per student, on average.

Table 2: The Effect of Accountability Pressure on Being in School

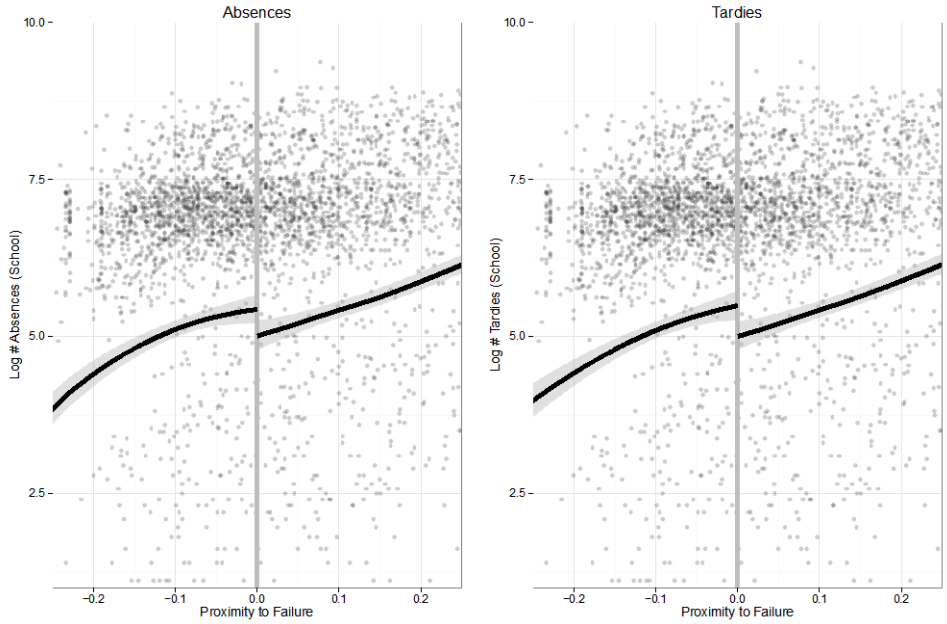
	Without Controls		With Imbalanced Controls	
	DV: log(Absences)	DV: log(Tardies)	DV: log(Absences)	DV: log(Tardies)
School Failure (Optimal Bandwidth)	-0.616***	-0.237**	-0.705***	-0.273**
	(0.205)	(0.111)	(0.210)	(0.111)
Number of students	4,658,783	4,951,462	4,339,160	4,600,265
Number of schools	9,490	10,059	8,342	8,815
School Failure (Half optimal)	-0.277	-0.154	-0.348	-0.163
	(0.271)	(0.144)	(0.275)	(0.143)
Number of students	3,111,584	3,651,008	2,918,150	3,418,482
Number of schools	6,317	7,464	5,621	6,632
School Failure (Twice optimal)	-0.927***	-0.353***	-1.050***	-0.380***
	(0.175)	(0.099)	(0.179)	(0.099)
Number of students	5,265,619	5,288,959	4,892,095	4,915,035
Number of schools	10,612	10,670	9,303	9,359
School Failure (Full)	-0.904***	-0.339***	-1.036***	-0.353***
	(0.165)	(0.096)	(0.170)	(0.096)
Number of students	5,289,083	5,289,083	4,915,159	4,915,159
Number of schools	10,671	10,671	9,360	9,360

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Cluster-robust standard errors (at the school-year level) reported below coefficient estimates. Running Variable: local-linear. Optimal bandwidth chosen using the procedure suggested by Imbens & Kalyanaraman (2012). Controls: Those showing imbalance in Table 1 including: pupil/teacher ratio, % migrant, and whether a school failed previously.

The effect of accountability pressure on absences and tardies is shown visually in figure 2.[36] The figures show the discrete decline in absences and tardies just on the failing side of the AYP cutoff. In short, when schools feel accountability pressure, they respond with efforts that end up helping students show up to school and to come to class on time.

[36] These figures draw their predictions from our models without controls.

Figure 2: Accountability Pressure, Absences & Tardies



Notes: In figure 2 individual points represent individual school-year observations. A dose of accountability pressure is administered to schools on the right side of the cutoff—where schools fail. The causal effect is the vertical difference between the two color-corresponding lines at the cutoff. The causal effect is from the estimates from the optimal bandwidth (Imbens & Kalyanaraman 2012). Local-linear regression is used to model the relationship between the running variable and the outcome.

Given that high rates of absenteeism and tardies bode poorly for all aspects of student learning, the observed decline in absences represents a normatively positive effect of NCLB.

7.2 The Effect of Failure on Misbehaving in School

Table 3 shows the impact of failure on ten measures of student misbehavior. Recall that three of these are individual measures of offences (fighting, and in and out of school suspensions) and the other seven are constructs that we have labeled possession of controlled items, violent, sexual, weapons-related, disruptive, falsification-related, and reportable offenses. For the sake of ease in viewing the results, we just present the results from our RD models with controls. The results without controls are substantively similar.

Table 3 shows that across a variety of measures, accountability pressure appears to induce students to misbehave more than they otherwise would. While lacking precision in narrower bandwidths, our estimates reach statistical significance when more data are employed in wider bandwidths. Some of the

estimates are larger in wider bandwidths, however, estimates from these wider bandwidths are generally statistically indistinguishable from those from narrower bandwidths.

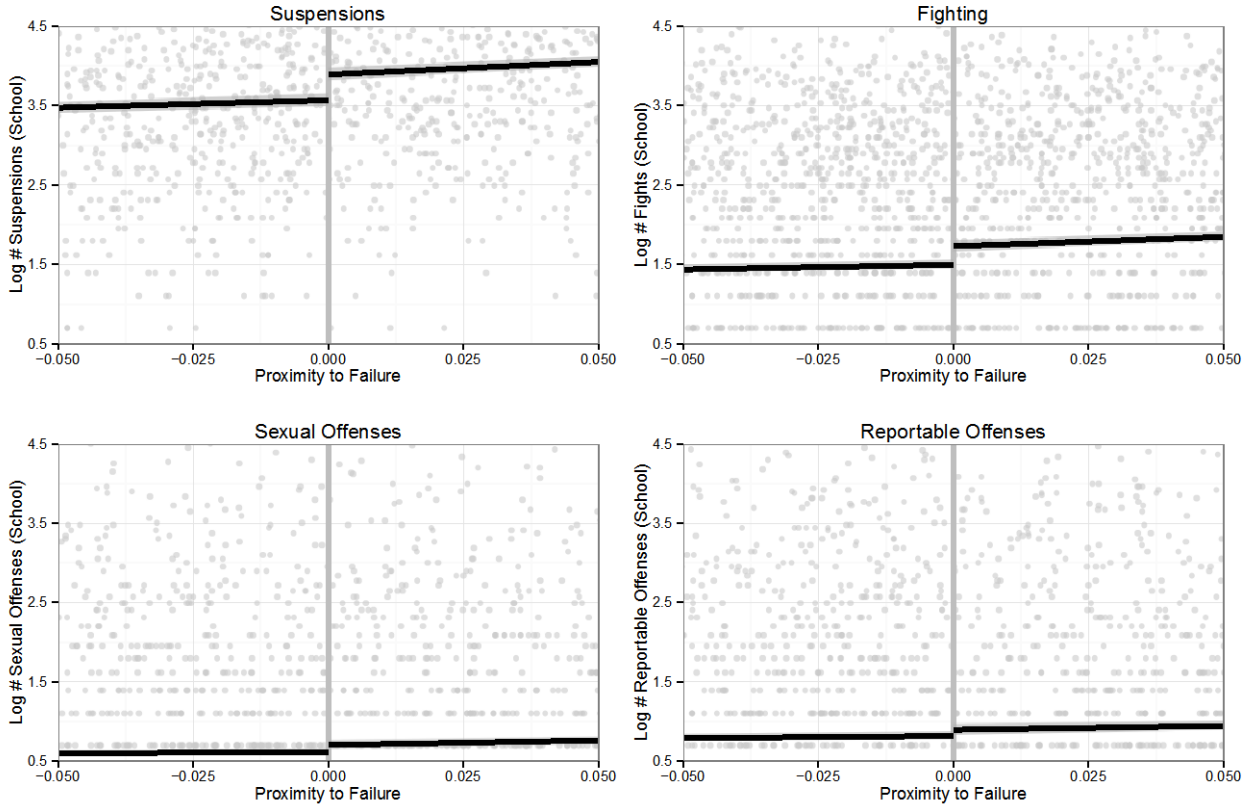
Our models indicate that failure causes an increase in the number of out-of-school suspensions by about 16% on average. This increase is equivalent to about a 0.10 σ increase in suspensions, or about 21 more per school, on average. We also see that pressure from failure increases the number of fights, by about 14% on average (≈ 0.08 σ; ≈ 1.4 more per school). In addition, there is some evidence of an increase in the number of sexual offenses (≈ 13% increase; ≈ 0.10 σ; ≈ 5 more per school), possession related offenses (≈ 12% increase; ≈ 0.10 σ; ≈ 0.6 more per school), disruption offenses (≈ 20% increase; ≈ 0.03 σ; ≈ 1 more per school), and offenses reportable to law enforcement (≈ 14% increase; ≈ 0.10 σ; ≈ 13 more per school). Figure 3 visually depicts some of estimates from table 3.

