Economics of Education Review 58 (2017) 55–67

Contents lists available at ScienceDirect

Economics of Education Review

journal homepage: www.elsevier.com/locate/econedurev

Accountability pressure: Regression discontinuity estimates of how No Child Left Behind influenced student behavior

John B. Holbein[a,*], Helen F. Ladd[b]

[a] Postdoctoral Research Fellow, Princeton University, 307 Robertson Hall, Princeton, NJ 08544, USA
[b] Sanford School of Public Policy, Duke University, 214A Sanford Bldg., Durham NC 27708, USA

ARTICLE INFO

Article history:
Received 15 June 2015
Revised 21 March 2017
Accepted 23 March 2017
Available online 29 March 2017

JEL classification:
I00
I28
I21

Keywords:
Performance-based accountability
No Child Left Behind
Non-cognitive skills
Regression discontinuity

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 this form 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 appears to have the unintended effect, however, of increasing the number of student misbehaviors. Further, we find some evidence that this negative response is most pronounced among minorities and low performing students: those who are the most likely to be left behind.

1. Introduction

In recent years, education policy in the U.S. 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. Yet, the many empirical studies evaluating 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 whether accountability pressure has effects on 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 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 misbehave while in school. Specifically, we explore the effect of accountability pressure that originates from schools' failure to make Adequate Yearly Progress (AYP).[1] While it is true that under NCLB all schools—regardless of their performance—face some pressure simply because they are a part of this performance accountability system, at the AYP cutoff marginally failing schools face an added dose of accountability pressure. This added pressure arises because marginally failing schools face a social stigma from being labeled "failing" and because they are exposed to a discrete jump in the likelihood of sanctions in future periods.[2] To identify the causal impact of this specific type of accountability pressure, we use a regression discontinuity design (RDD) that leverages exogenous variation at the arbitrary AYP cutoff. This approach isolates the causal effect of AYP failure apart from other observable and non-observable characteristics of students, schools, classrooms, and communities.

We find that failure-induced accountability pressure produces predictably mixed results for the non-achievement student behav-

* Corresponding author.
E-mail addresses: jholbein@princeton.edu (J.B. Holbein), hladd@duke.edu (H.F. Ladd).

[1] As we describe in further detail below, AYP is the criterion used to categorize schools based on student performance on standardized tests and other academic indicators. Schools that do not make AYP are labeled "failing" and face sanctions if they fail multiple times consecutively.

[2] Indeed, the threat of future sanctions may be particularly salient. For example, in our data for North Carolina described below, we find the increased threat of future sanctions for marginally failing schools is non-negligible in that a school that marginally fails in one year is 7-14 percentage points more likely to face sanctions in the next year ($p < 0.02$) than all-else-equally schools that marginally pass.

http://dx.doi.org/10.1016/j.econedurev.2017.03.005
0272-7757/

iors we examine. On the one hand, when schools face accountability pressure, students respond—as NCLB intends—by showing up to school and doing so on time. On the other, pressure leads to unintended and perhaps undesirable effects on how students behave when they are in school. Our measures of misbehavior include externalizing behavior that lead to suspensions, sexual offenses, and offenses that are required to be reported to law enforcement, among others. In addition to these overall results, we find important heterogeneities across school and student characteristics. Student responses vary depending on the non-achievement measures that NCLB requires schools to report and whether the school is under the direct threat of sanction with one or more previous failures. In addition, we find some evidence 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 appears to potentially harm students in other important ways.

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. Our results provide policymakers with causal evidence that accountability pressure may generate unintended effects on student behaviors outside of what is picked up by standardized tests of academic achievement. 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 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 work informs the growing body of research involving so called 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., Carneiro, Crawford, & Goodman, 2007; Gilman, Dooley, & Florell, 2006; Heckman, 2000; Jacob, 2002; Jackson, 2012). Despite this growing literature, we still know relatively little about the targeted policies that can help nurture—or alternatively, harm—the development of these skills. Our analysis suggests that education policies primarily targeted towards the development of cognitive skills (like NCLB) may also affect non-achievement metrics of student success, sometimes in undesirable ways. Instead of leaving no child behind, performance-based accountability policies appear to sometimes harm and perpetuate inequalities in the attributes not captured by test scores shown to be so important in school and beyond.

2. Background

In recent years, policymakers have implemented performance accountability systems widely, with these now an integral part of health, agriculture, law-enforcement, nonprofit, environment, foreign policy, and education sectors (Stecher et al., 2010). 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 to be reported. Under the second, individual actors' performance results are published. Finally, if the relevant actors fail to meet set standards, they face sanctions or 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" (Dee & Jacob, 2011, 149), with the law substantially altering the education system by implementing universal performance-based accountability. Under this system, student performance is evaluated primarily using student test scores. Schools whose students fail to meet arbitrary performance thresholds that determine their level of "adequate yearly progress" are labeled failing. Additionally, a less-publicized NCLB provision requires schools to measure and report "other academic indicators" (or, OAIs for short), which in many states include attendance or graduation rates.[3] 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 sanction periods, district officials alter schools' leadership structure by removing administrators or implementing school-takeover.[4] 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 are specifically directed at students, such as mandated exit exams or promotion requirements, NCLB applies pressure on schools, not students. As a result, accountability pressure is likely to affect school administrators and teachers most. Nonetheless, pressure may also be likely to spillover to students who, in turn, may react in positive or in negative ways.

2.1. Previous research on accountability pressure

Previous empirical studies of NCLB—and the similar local performance accountability reforms that preceded it—have focused primarily on how accountability pressure affects student test scores. Scholars studying such impacts have used a variety of panel and quasi-experimental techniques that can be divided into two types. The first includes studies that leverage differences *between* systems: comparing schooling units with and without accountability regimes. Analyses of this type are designed to shed light on the "full effect" of accountability pressure on student or school outcomes. However, inherent difficulties of establishing all-else-equal comparison groups make it difficult to identify causal effects with between systems designs. Seeking to address this limitation, studies of the second type compare schools *within* a given performance-accountability system. Under the NCLB context, for example, this approach compares outcomes for students who are in schools that fail to make AYP to those students in schools that do not, under the logic that failing schools face a higher dose of accountability pressure. A major advantage of this approach is that it permits all-else-equal comparisons: allowing us to compare schools as-good-as randomly assigned to failing to those as-good-as randomly assigned to passing. Its primary limitation, however, is that it focuses only on one piece, albeit an important one, of the total accountability pressure generated by a given accountability system.

