The Unintended Consequences of Academic Leniency*

A. Brooks Bowden Viviana Rodriguez Zach Weingarten

August 21, 2023

Abstract

In response to widening achievement gaps and increased demand for post-secondary education, local and federal governments across the US have enacted policies that have boosted high school graduation rates without an equivalent rise in student achievement, suggesting a decline in academic standards. To the extent that academic standards can shape effort decisions, these trends can have important implications for human capital accumulation. This paper provides both theoretical and empirical evidence of the causal effect of academic standards on student effort and achievement. We develop a theoretical model of endogenous student effort that depends on grading policies, finding that designs that do not account for either the spread of student ability or the magnitude of leniency can increase achievement gaps. Empirically, under a research design that leverages variation from a statewide grading policy and school entry rules, we find that an increase in leniency mechanically increased student GPA without increasing student achievement. At the same time, this policy induced students to increase their school absences. We uncover stark heterogeneity of effects across student ability, with the gains in GPA driven entirely by high ability students and the reductions in attendance driven entirely by low ability students. These differences in responses compound across high school and ultimately widen long-term achievement gaps as measured by ACT scores.

JEL Classifications: I24, J24, I28

Keywords: Academic Standards, Student Effort, Human Capital

*We would like to thank Anika Alam for excellent research assistance. We also would like to thank Brian Jacob, Shaun Dougherty, Samson Alva, Alper Hayri, participants at the Southern Economic Association, the Advanced Quantitative Methods and Analytics for Public Policy Support, the Causal Inference in Education Research Seminar, the Association for Education Finance and Policy, the UTSA Brown Bag, and the CBCSE Brown Bag for helpful comments and suggestions. This paper was supported by the National Science Foundation Graduate Research Fellowship Program and the Institute of Education Sciences, U.S. Department of Education, through Grant R305B200035 to the University of Pennsylvania. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education. All remaining errors are our own. Contact information: Bowden: University of Pennsylvania, bbowden@upenn.edu. Rodriguez: University of Texas at San Antonio, viviana.rodriguez@utsa.edu. Weingarten: University of Pennsylvania, zachwein@upenn.edu.

1 Introduction

Education is often considered "the great equalizer" for its unique capacity to promote social mobility through the accumulation of human capital. In response to widening socioeconomic achievement gaps and increased demand for high quality post-secondary education, state and federal governments have enacted policies to promote educational attainment. These efforts have had reported success. The last decade has seen steady increases in graduation rates and grade point averages in U.S. high schools (Murnane, 2013), yet student achievement—as measured by NAEP, ACT, and SAT scores—has stagnated, while college enrollment rates have decreased. Together, this suggests that academic standards have declined over time (Harris et al., 2023; Hurwitz and Lee, 2018; Blagg and Chingos, 2016). At face value, a decline in academic standards for graduation can naturally lead to more equitable outcomes since a larger share of students meet graduation requirements and obtain a high school diploma. However, this simplified view ignores the possible endogenous role that academic standards have on students' academic decisions, which directly impacts their accumulation of human capital.

This paper explores the relationship between academic standards, student effort, and human capital accumulation in two distinct ways: First, we develop a theoretical model of student achievement with endogenous effort in the presence of changing grade policies. Our simple model reconciles mixed findings in the literature and provides insight for why these policy effects are theoretically ambiguous. In particular, we show that an implementation which does not account for either the spread of student ability or the magnitude of the relative leniency can exacerbate the achievement gap in public schools. Motivated by the predictions of our model, we then empirically test for the effects of academic leniency on student effort and achievement in a setting where high school students faced an explicit reduction in standards. We recover the causal effect of this change on student outcomes by constructing a mechanism that leverages the implementation timing of a statewide policy in conjunction with a separate assignment rule that determines high school entry eligibility.

Understanding the effects of academic standards on student outcomes is challenging for two reasons. First, standards are intrinsic to teachers, schools, and districts. As students select teachers and schools, the effects of grade leniency on student outcomes becomes confounded by sorting (Figlio and Lucas, 2004). Second, changes in academic standards often occur gradually over time. Meaningful changes in the distribution of grades necessarily happens over long periods of time, making it impossible to disentangle treatment effects from unobserved time trends. We overcome these challenges by exploiting two sources of variation. First, we exploit a statewide shift in grading standards implemented by the State of North Carolina. In the Fall of 2014, the North Carolina State Board of Education voted to standardize high school grading policies to a 10-point scale in an effort to increase comparability between districts and increase the competitiveness of students applying to colleges. This policy effectively decreased the associated numeric threshold for all letter grades for *all* students across all subjects, courses, and teachers, thereby moving all public schools to a more lenient grading scale.