Table 3: The Effect of Accountability Pressure on Externalizing Behaviors

	(1) ISS	(2) OSS	(3) Fights	(4) Possession	(5) Violence	(6) Sexual	(7) Weapons	(8) Disrupt	(9) Falsify	(10) Reported
School Failure (Optimal Bandwidth)	-0.10	0.16***	0.05	0.12	0.04	0.11*	0.04	0.13	-0.01	0.13
	(0.08)	(0.06)	(0.07)	(0.08)	(0.08)	(0.06)	(0.07)	(0.11)	(0.06)	(0.08)
N (students)	4,701,618	4,435,662	5,494,486	5,390,376	5,249,466	5,274,093	5,469,494	5,575,214	5,365,102	5,748,304
N (schools)	9,305	8825	10,920	10,720	10,470	10,511	10,869	11,075	10,673	11,369
School Failure (Half Optimal)	-0.09	0.11	0.05	0.12	0.12	0.13	0.03	0.12	0.03	0.13
	(0.09)	(0.08)	(0.09)	(0.10)	(0.11)	(0.08)	(0.06)	(0.14)	(0.08)	(0.09)
N (students)	3,875,550	3,131,819	4,047,449	3,801,247	3,554,589	3,583,256	3,980,406	4,260,554	3,739,590	4,847,061
N (schools)	7,757	6,275	8,148	7,632	7,116	7,188	8,016	8,573	7,492	9,712
School Failure (Twice Optimal)	-0.08	0.16***	0.12*	0.13**	0.08	0.13**	0.05	0.19**	0.04	0.14*
	(0.07)	(0.05)	(0.06)	(0.06)	(0.07)	(0.05)	(0.04)	(0.10)	(0.06)	(0.07)
N (students)	4,805,464	4,800,319	5,847,876	5,847,058	5,832,599	5,834,992	5,847,876	5,853,373	5,844,412	5,853,373
N (schools)	9,499	9,476	11,546	11,542	11,518	11,523	11,546	11,570	11,539	11,570
School Failure (Full Bandwidth)	-0.08	0.17***	0.14**	0.12**	0.10	0.13**	0.06	0.20**	0.05	0.14*
	(0.07)	(0.05)	(0.06)	(0.06)	(0.07)	(0.05)	(0.04)	(0.10)	(0.05)	(0.07)
N (students)	4,805,464	4,805,464	5,853,373	5.853,373	5,853,373	5,853,373	5,853,373	5,853,373	5,853,373	5,853,373
N (schools)	9,499	9,499	11,570	11,570	11,570	11,570	11,570	11,570	11,570	11,570

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Cluster-robust standard errors (at the school-year level) reported below coefficient estimates. Outcomes are logged to deal with skew. Running Variable: local-linear. Optimal bandwidth chosen using the procedure suggested by Imbens & Kalyanaraman (2012). Controls: Those showing imbalance in Table 1 including: pupil/teacher ratio, % migrant, and whether a school failed previously. 1.) *ISS*=In school suspensions. 2.) *OSS*=Out of School Suspensions; 3.) *Fights*=The number of student fights. 4.) *Possession*= drug-related offenses 5.) *Violence*=offenses involving violence 6.) *Sexual* = sexual-related offenses 8.) *Disrupt* = misbehaviors involving disruption during school. 9.) *Falsify* = offenses involving falsification (extortion, theft, etc.).10.) *Reportable*= offenses reported to law enforcement.

Figure 3: Accountability Pressure & Externalizing Behaviors



Notes: In each of the panels of figure 4, individual points represent individual school-year observations. An additional dose of accountability pressure is administered to schools on the right side of the cutoff—where schools fail. The causal effect is the vertical difference between the two color-corresponding lines at the cutoff. Local-linear regression is used to model the relationship between the running variable and the outcome.

Although not all of the estimated coefficients are statistically significant at standard levels, most of them are positive. For instance, our estimates involving violence-related offenses (10% increase), weapons related offenses (6% increase), and falsification offenses (5% increase) provide some additional, albeit much more limited, support for the conclusion that accountability pressure generates more misbehavior.

These findings generally align with our expectations. As a multitasking framework predicts, when administrators are faced with a policy that incentivizes some outcomes but not others, they divert resources towards incented outcomes. In practice this means that school administrators target absences for reduction, and succeed in doing so. This additional attention spills over into the related construct of tardies, with pressure causing students to show up to class on time. Apparently, however,

administrators do not divert resources or are unable to positively move non-incentivized outcomes, such as how students behave when they are in school. The finding that students respond to school-level accountability pressure by misbehaving more than they otherwise would is a clear downside of accountability pressure under NCLB. Increased student misbehavior—regardless of the mechanism driving this increase—may harm student learning by distracting students, increasing stress, and putting additional strain on school officials.

7.3 Refinements and Extensions

An additional way to test the multitasking framework outlined earlier is to utilize variation in reporting requirements under NCLB. Although absence rates are a direct input into the determination of the AYP status for elementary and middle schools in North Carolina, that is not the case for high schools. So far we have grouped both types of schools together. By comparing the effect of accountability pressure in schools where "showing up" is incentivized to those where it is not, we can explore the multitasking framework in more detail. Theory would predict that absences would decline in elementary and middle, but not in high schools.

Table 4 makes this comparison. In the first row, we report results for elementary and middle schools alone[37], and in the second, for high schools. The results indicate that failing causes a noticeable decrease in the number of absences, but only when attendance is measured and incentivized. In contrast, when schools are not required to measure and report attendance, there is little evidence of declines in absences. The estimates are statistically distinct. Thus, for accountability to decrease absences it appears that direct measurement and incentives are needed.

[37] We categorize schools according the number of grades provided in the school ranges (Elementary: K-5, Middle: 6-8, High: 9-12). When there are ties in the number of grades provided, we categorize the school as the higher category.

Table 4: Being in School, Across School-Levels

	(1) DV: Absences
Optimal BW	
Failure in Elementary/Middle Schools (Attendance Incentivized)	-0.244*
	(0.149)
Failure in High Schools (Attendance Not Incentivized)	-0.032
	(0.094)
P ($\beta_{elem./middle} = \beta_{high}$)	0.07
Half Optimal BW	
Failure in Elementary/Middle Schools (Attendance Incentivized)	-0.135
	(0.187)
Failure in High Schools (Attendance Not Incentivized)	-0.083
	(0.112)
P ($\beta_{elem./middle} = \beta_{high}$)	0.73
Twice Optimal BW	
Failure in Elementary/Middle Schools (Attendance Incentivized)	-0.403***
	(0.128)
Failure in High Schools (Attendance Not Incentivized)	-0.110
	(0.091)
P ($\beta_{elem./middle} = \beta_{high}$)	0.00
Full BW	
Failure in Elementary/Middle Schools (Attendance Incentivized)	-0.448***
	(0.119)
Failure in High Schools (Attendance Not Incentivized)	-0.086
	(0.099)
P ($\beta_{elem./middle} = \beta_{high}$)	0.00

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Cluster-robust standard errors below coefficient estimates. below coefficients. Running Variable: local-linear. Bandwidth: full. Controls: Those showing imbalance in table 1 including: pupil/teacher ratio, % migrant, and whether a school failed previously. P-values of coefficient differences are based on the Wu-Hausman technique (Wu 1973; Hausman 1978).