While a complete review of the studies examining the effects of accountability pressure on student test scores is beyond the scope of this paper, we provide a short overview here, as this work provides important context for the present study. In an early study leveraging a between systems comparison, Ladd (1999) benchmarked students in Dallas' pre-NCLB accountability reforms to stu-

[3] The OAIs we mention are those in place in North Carolina: the state we explore below. More generally, NCLB requires that schools report at least one OAI. However, the law gives states leeway in deciding what measure to use (NCLB, Part A, Subpart 1, Section 111, 2CVii).

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

dents in observationally similar Texas school districts. She found that enhanced levels of accountability pressure in Dallas were associated with test score gains for Hispanic and Caucasian middle school students. Similarly, Jacob (2005) compared trends in the Chicago Public School System—which had also implemented a performance accountability reform before NCLB—to other large midwestern cities that did not institute accountability policies over the same time period. He found that accountability pressure increased test scores in Chicago, under that city's pre-NCLB accountability policy, with important caveats.[5] In a similar study, Neal and Schanzenbach (2010) showed some evidence of test score gains in response to accountability in Chicago, but primarily for those students close to the proficiency cutoff. Finally, Dee and Jacob (2011) used a comparative interrupted time series approach which benchmarked states that had implemented performance accountability reforms prior to No Child Left Behind (i.e. their "control" group) to states that had not (i.e. their "treatment" group). They found that the accountability pressure led to noticeable gains in fourth grade and, possibly eighth grade, math scores but no gains in reading.

Using within-systems comparisons—the approach we use in this study—various scholars have found that accountability pressure, albeit just the component of accountability pressure associated with school failure, also leads to higher test scores. As we mentioned briefly earlier, accountability pressure of this type arises because of the social stigma associated with being labeled as failing and the increased probability of sanctions in current and future time periods.[6] Comparing schools that fell on either side of Florida's letter grade cutoffs, Rouse, Hannaway, Goldhaber, and Figlio (2007) found that accountability pressure increased math scores by 0.06–0.14σ and reading scores by 0.06–0.10σ. Using an analogous approach, Reback, Rockoff, and Schwartz (2011) found modest positive gains in reading, math, and science tests in response to increases in accountability pressure when schools failed to make AYP. Likewise, Ahn and Vigdor (2014a) using a regression discontinuity design and data from North Carolina found positive effects in schools first entering the NCLB sanction regime and for schools entering higher sanction levels—with their estimates being somewhere between 0.05–0.08σ in these years. Finally, Traczynski and Fruehwirth (2014), using a similar approach to Ahn and Vigdor (2014a), found that, indeed, the accountability pressure that comes from marginally failing to make AYP increases student test scores downstream (by about 0.02–0.06σ).[7]

In short, previous work has shown that accountability pressure increases student test scores. This conclusion holds for estimates of the full effects of accountability systems (i.e. between comparisons) and for estimates of the marginal effect associated with school failure (i.e. within comparisons). While debate continues about the policy significance of the magnitude of these effects (with many observers arguing these are, at best, small to modest), whether they are consistently realized across different subgroups, and whether test score gains are worth the potential costs or unintended consequences that come with these, the overall pattern is relatively clear—accountability pressure leads to some increases in student test scores.[8]

2.2. Accountability and non test score outcomes

Evidence from a related body of literature suggests that accountability pressure is sufficient to elicit important changes in how schools are run. Most relevant to the topic at hand, studies have confirmed that the accountability pressure transferred to schools at the AYP failure cutoff elicits a response from school officials. For example, using a RDD Ahn and Vigdor (2014) find that marginal school failure causes higher levels of administrator turnover, perhaps as a result of decreased levels of perceived job security. With a similar approach, Chiang (2009) finds that AYP failure is sufficient to lead to significant changes in school policy, pedagogy, and practice. These studies, along with those on student test scores, suggest that the pressure marginally failing schools face drives them to shift their behavior in important ways.[9] The remaining question is whether such pressure also leads to changes in student behaviors beyond test scores.

Only a few researchers to date have addressed this question, and typically these examinations are only small parts of larger studies. For example, Ladd (1999) concluded that accountability pressure under Dallas' accountability system reduced the dropout rate of high school students; Chiang (2009) briefly examined the impact of Florida's accountability system on absences and disciplinary incidents as a part of his auxiliary analyses; and Reback et al. (2011) showed that accountability pressure had little impact on enjoyment of learning and may decrease students' anxiety towards testing.[10] In addition, several qualitative studies support the view that accountability pressure can affect students' non-achievement behaviors. For example, based on in-school observations and interviews, Wheelock, Haney, and Bebell (2000) find evidence that students 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) provide qualitative evidence that students subject to accountability pressure often exhibit stomachaches and headaches indicative of increased levels of anxiety. These qualitative studies are supported by teacher-level surveys, in which teachers report that students respond to school failure with increased levels of anxiety and lower levels of confidence and "love of learning" (Jones et al., 2007).

The relative scarcity of studies examining the impact of accountability pressure on non-test score measures of student behavior is unfortunate given the growing literature documenting the significance of what many researchers refer to as "non-cognitive" or psychosocial skills.[11] These attributes capture the learned attitudes, behaviors, and strategies that help children assimilate in society, but that are not captured by standardized tests. These skills are increasingly recognized as central to student performance in school and beyond. These so called non-cognitive skills may act

[5] That the gains emerged only for high-stakes tests suggests that strategic behavior—or "teaching to the test"—occurred.

[6] Scholars have long argued that school failure has a strong negative stigma attached to it (e.g., Albrecht and Joles 2003, 87). Indeed, empirical work is consistent with the fact that communities do not like having their schools being marked as failing and, as a result, people respond to such failure signals (e.g., Black 1999; Figlio and Lucas 2000; and Holbein 2016).

[7] Traczynski and Fruehwirth (2014) show dynamic effects that potentially diminish over time.

[8] Some of the small effects on academic achievement may be traced back to the nature of what these tests are measuring—with achievement being closely linked to the somewhat rigid constructs of socioeconomic status and cognitive ability.

[9] Other studies that use comparisons across accountability regimes find similar results. Clotfelter et al. (2004) find that accountability pressure made it more difficult for low performing schools to retain teachers. Feng et al. (2010) find similar results in Florida. Reback, et al. (2011) also find that accountability pressure lowered teacher perceptions of job security and increased the number of hours untenured teachers in high-stakes grades worked.

[10] They speculate that the latter occurs because failing schools prepare students for exams by using practice exams.