A simple comparison of treated and untreated cohorts might lead to concerns over selection into treatment if students are endogenously retained in school (Deming and Dynarski, 2008). Thus, we require a mechanism which (quasi-)randomly assigns observably similar students to either the stricter grade standard or the more lenient one. In North Carolina, a separate policy which sorts students to a high school entry year provides this necessary second source of variation. Under a fuzzy regression discontinuity research design, we compare students who happen to belong to the first cohort of 9th graders exposed to more lenient grading because their birth date was just to the right of the kindergarten entry cutoff, to students who happen to belong to the previous 9th grade cohort because their birth date was just to the left.

We rely on rich administrative data on the universe of North Carolina public school students for information on achievement and attendance during high school. We merge these data with exact birth date records from the North Carolina Department of Public Instruction to generate a measure of relative distance to the North Carolina kindergarten entry rule cut point. We use this distance measure as the running variable in our regression discontinuity specification. We find that increased grade leniency led students to obtain GPAs that were 0.27 points higher (11% increase), with no effects on student achievement. Furthermore, students that faced more lenient grading standards increased their school absences by 1.3 days per year (20% increase), which drove students to become more chronically absent.

In unpacking the results, we uncover stark heterogeneity of effects across observed student ability. Students at the top of the 8th grade performance distribution were the main drivers of GPA gains, but saw no increases in absences. At the same time, students on the bottom end of the distribution saw no GPA gains from greater grade leniency and were the main drivers of increased absenteeism. The lack of GPA gains for students on the lower end of the test score distributions is striking for two reasons. First, the policy mechanically increases GPA for all students. Second, the North Carolina policy created a larger buffer for students at the margin of passing a class.

Therefore, *a priori*, we would expect this policy to boost GPA for lower achieving students the most.

Finally, we explore whether heterogeneous effects compound over time. On one hand, the initial exposure to this policy could have caused gaps which subsequently disappeared as students adjusted to the new grading standard. On the other hand, gaps may widen each year if the effects of the policy generated lasting changes in the education trajectories of students. We find the latter to be true. Increases in student absences persist and compound over time only for students on the lower end of the test score distribution which is later reflected in decreased ACT achievement for this group. Thus, academic leniency exacerbated achievement gaps and lessened human capital accumulation for students already at a deficit.

This paper relates to three main strands of the literature. We first contribute to a small and growing literature that explores the role of student effort in human capital accumulation. An empirical side of this literature has found that student effort is a key contributor to academic success. Students who are absent and invest less time and effort studying, have been found to have lower achievement and attainment (Durden and Ellis, 1995; Stinebrickner and Stinebrickner, 2008; Metcalfe, Burgess and Proud, 2019). We contribute to this literature by documenting how a policy that induced students to reduce their effort in school also led to decreases in achievement.

A theoretical side considers the strategic decision-making process of students under the incentives produced by educational grading standards (Becker and Rosen, 1992; Betts, 1998). This paper makes a contribution to this literature by developing a model in which students of different ability types, with different returns and costs to effort, choose effort optimally in response to a given grade regime. The key takeaway of this model is that student response depends in part on the magnitude of academic leniency induced by the policy and the *discrepancy and spread* of the student score distribution, which can generate ambiguous effects of grading policy on student investments and learning in school. The predictions of our model help to reconcile competing and incongruous findings in the literature, which has found that relaxing grading standards may lead students to decrease their effort in some contexts (Babcock, 2010; Figlio and Lucas, 2004; Hvidman and Sievertsen, 2021), while at the same time motivate and benefit students in other settings (Ahn et al., 2019; Dee et al., 2016).

Second, we contribute to an established literature on the unintended consequences of tightening accountability measures for public schools. Accountability in education has traditionally been designed to shift incentives for educators and schools. In response to these policies, numerous types of unintended responses to accountability have been documented, such as the effects of teaching to the test (Koretz, 2002; Jacob, 2005; Glewwe, Ilias and Kremer, 2010), manipulating who takes high-stakes exams (Jacob, 2005; Figlio, 2006; Cullen and Reback, 2006), and inflating pass rates of important exams (Dee et al., 2016). However, education policies also induce endogenous responses from students, which may distort the association between educational achievement and human capital accumulation. Our paper speaks to the unintended response of policies in shaping student behavior. Consistent with the literature on unintended consequences of education policies, we find a concentration of gains for students at the top of the ability distribution with negative impacts for students at the bottom end of the distribution.