As another means of testing the multitasking framework, we leverage variation in exposure to sanctions. As was mentioned earlier, in addition to labeling schools as failing, NCLB applies sanctions to schools that repeatedly fail. Schools do not begin sanctions until they fail twice consecutively. Thus, while those at failing schools feel accountability pressure (due to the anticipation of sanctions), those that fall under sanctions receive an even greater dose of pressure. As incentives increase in their salience and influence, we would expect larger gains on the incentivized metric of absences.

Table 5 shows this comparison of how accountability pressure affects "showing up" and externalizing behavior between schools with and without sanctions immediately at stake. There is some evidence that increasing accountability pressure leads to larger declines in absences. Most of the decline we estimated in absences comes from schools that are under some sanctions, as opposed to schools that fail when sanctions will not be at stake. This pattern also emerges for some of our externalizing behaviors. Out- of- school suspensions and reportable offenses see the largest increases when schools marginally fail and sanctions ensue. However, sexual and falsification offenses are in the opposite direction than what we would predict.

Table 5: Accountability Pressure, by School Sanction Status

	(1) Absences	(2) Tardies	(3) ISS	(4) OSS	(5) Fights	(6) Possession	(7) Violence	(8) Sexual	(9) Weapon	(10) Disruptive	(11) Falsification	(12) Reportable
Fail (sanctions)	-0.97***	-0.26**	-0.13	0.16**	0.00	0.12	0.00	0.07	0.04	0.17	-0.09	0.14
	(0.23)	(0.13)	(0.09)	(0.07)	(0.09)	(0.09)	(0.10)	(0.08)	(0.05)	(0.13)	(0.08)	(0.11)
Fail (no sanctions)	-0.40	-0.27	-0.07	0.00	0.09	0.07	0.10	0.15*	0.08	-0.02	0.03	-0.03
	(0.34)	(0.19)	(0.14)	(0.11)	(0.12)	(0.11)	(0.13)	(0.08)	(0.08)	(0.19)	(0.11)	(0.13)
P ($\beta_{sanction} = \beta_{nosanction}$)	0.02	0.94	0.58	0.08	0.27	0.35	0.21	0.00	0.46	0.17	0.09	0.03

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Cluster-robust standard errors below coefficient estimates. Running Variable: local-linear, Bandwidth: optimal, Controls: Those showing imbalance in table 1 including: pupil/teacher ratio, % migrant, whether the school failed previously. P-values of coefficient differences are based on the Wu-Hausman technique (Wu 1973; Hausman 1978).

To recap, accountability pressure produces gains in attendance-related metrics, but only at the elementary and middle school levels for which attendance is incentivized and mostly when sanctions are immediately at stake. With respect to misbehavior, there is mixed evidence that the additional pressure that comes from immediate sanctioning plays a role in increasing misbehavior.

7.4 Subgroup Differences

In this section we explore whether the effects of accountability pressure differ across groups of students, defined primarily by their previous performance and racial statues. We might expect to observe variation in effects across groups of students whenever there are group-based differences in the abilities of students to process accountability pressure. These comparisons are directly relevant to policy discussion about NCLB, which specifically place accountability pressure on subgroups of students.

We emphasize that in contrast to the previous results that can be interpreted as causal, the heterogeneities discussed in this section cannot be interpreted as fully causal. Unlike AYP failure at the margin, our subjects are not randomly or as-good-as randomly assigned to the student characteristics we explore. As such, we are careful to describe these findings as suggestive of subgroup differences in responsiveness to accountability pressure. Future testing is required to determine whether these subgroup results are indeed causally robust.

Still, these differences are informative. Below we show that accountability pressure appears to differ by previous levels of student performance and by race/ethnicity. In addition to differences across

these subgroups, we have tested for differences by student gender and socioeconomic status. For the most part, we find no statistically significant differences by across gender and SES.

7.4.1 Accountability Pressure by Student Academic Performance

To examine potential differential effects on students grouped by their previous end of grade (EOG) test score performance, we focus on performance in reading, but our results are similar for math performance. Such an analysis restricts our sample to students in grades three through eight because of the annual testing in those grades. Table 5 shows our results.

On metrics of "showing up", we find that schools find ways to reduce absences among all performance subgroups. When schools face accountability pressure, they pressure their students to come to school. This reduction varies, however, in its substantive significance based on the underlying performance of students. Higher performing students in the top test score quartile reduce their absences at a rate more than twice as high as students in the bottom quartile. The difference between these two groups is statistically and substantively meaningful. This pattern also holds with the comparison between the top and the second to the top and the third quartiles. Failure induced accountability pressure encourages attendance, but perhaps unequally so. Schools targeting resources towards higher performing students may explain this finding. Or, higher performing students may be more malleable at baseline, perhaps given lower baseline absenteeism, to improve their attendance patterns. Either way, this finding is important and suggests that the potential benefits of accountability pressure may distribute themselves unequally in a way that leaves low-performing children behind.

A similar, but less pronounced, pattern emerges with tardies. When schools face accountability pressure from failing, there is some evidence that higher performing students are more responsive. The differences are smaller for tardies than absences, but are still substantively meaningful. Students in the top quartile reduce their absences at a rate 11 percentage points higher than third quartile, 20 percentage points higher than the second quartile, and 16 percentage points higher than the lowest

quartile. Unfortunately, we are not able to detect these differences with statistical precision at traditional levels, except between second and the top quartile (p=0.04). Still, there is some evidence that accountability pressure affects higher performing students in a similar way as it does for absences. When schools face accountability pressure, their students are more likely to show up and be on time, but higher performing students improve their behavior more than lower performing students.

With respect to measures of externalizing behavior we see an interesting, u-shaped pattern. Students at the bottom and the top of the test score distribution see notable increases in misbehavior, while those in the middle quartiles do not. For students in the lowest quartile of student performance, misbehaviors in the form in-school suspensions, out-of school suspensions, and violence-related offenses all increase in statistically significant and substantively important ways. For the highest quartile, misbehaviors in the form of out-of school suspensions, sexual-offenses, disruptive offenses, and those offenses reportable to law enforcement all rise in response to accountability pressure. In the middle two quartiles, we see very few estimates that approach a substantive size or statistical significance. Comparing estimates, we can see that on seven of the ten metrics of student misbehavior the top and bottom quartiles depart from the middle quartiles. In short, when schools face accountability pressure, the highest and lowest performing students are the groups most likely to be negatively affected.

What explains this u-shaped pattern? While compelling causal evidence explaining this phenomenon is difficult to bring to bear, the explanation may have something to do with the unequal distribution of accountability pressure and inequalities in students' capacity for improvement of their non-achievement behavior. According to this view, when schools fail all students face pressure. Those at the top, however, face a greater level of pressure because school officials target higher performing students as a way to raise their school's performance, both by reducing absenteeism and potentially

raising test scores. As such, higher performing students face a higher level of pressure to perform. In response they act out negatively by manifesting externalizing behaviors.

Students at the bottom of the test score distribution face a lower level of pressure to show up and to do so on time than their higher performing counterparts. Yet, these students have a lower capacity—because of having lower resources—to deal with such pressure beneficially. These students, thus, react negatively to the increased anxiety and stress that comes when school failure introduces a higher level of accountability pressure. In contrast, students in the two middle quartiles respond more favorably because they face lower levels of pressure than those students at the top of the distribution and have supportive structures that help them cope with increased levels of stress and anxiety.