[11] These skills go by many other names, including: soft skills, character skills, emotional intelligence, social cognitive abilities, meta-cognitive learning skills, and socio-emotional skills.

as complements to cognitive ability—capturing abilities that contribute to student performance on standardized tests and on other academic tasks.

The measures that we use in this paper—including absences, tardies, and misbehaviors—have been used in various contexts as proxies of students' non-cognitive skills. For example, Heckman, Humphries, Urzua, and Veramendi (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. This approach follows the observed-behaviors technique to understanding students' skills (e.g., Carneiro, Hansen, & Heckman, 2003; Heckman, Pinto, & Savelyev, 2013; Heckman, Stixrud, & Urzua, 2006; Heine, Buchtel, & Norenzayan, 2008; Jacob, 2002). Scholars who use this approach argue that observed human behaviors are informative of the underlying set of skills regulating those behaviors (Heckman & Kautz, 2013, 13–21).[12]

Our measures are specifically relevant for two broader constructs: children's ability to "be in school when they are supposed to be" and to "behave while they are in school." The first two measures—absences and tardies—are proxies for 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 misbehavior measures we examine include: behaviors that lead to suspensions, incidents of fighting, and various types of offenses such as those related to the possession of drugs, violence, risky sexual behavior, weapons, disruption, and falsification. Avoiding these externalizing behaviors is also important for performance in school and beyond.

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 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 some extent, the second two tasks could contribute to the basic goal of raising student achievement so that one might view the challenge as being one of investing in increasing attendance and improving student behavior up to the point at which the marginal productivity in terms of student achievement are equalized. However, the connection between achievement and these two behaviors is noisy and indirect. Underlying our multitasking approach, is the assumption that the three tasks are each valued in their own right and that marginal returns in the form of student achievement are far higher for investments in the teaching of cognitive skills directly than for investments in either of the other two non-cognitive skills that could lead to indirect gains in achievement. This assumption finds some empirical support; indeed, while investments in students' non-cognitive skills sometimes increase achievement (e.g., Eckenrode et al., 2010; Gertler et al., 2013), this result is by no means guaranteed (e.g., Chetty, Friedman, Hilger, Saez, Schanzenbach, & Yagan, 2011; CPPRG 1999; CPPRG 2011; Niles, Reynolds, & Nagasawa, 2006; Sorensen and Dodge 2016). In practice, the three tasks may require different types of targeted investments. For example, investments in tutoring programs or new pedagogical approaches that target cognitive skills may be best situated to improve achievement, whereas interventions that teach children skills such as self-regulation may be most appropriate for decreasing absenteeism and misbehavior. To accomplish the three tasks requires resources—including time, money, and personnel—that are in limited supply. Hence, school administrators face tradeoffs in how to allocate resources among their somewhat competing priorities.

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 (Fryer & Holden, 2012; Gibbons, 1998; Holmstrom & Milgrom, 1991; Laffont & Martimort, 2009). NCLB incentivizes school officials to place attention on raising scores on standardized tests and other academic indicators such as attendance, while not incentivizing others, such as how students behave while in school.

Educators have some clear tools for raising student test scores. Some of these tools—such as improvements to curriculum or teaching practices or increased access to after school tutoring services—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 to make it appear that scores are rising (e.g., Figlio & Winicki, 2005).

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 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 goal 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 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. To be sure, as we just noted, good behavior could have indirect effects on the incentivized outcome of student achievement, but in most cases, resources directly devoted to that task are likely to be more productive than those directed at student behavior.

[12] 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 avoids problems of survey-based measures such as reference bias and survey item nonresponse (Heckman and Kautz 2013).

[13] Throughout the paper we use the terms "behaving in school," "anti-social behaviors," and "externalizing behaviors" interchangeably to specifically describe student misbehaviors. 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] We acknowledge that these are a smaller subset of the tasks school officials face, including teaching democratic values, a love of learning, and practical skills for the workforce.

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 for elementary and middle schools, we predict that schools facing accountability pressure will find ways to decrease absenteeism. Still, this increased emphasis by school officials may or may not come to fruition because students may be resistant to such efforts.

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 predicted effect of accountability pressure on the number of tardies is unclear.

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 reported misbehaviors. This may occur if school administrators invest in student behavior as an indirect means of increasing achievement. 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 educators is transferred to 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 types 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 appropriate attention on disadvantaged students. Indeed, this was part of the law's original focus to "leave no child behind".

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

4. Data

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

$$P_{st} = f(F_{st-1},\ I_{st}) \tag{1}$$

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

Schools that fail to make adequate yearly progress ($F_{st\text{-}1}$) in a given year face a discrete jump in accountability pressure during the next school year primarily 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, or they have failed previously and are consequently subject to more stringent sanctions (I_{st}).

As we mentioned earlier, we use the term "accountability pressure" throughout this article as shorthand for the pressure that originates from school failure. We emphasize again that all schools, regardless of their performance, face some accountability pressure because they are all subject to NCLB's requirements but failing schools undoubtedly face an added dose of accountability pressure. Although it does not capture the total pressure imposed by the implementation of accountability systems, this added dose of pressure is likely to be meaningful.

Failure under NCLB is determined by a complex formula. At its most basic level, students in a given 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.[18] 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.

In this article, information on the accountability pressure individual schools face is based on North Carolina's public school per-

[15] Science tests are taken in NC schools. These, however, do not influence AYP.

[16] Reporting exemptions are given to schools when they have less than forty 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.

[18] While relatively few (≈3%) of schools fail according on the OAIs in a given year, the threat for failure remains salient. Many schools are very close to the cutoff.

formance data from 2006–2011. These data 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 depending on the year. The most commonly failed subgroup categories come, perhaps unsurprisingly, from the performance of low-SES and minority subgroups: groups with historically low performance.

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.[19] 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.[20] In practice, this approach means that we use matched school performance data from 2006–2011 to student behaviors from 2007–2012.[21] Our full estimation sample consists of about five million student-year observations nested in about 11,000 school-year observations. To account for this hierarchical structure, in our models below we collapse our observations to the school-year level: making our unit of observation a (weighted) school year.

In our analyses we examine absences and tardies separately.[22] Measures of these outcomes come from the NCERDC records for students in grades 3–12. The separate analysis of each of these outcomes 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 indicator 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.[23] Because absences and tardies are count measures that are right skewed at the school level, we transform them into logarithmic form.[24]

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 A2 in the appendix for a full list). These are documented for students in all grades from Kindergarten through Twelfth Grade, and are generated at the time of offense. Using each of these as separate outcomes is unpalatable due to the problems associated with multiple hypothesis testing. Following previous work examining student misbehaviors (e.g., Flay, Graumlich, Segawa, Burns, & Holliday, 2004), we group offenses together into a set of logically coherent categories, aggregating the individual measures into seven separate additive scales.[25] 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 measures individually because of their frequency and severity. These include a behavioral measure of the number of fights and two punishment measures: in-school suspensions and out-of-school suspensions.[26]

Some of our measures such as sexual and possession related offenses, are relatively rare—particularly in elementary schools. As a result, some schools reported no offenses of these types in a given year. Rather than deleting these schools from our sample—as logging these outcomes (to address skew) would do—we instead collapse our offenses outcomes to indicators for whether the school had an offense of the given type. We do this for all of our offense outcome measures except for suspensions, as these are much more common than our other offense outcomes.[27]

Throughout this paper, we interpret changes in our measures of misbehavior as actual changes in student behavior. That is, if we find evidence that accountability pressure led to a higher number of reported fights, we take that as evidence that students felt the pressure and responded by fighting more. In fact, there could well be an alternative interpretation of such a finding, namely that school-level educators responded to accountability pressure by using more rigorous reporting standards for student behavior than they otherwise would have. The correct interpretation is probably some combination of both. Regardless of whether the reported changes reflect actual changes in student behavior or changes in reporting standards, however, the basic conclusion would be the same. Under both interpretations, an increase in reported misbehaviors would be indicative that accountability pressure directed at schools and educators is transmitted down to the student level in potentially undesirable ways. Under one interpretation, the adverse effects show up in the form of greater student misbehavior. Under the other interpretation, the adverse effects show up in negative blots on student records and, in some cases, loss of learning time. When students are suspended, for example, they miss valuable instructional time, regardless of the reasoning behind the suspension.

5. Methods

We leverage a discontinuity in the determination of AYP and use regression discontinuity models to isolate the causal effect of failure-induced 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 & Lemieux, 2008; Lee & Lemieux, 2010; Lemieux & Milligan, 2008). Under this method, observations close to an arbitrary cutoff are separated by exogenous shocks (Butler & 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

[19] 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 for all our dependent variables. For example, a small number of schools (about 500 or $<$ 2% of the entire sample) report very few absences in a given year. This is especially prevalent in middle and, even more so, high schools. In our models below, we use mean imputation for these schools—replacing these with the mean level of absences in corresponding school types in the district in a given year.

[20] Failure status is published the summer after a given testing year. Anticipating marginal failure is difficult, because many measures go into determining AYP.

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

[22] At the individual level, tardies are conditional on attendance.

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

[24] One might be tempted to argue that school size would bias our estimates for absences and tardies. However, around the AYP cutoff, school size is balanced (see Table A3 in the Online Appendix).

[25] 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. 2012).

[26] 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.

[27] We could instead model our offense outcomes with a more complex two-part model. If we do so, we find results that run parallel to those outlined below.

events that push them to one side of the arbitrary cutoff or the other. Regression discontinuity models take advantage of this exogenous variation, using data on either side of the cutoff to estimate the change in an outcome. While this approach generates good internal validity, it may come at the cost of limiting generalizability to units around the cut point.

The starting point for any regression discontinuity 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 AYP and control consists of schools making AYP. As school failure status is readily 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 standard approach proposed by Ahn and Vigdor (2014a) for their study of how accountability pressure affects student test scores.[28] 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 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 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.[29]

We use the Ahn and Vigdor approach because it conceptually mirrors the process of determining failing/passing under NCLB.[30] 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 about 80% of schools.[31] Moreover, this approach benchmarks well with slightly different methods of specifying of the running variable in the NCLB context (Traczynski and Fruehwirth, 2014).[32] 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. We show the two-stage form of this model here to illustrate how this model is specified.

$$\mathrm{F}_{st-1} = \gamma_0 + \gamma_1 \mathrm{P}_{st-1} + g\ (\mathrm{R}_{st-1}) + \gamma_2 \mathrm{X}_{st-1} + \xi_{st} \tag{3}$$

$$\mathrm{O}_{st} = \beta_0 + \beta_1 \hat{\mathrm{F}}_{st-1} + g(\mathrm{R}_{st-1}) + \beta\ \mathrm{X}_{st-1} + \varepsilon_{st} \tag{4}$$

Equation [3] models observed actual failure ($\mathrm{F}_{st\text{-}1}$) as a function of the excluded instrument determined by the running variable ($\mathrm{P}_{st\text{-}1}$)—an indicator for passing the threshold in the running variable—and the running variable ($\mathrm{R}_{st\text{-}1}$). In equation [4], we estimate each outcome variable (O_{st}) as a function of the instrumented failure variable, ($\widehat{\mathrm{F}}_{st-1}$), proximity to failure ($\mathrm{R}_{st\text{-}1}$)—which we model with a flexible non-parametric form, denoted by g(•) that allows flexibility in both stages of the model—and a set of controls ($\mathrm{X}_{st\text{-}1}$).[33,34] To increase precision and to ensure that any slight covariate imbalances at the failure discontinuity do not confound our results, our models also include statistical controls for variables that show some sign of being imbalanced at the failure cutoff.

Below we show that our models are robust to a variety of necessary modeling choices. One choice relates to the size of the bandwidth. It has long been known that choosing a bandwidth comes with a bias/efficiency tradeoff: with observations closer to the cutoff producing less bias, but more statistical uncertainty (Lee & Lemieux, 2010). Our preferred results that we outline below are based on the optimal bandwidth recommended by Imbens & Kalyanaraman (2012), as these estimates have been shown to balance this tradeoff between a lack of statistical power (something we are conscientious about because of our nested data structure) and bias. We note, however, that the bandwidth choice is, in some ways, less important to the specification of the RDD models than the modeling of the running variable (Gelman & Zelizer, 2015; Lee & Lemieux, 2010). Indeed, as long as we are correctly approximating the underlying function of the running variable, it matters much less how much data around the cutoff we are using. We also report results for other bandwidths, noting when dissimilarities arise across the bandwidths. In addition, in the Online Appendix

[28] 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 (2016) for applied examples of similar approaches.

[29] 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.

[30] Some work has begun to grapple with the issue of specifying the running variable with multiple inputs. However, this work explicitly deals with multiple treatments (Papay et al. 2011, 204).