Table 5: Accountability Pressure, by Student Performance

	(1) Absences	(2) Tardies	(3) ISS	(4) OSS	(5) Fights	(6) Possession	(7) Violence	(8) Sexual	(9) Weapons	(10) Disruptive	(11) Falsification	(12) Reportable
Failure (Lowest Quartile)	-0.32***	-0.12	0.14*	0.26***	0.20***	-0.01	0.18***	0.00	-0.01	0.09	-0.01	0.04
	(0.10)	(0.12)	(0.07)	(0.07)	(0.07)	(0.04)	(0.07)	(0.05)	(0.01)	(0.09)	(0.05)	(0.07)
Failure (Second Quartile)	-0.46***	-0.08	-0.05	0.08	0.08	-0.02	0.02	0.01	0.00	-0.08	-0.07	-0.07
	(0.09)	(0.11)	(0.07)	(0.06)	(0.05)	(0.03)	(0.06)	(0.04)	(0.00)	(0.07)	(0.04)	(0.05)
Failure (Third Quartile)	-0.46***	-0.17	-0.09	-0.06	-0.03	-0.01	-0.09	-0.01	0.00	-0.10	-0.07*	-0.06
	(0.11)	(0.12)	(0.07)	(0.06)	(0.05)	(0.02)	(0.06)	(0.04)	(0.00)	(0.08)	(0.04)	(0.05)
Failure (Top Quartile)	-0.76***	-0.28***	0.00	0.28***	0.11	0.14	0.10	0.15*	0.05	0.25**	0.11	0.19*
	(0.14)	(0.06)	(0.10)	(0.08)	(0.07)	(0.09)	(0.07)	(0.09)	(0.03)	(0.11)	(0.08)	(0.11)
P ($\beta_{lowest} = \beta_{second}$)	0.00	0.42	0.00	0.00	0.00	0.57	0.00	0.67	0.21	0.00	0.05	0.01
P ($\beta_{lowest} = \beta_{third}$)	0.00	0.12	0.00	0.00	0.00	0.98	0.00	0.67	0.34	0.00	0.06	0.03
P ($\beta_{lowest} = \beta_{top}$)	0.00	0.13	0.04	0.62	0.00	0.07	0.00	0.03	0.04	0.01	0.04	0.08
P ($\beta_{second} = \beta_{third}$)	0.77	0.12	0.02	0.00	0.00	0.43	0.00	0.08	0.52	0.15	0.93	0.44
P ($\beta_{second} = \beta_{top}$)	0.00	0.04	0.62	0.00	0.56	0.06	0.04	0.06	0.10	0.00	0.01	0.01
P ($\beta_{third} = \beta_{top}$)	0.00	0.30	0.25	0.00	0.01	0.09	0.00	0.03	0.09	0.00	0.01	0.01

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Cluster-robust standard errors below coefficient estimates. Running Variable: local-linear, Bandwidth: optimal, Controls: Those showing imbalance in table 1 including: pupil/teacher ratio, % migrant, whether the school failed previously. P-values of coefficient differences are based on the Wu-Hausman technique (Wu 1973; Hausman 1978).

7.4.2 Accountability Pressure by Race and Ethnicity

Table 6 shows the effect of accountability pressure by student race/ethnicity subgroups. It shows that accountability reduces absences most for white students. Although African American and Hispanic students still exhibit some reductions, they are not as large as the reduction that emerges for white students. This difference is statistically distinct from zero at the 95% confidence level. When it comes to tardies, however, the reaction is similar across the subgroups—white, African American, and Hispanic students all see declines in tardies, on average.

With respect to reported student misbehaviors, the evidence suggests that the minorities, and especially black students, respond to pressure by increasing their misbehavior more than white students. This difference is clearer in models using wider bandwidths, where our subgroup estimates are more precisely estimated, but holds in directionality and in substantive meaning for the narrower bandwidths as well. With respect to in-school suspensions, accountability pressure appears to reduce the problem for white students but not for African American and Hispanic students. For many of other behavior categories, including ,for example, fighting, violent offenses, disruptive behavior and reportable offences, the evidence suggest that accountability is associated with higher rates of misbehavior for black students than for Hispanic or white students.

Table 6: Accountability Pressure by Race

	(1) Absences	(2) Tardies	(3) ISS	(4) OSS	(5) Fights	(6) Possession	(7) Violence	(8) Sexual	(9) Weapons	(10) Disruptive	(11) Falsification	(12) Reportable
Optimal Bandwidth												
Failure (black)	-0.60***	-0.23*	-0.08	0.14	0.00	0.10	0.05	0.10	0.01	0.10	0.04	0.14
	(0.23)	(0.13)	(0.11)	(0.09)	(0.10)	(0.10)	(0.09)	(0.08)	(0.02)	(0.14)	(0.09)	(0.13)
Failure (white)	-0.75***	-0.21*	-0.21**	0.07	-0.06	0.10	-0.04	0.08	0.01	0.00	-0.10	0.04
	(0.21)	(0.12)	(0.09)	(0.07)	(0.08)	(0.10)	(0.08)	(0.07)	(0.01)	(0.10)	(0.07)	(0.11)
Failure (hispanic)	-0.36***	-0.13	-0.11	0.09	-0.03	-0.05	-0.08	0.05	0.00	0.05	-0.04	0.13*
	(0.20)	(0.11)	(0.10)	(0.08)	(0.08)	(0.04)	(0.09)	(0.08)	(0.00)	(0.13)	(0.06)	(0.07)
P ($\beta_{black} = \beta_{white}$)	0.04	0.75	0.06	0.28	0.39	0.82	0.10	0.67	0.96	0.32	0.02	0.14
P ($\beta_{Hispanic} = \beta_{white}$)	0.00	0.17	0.00	0.71	0.00	0.11	0.35	0.37	0.30	0.50	0.13	0.27
P ($\beta_{black} = \beta_{Hispanic}$)	0.02	0.24	0.67	0.19	0.68	0.11	0.00	0.23	0.52	0.46	0.30	0.92
Full Bandwidth												
Failure (black)	-0.76***	-0.33***	0.00	0.20***	0.18**	0.16**	0.16**	0.10	0.01	0.23**	0.08	0.17*
	(0.18)	(0.10)	(0.09)	(0.07)	(0.07)	(0.08)	(0.07)	(0.07)	(0.03)	(0.10)	(0.07)	(0.09)
Failure (white)	-1.07***	-0.34***	-0.22**	0.01	-0.06	0.08	-0.09	0.05	-0.01	0.02	-0.08	-0.01
	(0.18)	(0.10)	(0.08)	(0.06)	(0.06)	(0.08)	(0.06)	(0.07)	(0.01)	(0.10)	(0.06)	(0.09)
Failure (hispanic)	-0.74***	-0.18**	-0.02	0.16**	0.02	0.00	-0.02	0.07	0.00	0.15	-0.01	0.09*
	(0.15)	(0.08)	(0.08)	(0.07)	(0.05)	(0.03)	(0.06)	(0.05)	(0.00)	(0.09)	(0.04)	(0.05)
P ($\beta_{black} = \beta_{white}$)	0.00	0.63	0.00	0.00	0.00	0.00	0.00	0.04	0.44	0.00	0.00	0.00
P ($\beta_{Hispanic} = \beta_{white}$)	0.00	0.00	0.00	0.00	0.00	0.28	0.00	0.53	0.34	0.00	0.07	0.17
P ($\beta_{black} = \beta_{Hispanic}$)	0.77	0.01	0.61	0.19	0.00	0.03	0.00	0.59	0.62	0.07	0.11	0.28

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Running Variable: local-linear, Controls: Those showing imbalance in table 1 including: pupil/teacher ratio, % migrant, whether the school failed previously. P-values of coefficient differences are based on the Wu-Hausman technique (Wu 1973; Hausman 1978).

8. Conclusion

This study provides evidence that school-level accountability pressure from NCLB affects students in various ways that have been understudied in the literature. We show, first, that accountability pressure improves student behaviors that are directly or tangentially incentivized, namely, improving students' ability to "be where they are supposed to be when they are supposed to be there." On our behavioral proxies for these behaviors —absences and tardies—accountability pressure appears to be beneficial. As a multitasking framework would predict, these positive effects emerge only in elementary and middle schools, which face incentives to reduce absentee rates, and not in high schools, where they are not incentivized. Moreover, as would be expected, the estimated effects are typically larger in failing schools that are under sanction than in those that have not previously failed.