[31] 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.

[32] Traczynski and Fruehwirth (2014) 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.

[33] We use a triangle kernel that places greater weight on points around the cutoff.

[34] With our continuous outcomes, we choose non-parametric regression because it is deemed the best practice in the RDD literature given its flexibility (Hahn, Todd, and Vander-Klaauw 2001; Lee and Lemieux 2010). For our dichotomous variables we use a probit specification with average marginal effects. Models with probit model the running variable as flexibly linear, given the potential for over-fitting with higher-order polynomials (Gelman and Imbens 2014; Gelman and Zelizer 2015).

Table 1
The effect of accountability pressure on being in school.

	Without controls		With controls	
	(1) DV: Absences	(2) DV: Tardies	(3) DV: Absences	(4) DV: Tardies
IK optimal bandwidth				
School failure	-0.334^{**}	-0.295	-0.314^{**}	-0.361^{*}
	(0.146)	(0.188)	(0.145)	(0.187)
μ_{School}	1278.46	386.56	1278.46	386.56
Number of students	3979,474	2312,539	3979,474	2312,539
Number of schools	7936	4279	7936	4279
Half IK optimal bandwidth				
School failure	-0.229	-0.207	-0.189	-0.258
	(0.237)	(0.266)	(0.238)	(0.267)
μ_{School}	1268.17	345.95	1268.17	345.95
Number of students	2729,550	1780,070	2729,550	1780,070
Number of schools	5510	3435	5510	3435
Twice IK optimal bandwidth				
School failure	-0.444^{***}	-0.286^{*}	-0.412^{***}	-0.370^{**}
	(0.114)	(0.165)	(0.114)	(0.163)
μ_{School}	1374.38	401.23	1374.38	401.23
Number of students	4460,776	2431,347	4460,776	2431,347
Number of schools	8785	4460	8785	4460
Full bandwidth				
School failure	-0.457^{***}	-0.263	-0.422^{***}	-0.349^{**}
	(0.106)	(0.161)	(0.104)	(0.158)
μ_{School}	1376.75	401.23	1376.75	401.23
Number of students	4472,163	2431,347	4472,163	2431,347
Number of schools	8806	4460	8806	4460

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Outcomes are logged to help adjust for skew. Cluster-robust standard errors (at the school-year level—the level of treatment) are listed below coefficient estimates. The dependent variables are modeled using local-linear regression. The optimal bandwidth was used (Imbens & Kalyanaraman 2012). μ_{School} represents the non-logged, school-level mean for the outcomes in the corresponding bandwidths. Controls include those showing any indication of imbalance in Table A1, including: pupil/teacher, % migrant, and whether a school had failed previously.

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 makes our unit of observation a weighted school-year. This approach follows that of other similar work involving NCLB (e.g., Ahn & Vigdor, 2014a; Traczynski and Fruehwirth, 2014) and allows us to preserve the virtues of our rich individual-level data, account for the clustered nature of our school-year level treatment, and model our outcomes in a way that is readily interpretable and justified by the distribution of our outcomes.[35] Under this approach, our estimates represent the causal effect provided the "as-good-as random" assignment of schools holds at the failure cutoff.

We present specification checks in the Online Appendix that explore whether this assumption is satisfied. These include a check for the covariate balance of many covariates, including checks of lagged versions of our outcome measures; the McCrary (2008) check for precise sorting at the cutoff; and a check of placebo treatments at other points along the running variables. These checks provide us with reasonable confidence that the AYP failure can be used to draw out internally valid estimates.

6. Results: absences & tardies

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 1 shows that, consistent with our expectations, accountability pressure causes a reduction in the number of reported student absences. Our preferred estimates are in the first row, which is based on the optimal bandwidth. Here we can see that the effect size is noticeable: failure causes absences to decline by about 30%–40%, on average, during the following school year. This effect is statistically different from 0 at the 5% level in the optimal bandwidth in both the models with and without controls. That this effect is not significant in very narrow bandwidths is likely a matter of a lack of statistical power rather than a lack of a true effect. While there is some variation in the size of this estimate across model bandwidths—with narrower bandwidths generally showing slightly smaller effect sizes—the direction of the effect does not vary and the narrower estimates are not statistically distinguishable from estimates in wider bandwidths. Our estimate from the optimal bandwidth represents about 20–30% of a standard deviation in 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 300–400 fewer absences in a school, on average.[36]

Table 1 also shows the effect of accountability pressure on tardies for which we do not have a clear prediction. 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 appears to dominate 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 20–30%, on average. This represents about 0.16 of a standard deviation, which equates to about 150 fewer tardies in schools, on average. Although these estimates are not statistically significant at standard levels, they are close—with p-values of 0.117 and 0.054 in the optimal bandwidth models without and with controls, respectively. In both models, the statistical precision increases with the expan-

[35] We cluster our standard errors at the level of treatment: the school-year.

[36] For reference, the average school size in our data is approximately 460 students.

sion of the bandwidth. Perhaps more illuminating, the substantive size of this effect does not vary by bandwidth. This pattern suggests that a lack of significant differences from 0 in the narrower bandwidths likely reflects a lack of power.[37]

In short, the weight of evidence suggests that when schools feel accountability pressure after failing to make AYP, they respond with efforts that end up helping students show up to school and (perhaps) to come to class on time. Given that high rates of absenteeism bode poorly for all aspects of student learning, this observed decline represents a normatively positive effect of NCLB.

6.1. Results: reported misbehaviors

Table 2 shows the impact of school failure on ten measures of reported 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.

Table 2 shows that for some of the measures, accountability pressure appears to induce students to misbehave more than they otherwise would. The results, however, are somewhat sensitive to the bandwidth we use. Based on the optimal bandwidth, we find, first, that school failure causes suspensions to rise by about 15–20% in the next year, on average. This is equivalent to a 0.16σ increase in suspensions, or about 20 more per school, on average.[38] Based on the optimal bandwidth we would also conclude that accountability pressure increases the probability of drug related offences (labeled “possession”), risky sexual behaviors, and offences that are reportable to law enforcement agencies by about 4–7 percentage points, but none of the other measures. Based on a wider bandwidth (see the panel for twice the optimal size), however, we would conclude that accountability pressure also appears to increase fights and disruptive behavior. Because the estimates based on the wider bandwidths are generally statistically indistinguishable from those based on narrower bandwidths, we believe that the lack of significance in narrower bandwidths (see results in panel for half the optimal bandwidth) is due to a lack of statistical power, rather than the absence of a true effect. In short, our results suggest that while accountability pressure appears to decrease absenteeism, as NCLB intended, it may have the negative unintended consequence of increasing student misbehaviors. The questions remain as to what is driving these effects and whether they are uniform across student subgroups.

6.2. Results: refinements and extensions

In this section we dig deeper into what is driving the results we just presented. To further explore the applicability of the multitasking framework, we test first for variation in responses across schools with differing 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, this is not the case for high schools. So far we have grouped both types of schools together. If our multitasking framework is correct, we would predict that absences would decline in elementary and middle schools but not in high schools.[39]

Table 3 displays the effects broken down by school level, with the important test statistic for the coefficient differences being shown in the last row. In the first row of the first panel, we report results for elementary and middle schools alone, and in the second, for high schools. The results indicate that school failure causes a noticeable decrease in the number of absences, but only in elementary and middle schools. In contrast, in high schools, which are not held accountable under NCLB for absentee rates, there is little evidence of a decline in absenteeism. These statistically distinct estimates are consistent with the view that school failure under NCLB leads to a decline in absences because it causes schools to attend to an incentivized metric. In short, incentives play a large role in the reduction of student absences.

Table 3 also provides similar estimates across school levels for our offense measures. These are not as informative to what is driving our effect estimates, as NCLB does not incentivize misbehaviors in any school level. Still, they highlight some potential heterogeneities in our effect estimates. The second panel in Table 3 shows the effect of school failure on our misbehavior measures, with elementary schools being separated out on the grounds that they serve younger children who are much less likely than older children to engage in some of the behaviors such as those related to drugs or to risky sexual behavior. As can be seen, the evidence suggests that the accountability pressure associated with the failure of an elementary school tends to tranlate into more fights, more disruptive behaviours, and more violence. In contrast, older students are significantly more likely to respond by being involved in drug related, sexual, and reportable offenses that one would expect to be more prevalent among older students.[40]

As a further test of our multitasking framework, we leverage variation in exposure to sanctions. As we mentioned earlier, in addition to labeling schools as failing, NCLB applies sanctions to schools that repeatedly fail. Sanctions are not imposed until a school fails twice consecutively and the sanctions are increasingly stringent. Hence, a failing school that is currently facing sanctions is likely to feel more accountability pressure than one that is not yet being sanctioned. Based on this logic, we would expect school failure to exert large effects in schools subject to sanctions than in other schools.

We find some support for this expectation, especially with respect to absences, in models that split the sample into schools facing and not facing sanctions (see Table A6 in the Online Appendix). The difference between coefficient estimates in the two samples is statistically significant at the 10% level for absences ($p \approx 0.09$) and the 5% level for tardies ($p < 0.01$).[41] In fact, most of the decline that we reported earlier for absences appears to come from schools that are facing some sanctions, rather than in schools that fail when sanctions are not at stake. Thus, sanctions appear to play a key role in dialing up the accountability pressure needed to reduce student absenteeism. This pattern also emerges for some of our externalizing behaviors, particularly out-of-school suspensions, but is much less muted overall.[42] Together, these school-level checks provide support for the conclusion that the specific incentives built into accountability programs matter for student behavior.

We next explore whether the effects of accountability pressure differ across student subgroups, defined primarily by their previous

[37] The effects of accountability pressure on absences and tardies are shown visually in Figure A3 in the Online Appendix.

[38] There is some evidence that there might be fewer in-school suspensions as a result; however, this effect is only significant at the 10% level.

[39] 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 at the higher level.

[40] One final pattern is notable in Table 3. In Table 2, the estimates for in-school suspensions were all negative, and only barely significant at the 10% in the full bandwidth. Yet, when we break our models by school levels, we see that these estimates gain statistical precsion—perhaps due the high level of variance in ISS across school levels. This effect suggests that some of the results we document here may be due to strategic behavior on part of school officials. Accountability pressure appears to decrease principals' use of ISS and increase their use of OSS.

[41] This difference holds if we restrict our sample to just elementary and middle schools.

[42] Indeed, sexual offenses are in the opposite direction than what we would predict.

Table 2
The effect of accountability pressure on externalizing behaviors.

	(1) Log (ISS)	(2) Log (OSS)	(3) Fights (0/1)	(4) Possess (0/1)	(5) Violence (0/1)	(6) Sexual (0/1)	(7) Weapons (0/1)	(8) Disrupt (0/1)	(9) Falsify (0/1)	(10) Reportable (0/1)
IK optimal bandwidth										
School failure	−0.18*	0.18*	0.02	0.04*	0.01	0.07***	0.00	0.01	0.02	0.04*
	(0.11)	(0.09)	(0.02)	(0.02)	(0.02)	(0.03)	(0.03)	(0.02)	(0.02)	(0.03)
μ_{School}	88.50	100.58	0.68	0.49	0.79	0.45	0.11	0.79	0.60	0.48
# of students	5066,742	4773,954	6549,060	6801,589	6284,521	6571,934	5760,856	6128,669	6855,686	6733,192
# of schools	10,582	10,038	14,104	14,597	13,563	14,165	12,500	13,269	14,696	14,466
Half IK optimal bandwidth										
School failure	−0.17	0.21	0.00	0.02	0.02	0.03	−0.02	0.01	0.01	0.00
	(0.14)	(0.14)	(0.03)	(0.03)	(0.04)	(0.04)	(0.05)	(0.04)	(0.03)	(0.04)
μ_{School}	84.10	95.66	0.69	0.49	0.79	0.45	0.11	0.79	0.59	0.48
# of students	4203,969	3346,642	4760,060	5331,295	4252,926	4812,855	3547,860	4030,929	5508,206	5180,745
# of schools	8848	7060	10,352	11,625	9244	10,480	7702	8748	11,993	11,271
Twice IK optimal bandwidth										
School failure	−0.17*	0.16**	0.04*	0.04*	0.03	0.07***	0.03	0.04**	0.02	0.04*
	(0.09)	(0.08)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
μ_{School}	89.90	108.11	0.69	0.50	0.78	0.45	0.12	0.79	0.60	0.48
# of students	5157,457	5170,884	7057,598	7058,490	7040,118	7057,820	6925,126	7016,843	7058,490	7058,490
# of schools	10,746	10,788	15,072	15,077	15,034	15,074	14,821	14,991	15,077	15,077
Full bandwidth										
School failure	−0.16*	0.15**	0.04*	0.04*	0.03	0.07***	0.03	0.04**	0.02	0.04*
	(0.09)	(0.07)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
μ_{School}	89.90	108.13	0.69	0.50	0.78	0.45	0.12	0.79	0.60	0.48
# of students	5157,457	5171,622	7058,490	7058,490	7058,490	7058,490	7058,490	7058,490	7058,490	7058,490
# of schools	10,746	10,792	15,077	15,077	15,077	15,077	15,077	15,077	15,077	15,077

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The first two column outcomes are logged to help address skew; the others are binary. Cluster-robust standard errors (at the school-year level—the treatment level) below coefficient estimates. The continuous outcomes are modeled using local linear regression, while the dichotomous outcomes are modeled with a probit specification with average marginal effects. The optimal bandwidth was chosen using the procedure suggested by Imbens and Kalyanaraman (2012). μ_{School} represents the non-logged, school-level mean for the outcomes in the corresponding bandwidths. The estimates above do not include controls; the results do not change if these are included.