Consistent with our multitasking framework, we have also shown that accountability pressure generates increases in anti-social behaviors on the part of students. Thus accountability pressure at the school level is transferred down to the student level, and not always in positive ways, presumably because schools devote time and resources to improving incentivized behaviors at the expense of ignoring other behaviors. Following school failure, schools experience noticeable increases in misbehaviors that lead to suspensions, sexual offenses, and reportable offenses that cannot be attributed to other aspects of the school. Further, we have shown that changes in misbehavior are exacerbated among minority and low-performing students—those that supporters of No Child Left Behind explicitly hoped would not be left behind.

Future research on school accountability programs would do well to explore other behaviors not directly incentivized by such programs. While direct incentives may improve easy-to-monitor variables such as absences and tardies, this study shows that such programs may do harm by increasing student misbehavior in school. A more complete understanding of how performance-based accountability

programs such as NCLB—the main goal of which is to raise student achievement—affect other student behaviors would help policymakers weigh any positive outcomes of such programs against the potential costs of damaging certain non-achievement behaviors vital for success in school and beyond. Our work shows that such policies deserve some hope but also some pause, given their potentially harmful effects on student attributes not captured by student test scores.

References

Ahn, T., & Vigdor, J. (2014a). The Impact of No Child Left Behind's Accountability Sanctions on School Performance: Regression Discontinuity Evidence from North Carolina. (No. w20511). National Bureau of Economic Research.

Ahn, T., & Vigdor, J. L. (2014b). When Incentives Matter Too Much: Explaining Significant Responses to Irrelevant Information (No. w20321). National Bureau of Economic Research.

Angrist, J. D., Pathak, P. A., & Walters, C. R. (2011). Explaining charter school effectiveness (No. w17332). National Bureau of Economic Research.

Bacolod, M., DiNardo, J., & Jacobson, M. (2009). Beyond Incentives: Do Schools Use Accountability Rewards Productively? (No. w14775). National Bureau of Economic Research.

Benda, B. B. (2005). The robustness of self-control in relation to form of delinquency. Youth & Society, 36(4), 418-444.

Butler, D. M., & Butler, M. J. (2006). Splitting the difference? Causal inference and theories of split-party delegations. Political Analysis, 14(4), 439-455.

Calonico, S., Cattaneo, M. D., & Titiunik, R. (2013). Robust data-driven inference in the regression-discontinuity design. The Stata Journal, ii, 1-34.

Carneiro, P., Hansen, K. T., & Heckman, J. J. (2003). Estimating Distributions of Treatment Effects with an Application to the Returns to Schooling and Measurement of the Effects of Uncertainty on College (No. w9546). National Bureau of Economic Research.

Carneiro, P., Crawford, C., & Goodman, A. (2007). The impact of early cognitive and non-cognitive skills on later outcomes.

Chiang, H. (2009). How accountability pressure on failing schools affects student achievement. Journal of Public Economics, 93(9), 1045-1057.

Cook, P. J., Dodge, K., Farkas, G., Fryer Jr, R. G., Guryan, J., Ludwig, J., Mayer, S., Pollack, H., & Steinberg, L. (2014). The (Surprising) Efficacy of Academic and Behavioral Intervention with Disadvantaged Youth: Results from a Randomized Experiment in Chicago (No. w19862). National Bureau of Economic Research.

Cullen, J. B., & Reback, R. (2006). Tinkering toward accolades: School gaming under a performance accountability system (Vol. 14, pp. 1-34). Emerald Group Publishing Limited.

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

Duckworth, A. L., Peterson, C., Matthews, M. D., & Kelly, D. R. (2007). Grit: perseverance and passion for long-term goals. Journal of personality and social psychology, 92(6), 1087.

Eaton, D. K., Kann, L., Kinchen, S., Shanklin, S., Flint, K. H., Hawkins, J., & Wechsler, H. (2012). Youth risk behavior surveillance-United States, 2011. Morbidity and mortality weekly report. Surveillance summaries (Washington, DC: 2002), 61(4), 1-162.

Feng, L., Figlio, D. N., & Sass, T. (2010). School accountability and teacher mobility (No. w16070). National Bureau of Economic Research.

Figlio, D. N., & Getzler, L. S. (2002). Accountability, ability and disability: Gaming the system (No. w9307). National Bureau of Economic Research.

Figlio, D. N., & Winicki, J. (2005). Food for thought: the effects of school accountability plans on school nutrition. Journal of public Economics, 89(2), 381-394.

Flay, B. R., Graumlich, S., Segawa, E., Burns, J. L., & Holliday, M. Y. (2004). Effects of 2 prevention programs on high-risk behaviors among African American youth: a randomized trial. Archives of Pediatrics & Adolescent Medicine, 158(4), 377-384.

Fryer Jr, R. G., & Holden, R. T. (2012). Multitasking, Learning, and Incentives: A Cautionary Tale (No. w17752). National Bureau of Economic Research.

Gibbons, R. (1998). Incentives in Organizations. Journal of Economic Perspectives, 12(4), 115-132.

Gilman, R., Dooley, J. & Florell, D. (2006). Relative levels of hope and their relationship with academic and psychological indicators on adolescents. Journal of Social and Clinical Psychology, 25, 166-178.

Grunbaum, J. A., Kann, L., Kinchen, S., Ross, J., Hawkins, J., Lowry, R., & Collins, J. (2004). Youth risk behavior surveillance--United States, 2003. Morbidity and mortality weekly report. Surveillance summaries (Washington, DC: 2002), 53(2), 1-96.

Hahn, J., Todd, P., & Van der Klaauw, W. (2001). Identification and estimation of treatment effects with a regression- discontinuity design. *Econometrica*, *69*(1), 201-209.

Hausman, J. A. (1978). Specification tests in econometrics. Econometrica: Journal of the Econometric Society, 1251-1271.

Heine, S. J., Buchtel, E. E., & Norenzayan, A. (2008). What do cross-national comparisons of personality traits tell us? The case of conscientiousness. Psychological Science, 19(4), 309-313.

Heckman, J. J. (2000). Policies to foster human capital. Research in economics, 54(1): 3-56.

Heckman, J. J., & Rubinstein, Y. (2001). The importance of noncognitive skills: Lessons from the GED testing program. American Economic Review, 145-149.

Heckman, J. J., Stixrud, J., & Urzua, S. (2006). The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior (No. w12006). National Bureau of Economic Research.

Heckman, J. J., Humphries, J. E., Urzua, S., & Veramendi, G. (2011). The effects of educational choices on labor market, health, and social outcomes. Unpublished manuscript, University of Chicago, Department of Economics.

Heckman, J. J., & Kautz, T. (2013). Fostering and measuring skills: Interventions that improve character and cognition (No. w19656). National Bureau of Economic Research.

Heckman, J., Pinto, R., & Savelyev, P. (2013). Understanding the Mechanisms through Which an Influential Early Childhood Program Boosted Adult Outcomes. American Economic Review, 103(6), 2052-86.

Hoffman, J. V., Assaf, L. C., & Paris, S. G. (2001). High-stakes testing in reading: Today in Texas, tomorrow?. *The Reading Teacher*, 482-492.

Holbein, J. (2014). Left Behind: Does Performance Information Promote Democratic Accountability? APSA 2014 Annual Meeting Paper.

Holmstrom, B., & Milgrom, P. (1991). Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. JL Econ. & Org., 7, 24.

Imbens, G., & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. Journal of Econometrics, 142(2), 615-635.

Imbens, G, & Zajonc, T. (2011). Regression discontinuity design with multiple forcing variables. Report, Harvard University [972].

Imbens, G., & Kalyanaraman, K. (2012). Optimal bandwidth choice for the regression discontinuity estimator. The Review of Economic Studies, 79(3), 933-959.

Jacob, B. A. (2002). Where the boys aren't: non-cognitive skills, returns to school and the gender gap in higher education. Economics of Education review, 21(6), 589-598.

Jacob, B. A., & Lefgren, L. (2004). Remedial education and student achievement: A regression-discontinuity analysis. Review of economics and statistics, 86(1), 226-244.

Jacob, B. A. (2005). Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago Public Schools. Journal of public Economics, 89(5), 761-796.

Jackson, C. K. (2013). Non-cognitive ability, test scores, and teacher quality: Evidence from 9th grade teachers in North Carolina.

Jones, B. D. (2007). The unintended outcomes of high-stakes testing. Journal of applied school psychology, 23(2), 65-86.

Kinsler, J. (2013). School discipline: a source of salve for the racial achievement gap? *International Economic Review*, *54*(1), 355-383.

Ladd, G. W. (2005). Children's peer relations and social competence: A century of progress. Yale University Press.

Ladd, H. F. (1999). The Dallas school accountability and incentive program: an evaluation of its impacts on student outcomes. Economics of Education Review, 18(1), 1-16.

Ladd, H. F., & Zelli, A. (2002). School-based accountability in North Carolina: The responses of school principals. Educational Administration Quarterly, 38(4), 494-529.

Laffont, J. J., & Martimort, D. (2009). The theory of incentives: the principal-agent model. Princeton University Press.

Lee, D. S., & Lemieux, T. (2010). Regression Discontinuity Designs in Economics. Journal of Economic Literature, 48, 281-355.

Lemieux, T., & Milligan, K. (2008). Incentive effects of social assistance: A regression discontinuity approach. Journal of Econometrics, 142(2), 807-828.

Lyubomirsky, S., King, L. & Diener, E. (2005). The benefits of frequent positive affect: Does happiness lead to success? Psychological Bulletin, 131, 803-855.

Matsudaira, J. D. (2008). Mandatory summer school and student achievement. Journal of Econometrics, 142(2), 829-850.

McNeil, L. M., Coppola, E., Radigan, J., & Heilig, J. V. (2008). Avoidable losses: High-stakes accountability and the dropout crisis. Education policy analysis archives, 16(3).

Neal, D., & Schanzenbach, D. W. (2010). Left behind by design: Proficiency counts and test-based accountability. The Review of Economics and Statistics, 92(2), 263-283.

Nichols, A. (2012). rd: Stata module for regression discontinuity estimation. Statistical Software Components.

Papay, J. P., Willett, J. B., & Murnane, R. J. (2011). Extending the regression-discontinuity approach to multiple assignment variables. Journal of Econometrics, 161(2), 203-207.

Pratt, T. C., & Cullen, F. T. (2000). The empirical status of Gottfredson and Hirschi's general theory of crime: A meta-analysis. Criminology, 38(3), 931-964.

Reardon, S. F., & Robinson, J. P. (2012). Regression discontinuity designs with multiple rating-score variables. Journal of Research on Educational Effectiveness, 5(1), 83-104.

Reback, R., Rockoff, J., & Schwartz, H. L. (2011). Under pressure: Job security, resource allocation, and productivity in schools under NCLB (No. w16745). National Bureau of Economic Research.

Rouse, C. E., Hannaway, J., Goldhaber, D., & Figlio, D. (2007). Feeling the Florida heat? How low-performing schools respond to voucher and accountability pressure (No. w13681). National Bureau of Economic Research.

Stecher, B. M., Camm, F., Damberg, C. L., Hamilton, L. S., Mullen, K. J., Nelson, C., & Leuschner, K. J. (2010). Toward a culture of consequences. Rand Corporation.

Tomer, J. F. (2003). Personal capital and emotional intelligence: an increasingly important intangible source of economic growth. Eastern Economic Journal, 453-470.

Traczynski, J. and Fruehwirth, J. C. (2014). Spare the Rod? The Dynamic Effects of No Child Left Behind on Failing Schools. Working paper.

Wheelock, A., Haney, W., & Bebell, D. (2000). What can student drawings tell us about high-stakes testing in Massachusetts? The Teachers College Record.

Whitney, C. (2013). The Effects of No Child Left Behind On Children's Socio-Emotional Outcomes. APPAM 2013 Paper.

Wong, V. C., Steiner, P. M., & Cook, T. D. (2013). Analyzing Regression-Discontinuity Designs With Multiple Assignment Variables A Comparative Study of Four Estimation Methods. Journal of Educational and Behavioral Statistics, 38(2), 107-141.

Worrell, F. C. & Hale, R. L. (2001). The relationship of hope in the future and perceived school climate to school completion. School Psychology Quarterly, 16, 370-388.

Wu, D. M. (1973). Alternative tests of independence between stochastic regressors and disturbances. Econometrica: journal of the Econometric Society, 733-750.

Appendix

Descriptive Statistics

Table A1 includes basic summary statistics for our outcome measures. It is important to note that in the text we used logged measures of these outcomes. Here we present the unlogged count versions for simplicity of interpretation.

Table A1: Outcome Measures Descriptive Statistics

(1) Variable	(2) Mean	(3) S.D.
In-school Suspensions	82.6	129.4
Out-school Suspensions	102.7	150.0
Fights	14.1	23.7
Possession-Related Offenses	4.5	12.0
Violence-Related Offenses	32.1	50.7
Sexual-Related Offenses	6.6	29.0
Weapons-Related Offenses	1.1	12.1
Disruption-Related Offenses	126.6	282.0
Falsification-Related Offenses	5.3	10.9
Reportable Offenses	11.5	56.8
Absences	1114	1402
Tardies	174.5	542

Note: Summary Statistics for our Outcomes. Unlike our outcomes, which are logged to address concerns functional form, here we provide raw counts at the school level. These are unweighted

Offense Measures Used

The NCERDC collects information from about 75 offense types. To minimize problems with multiple hypothesis testing and issues with varying reporting requirements, the individual measures were included into eight constructs based on measure similarity. Table A.1 shows the individual offenses used to construct the eight misbehavior measures used in the paper.

Table A2: Offense Measures

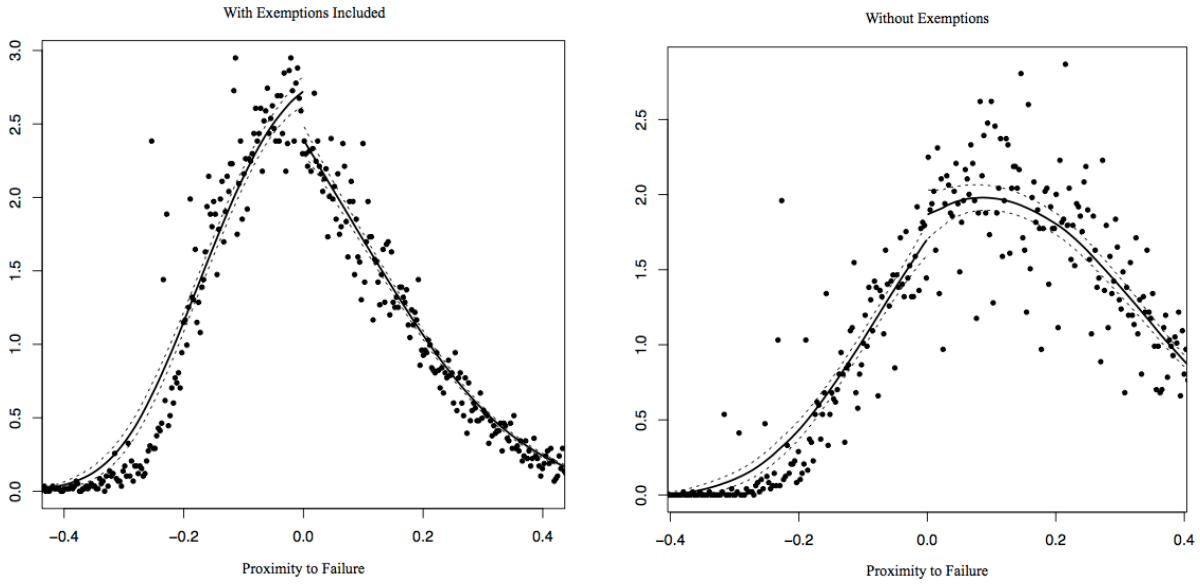
(1) "Possession"	(2) "Violence"	(3) "Sexual"	(4) "Weapon"	(5) "Disruptive"	(6) "Falsification"	(7) "Reportable"
Possession of alcoholic beverage, Possession of controlled substance in violation of law, Possession of tobacco, Possession of a prescription drug, Distribution of a prescription drug, Use of controlled substances, Use of alcoholic beverages, Use of narcotics, Inappropriate items on school property	Assault resulting in serious injury, Assault on school personnel (not resulting in serious injury), Assault involving use of a weapon, Communicating Threats, Assault on student w/o weapon and not resulting in injury, Assault on non-student w/o weapon not resulting in injury, Aggressive Behavior, Hazing, Bullying, Kidnapping, Affray	Rape, Taking indecent liberties with a minor, Sexual assault, Sexual offense, Harassment - sexual, Mutual sexual contact between two students, Excessive display of affection, Harassment, Physical Exam, Immunization	Bomb threat, Burning of a school building, Death by other than natural causes, Possession of a firearm or powerful explosive, False fire alarm, Weapon used or possessed, Robbery with a dangerous weapon, Robbery without a dangerous weapon	Inappropriate language or disrespect, Insubordination, Bus misbehavior, Disruptive behavior, Disrespect of faculty/staff, Honor Code violation, Dress code violation, Disorderly conduct, Cell phone use, Gang activity	Falsification of information, Extortion, Property damage, Theft, Possession of counterfeit items, Use of counterfeit items, Gambling	Other School Defined Offense, Offense 1 reported to law enforcement, Offense 2 reported to law enforcement, Legally reportable offense, Other offense resulting in OSS or expulsion