Table 3
Accountability pressure, by school-levels.

	(1) Log (Absences)	(2) Log (Tardies)		(3) Log (ISS)	(4) Log (OSS)	(5) Fights (0/1)	(6) Possess (0/1)	(7) Violent (0/1)	(8) Sex (0/1)	(9) Weapon (0/1)	(10) Disrupt (0/1)	(11) Falsify (0/1)	(12) Report (0/1)
Failure in elementary/ Middle schools (incentivized)	−0.316	−0.315	Failure in elementary schools	−0.173	−0.049	0.045	−0.042	0.047	−0.033	−0.014	0.045	−0.022	−0.033
	(0.078)	(0.150)		(0.071)	(0.066)	(0.022)	(0.024)	(0.017)	(0.024)	(0.016)	(0.017)	(0.023)	(0.024)
Failure high schools (not incentivized)	0.042	0.214	Failure high schools/ Middle schools	−0.176	0.068	−0.026	−0.004	−0.021	0.021	0.020	−0.004	−0.019	0.003
	(0.137)	(0.196)		(0.083)	(0.049)	(0.020)	(0.022)	(0.020)	(0.023)	(0.025)	(0.019)	(0.021)	(0.022)
p ($\beta_{elementary/middle} = \beta_{high}$)	0.00***	0.00***	p ($\beta_{elementary} = \beta_{middle/high}$)	0.94	0.01***	0.00***	0.00***	0.00***	0.00***	0.07	0.00***	0.76	0.00***

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$: stars for p-values below these thresholds only listed for tests of heterogeneities, to highlight these. The first four column outcomes are logged to help address skew; the others are dichotomous. Cluster-robust standard errors (at the school-year level–the level of treatment) below coefficient estimates. The optimal bandwidth was chosen using the procedure suggested by Imbens and Kalyanaraman (2012). Controls include those showing any indication of imbalance in Table 1, including: pupil/teacher ratio, % migrant, and whether a school had failed previously. P-values of coefficient difference are based on the Wu-Hausman technique (Hausman, 1978; Wu, 1973).

performance and racial status.[43] These comparisons are directly relevant to policy discussion about NCLB, which specifically placed emphasis on underperforming student subgroups. We note, however, that the heterogeneities discussed in this section cannot be interpreted as fully causal, as unlike school failure at the margin our subjects are not exogenously assigned to the student characteristics we explore.

To examine potential differential effects across initial levels of student performance, we first group students by their previous end-of-grade (EOG) test scores. Such an analysis restricts our sample to students in grades three through eight because of the annual testing in those grades. Table A7 in the Online Appendix shows our results, highlighting the important differences across subgroup coefficients.

On metrics of "showing up," we find that schools find ways to reduce absences and tardies among all performance subgroups. The estimates across student performance quartiles are in most cases not consistently different. In short, when schools face accountability pressure and are incentivized to target attention towards a certain behavior, they pressure all their students to come to school. The same pattern generally applies to reductions in tardies.

In contrast, we find interesting heterogeneities in our measures of externalizing behavior. On some of our measures, we see a u-shaped pattern—with only students at the bottom and the top of the test score distribution exhibiting notable increases in misbehavior. For students in the lowest quartile of student performance, accountability pressure increases 7 out of the 10 metrics of misbehavior more than the middle two categories. This is true in-school suspension, out of school suspension, fights, possession-related offenses, violence-related offenses, sexual-related offenses, and disruption-related offenses. A similar pattern can be seen among the highest quartile: with 6 out of the 10 misbehavior metrics including out-of school suspensions, possession-related offenses, sexual-related offenses, weapons-related offenses, falsification-related offenses, and those offenses reportable to law enforcement showing higher responsiveness than the middle two categories. These results show that when schools face accountability pressure, the highest and lowest performing students are most negatively affected.

What explains this pattern? At first glance, it may appear to contradict the finding of other researchers that accountability induces teachers to focus attention on students at the margin of being proficient—that is, those in the middle quartiles (Neal & Schanzenbach, 2010). In fact, though, because that view implies that teachers pay less attention to students at the bottom and the top of the performance distribution, these patterns are consistent with that view. High and low performing students who receive lower levels of attention may be more likely to act out and engage in the types of misbehaviors we document.[44] Whatever the reason for these patterns, we simply note that they represent an unintended and perhaps undesirable distributional effect of performance-based accountability systems.[45]

Table A8 in the Online Appendix shows the effect of accountability pressure by student race/ethnicity subgroups. Again, the important test statistics here are those testing differences across subgroups—those highlighted in the bottom three rows. Our results show that accountability reduces absences the most for white stu-

[43] Although we also tested for differences by student gender and socioeconomic status, we find few if any statistically significant differences along these dimensions.

[44] Indeed, there is some evidence of this in the slight decline in some student misbehaviors among the middle performance categories. These declines are relatively small, and tend to be significant only at the 10% level.

[45] This pattern may alternatively reflect a mismatch between the distribution of pressure and students' capacity for improving their non-achievement behavior.

dents. Although declines in absences for African American and Hispanic students are also noticeable, they are not as large as those for white students. This difference is statistically distinct from zero at the 95% confidence level. For tardies, the differences across the subgroups are less clear. White students appear to respond more than Hispanic students but there is little difference in responsiveness between White and Black students.

With respect to reported student misbehaviors, the evidence suggests that minorities, and especially Black students, respond to pressure by increasing their misbehavior more than White students. This pattern holds in 5 out of the 10 misbehavior metrics for African Americans and in only 2 out of 10 misbehavior metrics for Hispanics. 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 the other behavior categories, including, fighting, violent offenses, disruptive behavior and reportable offences, the evidence suggests that accountability pressure has a similar impact among Hispanic and Black students.