Supplemental Analyses

Precise Sorting

Here we provide the McCrary Density check for precise sorting at the cutoff (2008). As can be seen in Figure 1, there is a small cluster of schools on the passing side of the cutoff (i.e. the left side of the left panel). However, as Ahn and Vigdor note, this cluster is relatively small (2014a). Moreover, Traczynski and Fruehwirth show that this clustering is driven by NCLB's exceptions to passing AYP, not precise manipulation by schools by documenting that the cluster of schools just above passing disappears when the confidence interval and growth exemptions are left out of the running variable determination (N.d. 16, 48). In short, the cluster of marginally passing schools likely comes because NCLB pushes a group of schools above the cutoff by providing exemptions. We replicate their finding here by showing the distribution of the running variable without exemptions on the right panel of the graph. Under this specification, the clustering of schools on the passing side of the cutoff disappears. Given this and our performance on the covariate balance check, it appears that precise manipulation by individual school actors is of minimal concern. The discontinuity still appears to sort schools in an as-good-as random manner.

Figure 1: McCrary Density Test—Running Variable Distribution



Notes: McCrary density test (2008) for precise sorting at the failure cutoff. The left panel shows the check with the distribution of the running variable incorporating NCLB's exemptions (confidence interval and growth). It shows a group of schools just marginally making AYP. The right panel shows the distribution of the running variable when the exemptions are not incorporated. The group of schools just marginally making AYP disappears. This second panel reveals that the cluster on the just passing side of the cutoff is due to the exemptions NCLB grants, not due to precise sorting of schools across the failure cutoff.

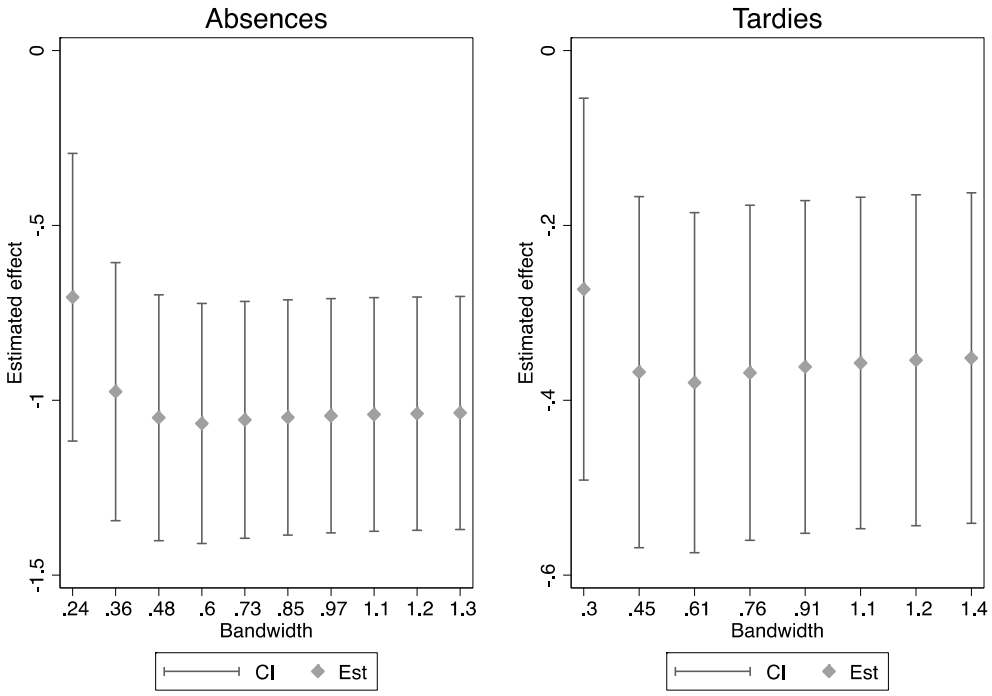
Effects by Bandwidths

A key decision in regression discontinuity models is the bandwidth of data used around the cutoff (Lee & Lemieux 2010). In the paper, we showed the overall estimates for the optimal bandwidth, half the optimal bandwidth, two times the optimal bandwidth. Here we display visually our overall estimates across a wide array of bandwidths. It is important to remember that the bandwidth is divided roughly equally on either side of the failure cutoff.

In the figures A1 and A2 we display the RD coefficient estimates on the log scale (y-axis) by the bandwidth used (x-axis). For each variable, we show bandwidths from the very narrow half the optimal bandwidth (on the far left) to that using the entire range of data (as close as possible, with as similar number of coefficients shown as possible across the graphs).

In general these show that are the results presented in the paper are generally robust to bandwidth selection. For absences and tardies, the estimated effect of pressure is slightly smaller in narrower bandwidths. These estimates, however, still remain negative and statistically distinct from zero.

Figure A1: Estimates of Table 2 by Bandwidth

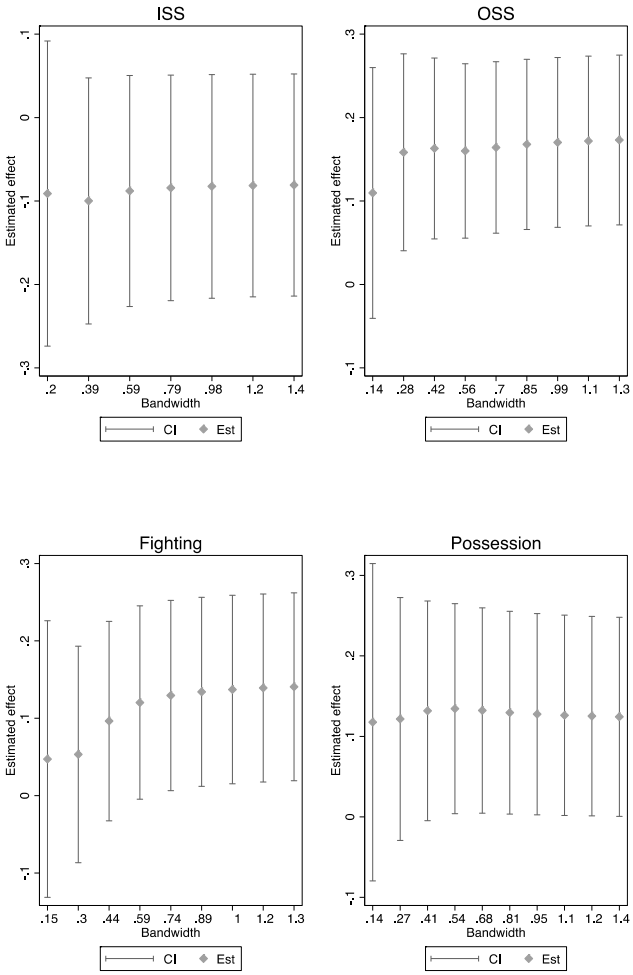


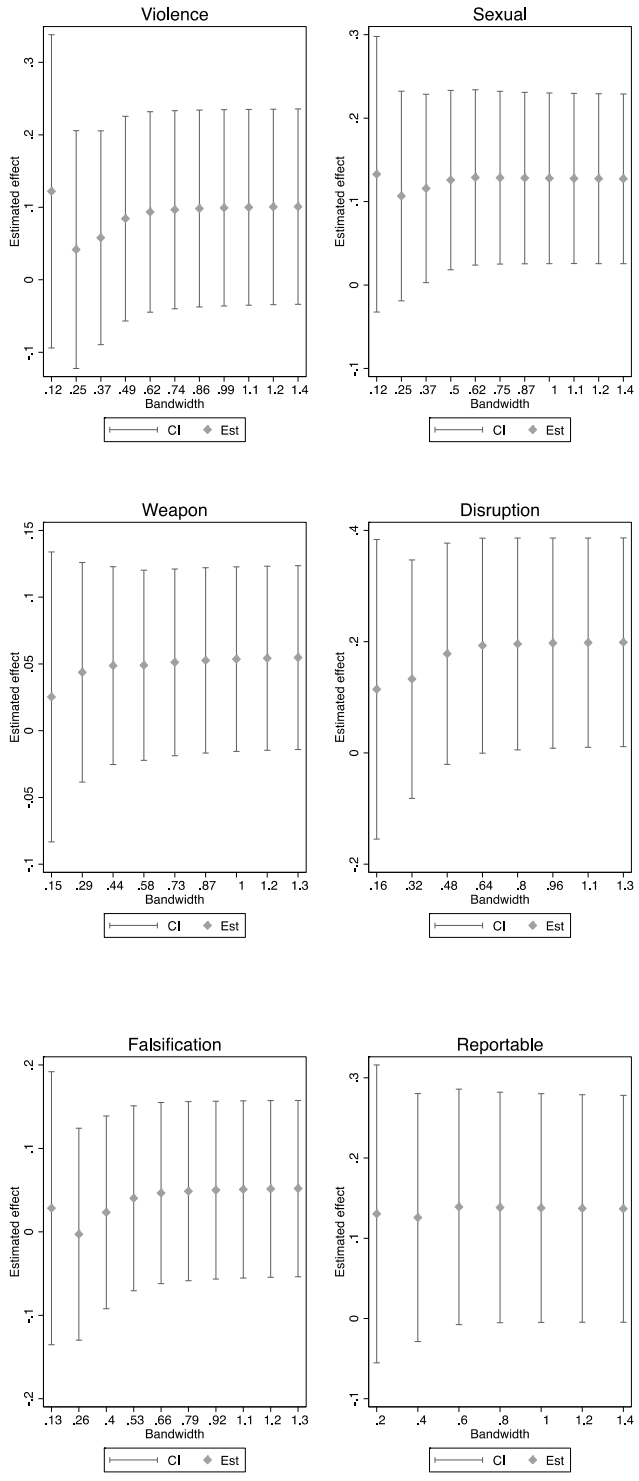
Note: Estimates for table 2 across bandwidths. Like in the paper, imbalanced controls are included. The furthermost left bandwidth is half of the optimal bandwidth suggested by Imbens & Kalyanaraman (2012). The bandwidth at the furthermost right is the full-bandwidth of schools.

A consistent pattern holds across our externalizing behavior outcomes. Out of school suspensions also remain remarkably consistent across the range of bandwidths. With the half of the

optimal bandwidth—7 points on either side of the cutoff—we still produce an estimate that is statistically indistinguishable from the other estimates. That this estimate is not distinct from zero in this very narrow bandwidth likely reflects the low amount of statistical power very close to the failure cutoff. Estimates for fighting are a little noisier, but remain statistically indistinguishable from one another across the bandwidths. Again, as more data is added in wider bandwidths, we are able to be more precise and detect an effect that is substantively meaningful and statistically distinct from zero. Such a pattern also holds with sexual-related offenses, disruption-related offenses, and offenses reportable to law enforcement.

Figure A2: Estimates of Table 3 by Bandwidth





Note: Estimates for table 3 across bandwidths. Like in the paper, imbalanced controls are included. The furthermost left bandwidth is the optimal. The furthermost to the right is the full-bandwidth. These show that narrower bandwidths reveal similar results, with less precision due to small sample sizes.

Alternate Model Specifications

Our preferred models provided in the paper include covariates and student-level frequency weights. This approach allows us to account for the possibility of underlying heterogeneity at the school failure cutoff and to draw inferences to the student-level. There are, however, other approaches for modeling the effect of failure on our outcomes at the discontinuity. Tables A2 and A3 display alternate specifications for tables 2 and 3, respectively. These include models that omit controls and weights.

As can be seen below, generally speaking, the results outlined in the paper are robust. In some instances, the inferences change according to the basic parameters used in the model. For instance, our estimate for absences is smaller when we do not include student-level weights or controls. However, this estimate is still substantively meaningful, in the same direction, and statistically distinct from zero. On our other 11 outcomes, our estimates are consistent, usually not statistically distinct from one another.

Table A3: Being in School, Alternate Model Specifications

	(1) Absences	(2) Tardies
With controls & weights (in text)	-1.04***	-0.35***
	(0.17)	(0.10)
No controls, with weights	-0.90***	-0.34***
	(0.17)	(0.10)
Controls, no weights	-0.91***	-0.28***
	(0.10)	(0.08)
No weights, no controls	-0.71***	-0.29***
	(0.10)	(0.08)

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Cluster-robust standard errors below coefficient estimates. The table above displays coefficients from a variety of alternate specifications for the models produced in the paper. In the table, the first row displays the coefficients displayed in the paper.

Table A4: Behaving in School, Alternate Model Specifications

	(1) ISS	(2) OSS	(3) Fights	(4) Possession	(5) Violence	(6) Sexual	(7) Weapon	(8) Disruptive	(9) Falsification	(10) Reportable
With controls & weights (in text)	-0.08	0.17***	0.14**	0.12**	0.10	0.13**	0.06	0.20**	0.05	0.14*
	(0.07)	(0.05)	(0.06)	(0.06)	(0.07)	(0.05)	(0.04)	(0.10)	(0.05)	(0.07)
No controls, with weights	-0.12*	0.12**	0.13**	0.10*	0.10	0.12**	0.04	0.18**	0.04	0.09
	(0.07)	(0.04)	(0.06)	(0.06)	(0.06)	(0.05)	(0.03)	(0.09)	(0.05)	(0.06)
Controls, no weights	-0.10*	0.11***	0.14***	0.05	0.11**	0.08***	0.02	0.17***	0.03	0.06
	(0.06)	(0.04)	(0.04)	(0.03)	(0.05)	(0.03)	(0.02)	(0.06)	(0.03)	(0.04)
No weights, no controls	-0.12**	0.10**	0.14***	0.04	0.14***	0.07**	0.01	0.20***	0.04	0.03
	(0.06)	(0.04)	(0.04)	(0.03)	(0.04)	(0.03)	(0.01)	(0.05)	(0.03)	(0.03)

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Cluster-robust standard errors used to create the 95% CI's shown below the coefficient estimates. The table above displays coefficients from a variety of alternate specifications for the models produced in the paper. In the table, the first row displays the coefficients displayed in the paper. Models with fixed effects (denoted FE above) include a school fixed effect. Bias correction that implemented using the "rdrobust" command in Stata as suggested by Calonico et al. (2013).