These patterns across student performance and race/ethnicity suggest that accountability pressure has heterogeneous effects. Although we cannot claim these are causally determined patterns and have not explored reasons for the differences, we believe the patterns are important.

7. Conclusion

This study provides evidence that school-level accountability pressure from NCLB affects students in various unintended ways that have been understudied in the literature. We show, first, that accountability pressure improves the non-achievement behaviors that are directly (or tangentially) incentivized, namely by reducing student absences (and tardies). As a multitasking framework would predict, these positive effects emerge most clearly in elementary and middle schools, which face incentives to reduce absentee rates, and less so in high schools—where reductions are not incentivized. This difference suggests that our results are driven by NCLB's emphasis on student absenteeism and not simply by a school's desire to raise student achievement. Moreover, as would be expected, the estimated effects are typically larger in failing schools with sanctions immediately at stake than in those that have not previously failed, suggesting that sanctions too have a role to play in producing lower rates of absenteeism. Given the broader consequences of student absenteeism, this reduction in absences is a distinctly positive normative finding.

Consistent with our multitasking framework, however, we have also shown that accountability pressure generates unintended increases in reported anti-social student behaviors. Thus, accountability pressure at the school level is transferred down to students, perhaps because schools devote time and resources to improving incentivized behaviors at the expense of ignoring other behaviors, or perhaps in part by tightening reporting standards for offenses. Following school failure, schools experience noticeable increases in reported 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 some of the reported misbehaviors are higher 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 or outcomes 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 unintended harm by increasing student misbehavior in school. A more complete understanding of how performance-based accountability programs such as NCLB affect non-achievement student outcomes would help policymakers weigh any benefits of such programs against the potential costs of damaging non-achievement behaviors vital for success in school and beyond.

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, D. Sunshine Hillygus, Jeffrey Traczynski, Jacob Vigdor, three anonymous reviewers, and participants at the Association for Public Policy Analysis and Management, the Association for Education Finance and Policy, and the Calder Research Conference for their thoughtful comments and feedback.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.econedurev.2017.03.005.

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*. National Bureau of Economic Research (No. w20511).

Albrecht, S. F., & Joles, C. (2003). Accountability and access to opportunity: Mutually exclusive tenets under a high-stakes testing mandate. *Preventing School Failure: Alternative Education for Children and Youth, 47*(2), 86–91.

Angrist, J. D., Dynarski, S. M., Kane, T. J., Pathak, P. A., & Walters, C. R. (2012). Who benefits from KIPP? *Journal of policy Analysis and Management, 31*(4), 837–860.

Bacolod, M., DiNardo, J., & Jacobson, M. (2009). *Beyond incentives: Do schools use accountability rewards productively?*. National Bureau of Economic Research (No. w14775).

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

Black, S. E. (1999). Do better schools matter? Parental valuation of elementary education. *Quarterly Journal of Economics*, 577–599.

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

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*. National Bureau of Economic Research (No. w9546).

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

Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W., & Yagan, D. (2011). How does your kindergarten classroom affect your earnings? Evidence from project STAR. *The Quarterly Journal of Economics, 126*(4), 1593–1660.

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

Conduct Problems Prevention Research Group. (1999). Initial impact of the Fast Track prevention trial for conduct problems: II. Classroom effects. *Journal of Consulting and Clinical Psychology, 67*(5), 648.

Conduct Problems Prevention Research Group. (2011). Child Development. *The effects of the Fast Track preventive intervention on the development of conduct disorder across childhood, 82*(1), 331.

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.

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

Eckenrode, J., Campa, M., Luckey, D. W., Henderson, C. R., Cole, R., Kitzman, H., et al. (2010). Long-term effects of prenatal and infancy nurse home visitation on the life course of youths: 19-Year follow-up of a randomized trial. *Archives of Pediatrics & Adolescent Medicine, 164*(1), 9–15.

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

Figlio, D. N., & Lucas, M. E. (2000). *What's in a grade? School report cards and house prices*. National Bureau of Economic Research (No. w8019).

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.

Figlio, D. N. (2006). Testing, crime and punishment. *Journal of Public Economics, 90*(4), 837–851.

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, R. G., Jr, & Holden, R. T. (2012). *Multitasking, learning, and incentives: A cautionary tale*. National Bureau of Economic Research (No. w17752).

Gelman, Andrew, & Imbens, Guido (2014). *Why high-order polynomials should not be used in regression discontinuity designs. No. w20405*. National Bureau of Economic Research.

Gelman, A., & Zelizer, A. (2015). Evidence on the deleterious impact of sustained use of polynomial regression on causal inference. *Research & Politics, 2*(1) 2053168015569830.

Gertler, P., Heckman, J., Pinto, R., Zanolini, A., Vermeersch, C., Walker, S., et al. (2013). *Labor market returns to early childhood stimulation: A 20-year followup to an experimental intervention in jamaica*. National Bureau of Economic Research (No. w19185).

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.

Gottfried, M. A. (2009). Excused versus unexcused: How student absences in elementary school affect academic achievement. *Educational Evaluation and Policy Analysis, 31*(4), 392–415.

Grunbaum, J. A., Kann, L., Kinchen, S., Ross, J., Hawkins, J., Lowry, R., et al. (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., Stixrud, J., & Urzua, S. (2006). *The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior*. National Bureau of Economic Research (No. w12006).

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*. National Bureau of Economic Research (No. w19656).

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), 2052–2086.

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. (2016). Left behind: Citizen responsiveness to government performance information? *American Political Science Review, 110*(2).

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. (2012). Non-cognitive ability, test scores, and teacher quality: Evidence from 9th grade teachers in North Carolina. No. w18624. National Bureau of Economic Research.

Jones, B. D. (2007). The unintended outcomes of high-stakes testing. *Journal of Applied School Psychology, 23*(2), 65–86.

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.

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.

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

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

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

Niles, M. D., Reynolds, A. J., & Nagasawa, M. (2006). Does early childhood intervention affect the social and emotional development of participants. *Early Childhood Research and Practice, 8*(1), 34–53.

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.

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*. National Bureau of Economic Research (No. w16745).

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

Sorensen, L. C., & Dodge, A. K. (2016). How does the fast track intervention prevent adverse outcomes in young adulthood? *Child development*.

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

Traczynski, J., & 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.

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.

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