Lastly, we contribute to a relatively large literature which documents grading standards and its subsequent effects on student welfare. A robust strand of work documents the impact of grading standards on student subsequent achievement (Hvidman and Sievertsen, 2021), how students sort into courses and programs (Bar, Kadiyali and Zussman, 2009; Butcher, McEwan and Weerapana, 2014; Ahn et al., 2019), how students endogenously change the level of effort they choose to exert (Babcock, 2010), and how students perform in the labor market (Hansen, Hvidman and Sievertsen, 2023). We contribute to this literature by recovering causal effects of academic leniency on student effort and learning and documenting large and persistent heterogeneous effects across student ability.

The remainder of the paper is structured as follows. Section 2 develops a theoretical model where students respond endogenously to grading policies. Section 3 describes the grading policy change implemented by the North Carolina State Board of Education in 2014. Section 4 describes the data used in the analysis and provides summary statistics for our final analysis sample. Section 5 establishes the research design. Finally, Section 6 presents our results, validated in Section 7. Section 8 concludes the paper and provides suggestions for future work.

2 Conceptual Framework

We first develop a model of student responses to grading standards. We construct this simple model to guide the interpretation of our empirical results and provide a stylized prediction for how students may respond differently to changes in grading standards or policies depending on their prior skills and experiences. This model additionally demonstrates the importance of policy implementation and the potential for unintentional consequences, such as widening achievement gaps.

2.1 Environment

Consider a high school student, i, defined by their latent ability, a_i. We discretize ability according to $a_i \in \{a^\ell, a^h\}$, which refers to low and high ability, respectively. While we do not formally model the evolution of ability and the impact of socioeconomic input variables (Todd and Wolpin, 2003; Heckman, 2006), we consider ability as the dynamically-produced realization at the time we observe students in our data. Given this, we assume students know their own type.

Schools are endowed with a grading policy, P, set forth by the district or state. Grading policies map numeric course averages (scores) into quality points. The mean of these quality points forms a student's grade point average (GPA). Formally, $P : \mathcal{S} \times \boldsymbol{p} \to \{0, 1, ..., 4\}$, where $\mathcal{S} \equiv [0, 100] \subset \mathbb{R}$ is the score space and $\boldsymbol{p} := \{p_A, p_B, p_C, p_D\}$ is the set of cut points. We assume this policy has identical threshold sizes for all grades above an F, meaning $100 - p_A = p_A - p_B$, as well as $p_A - p_B = p_B - p_C$, and so on[1]. As an example, a 10-point policy takes the form

$$P(s_i, \boldsymbol{p}) = \begin{cases} 4.0, & s_i \in [90, 100] \\ 3.0, & s_i \in [80, 90) \\ 2.0, & s_i \in [70, 80) \\ 1.0, & s_i \in [60, 70) \\ 0.0 & s_i \in [0, 60), \end{cases}$$

where s_i is student i's numeric final average, or score, in a class. In the above, $p_A = 90$, $p_B = 80$, $p_C = 70$, and $p_D = 60$. In general, a symmetric policy is an n-point one where n denotes the length of each passing grade range.

We assume class scores depend on student ability and exerted effort, e_i. In this paper, we focus on the student's problem of earning a grade in one class, although this model can easily be extended to accommodate a semester's worth of courses.[2] Formally, $s_i = s(a_i, e_i)$ for some concave score production function $s(\cdot)$, increasing in a_i and non-decreasing in e_i. We parameterize s_i in the

[1] This is a simplifying assumption that need not hold for the results to hold.

[2] The easiest way to do this would be to assume the semester's utility is the sum of each course's utility. The problem of the student would then change to account for the division of effort across course schedules, rather than the isolated decision to exert effort in any one class.

following way:

$$s_i = \mu + \beta a_i + \gamma \ln e_i + \xi_i, \quad \xi_i \sim F.$$

This function satisfies our assumptions for any $(\beta, \gamma) >> 0$. This form additionally features decreasing marginal returns to effort irrespective of ability. Given our functional assumption on score production, we further impose $e_i \in [1, \bar{e}]$.

The term ξ_i captures shocks to the production of scores. In practice, F can be generalized to any distribution belonging to the class of distributions which have bounded support Ω and are everywhere differentiable along that support. In other words, we assume that $P(\xi_i \in \Omega) = 1$ and that the pdf of F, $f(\cdot)$, exists and is continuous along Ω. We maintain an assumption of boundedness to prevent shocks from taking on extreme values, which would send scores beyond the range $[0, 100]$. In the discussion that follows, we explicitly parameterize the distribution $F \equiv \mathcal{U}(\underline{\xi}, \bar{\xi})$ for $\underline{\xi} := \inf(\Omega)$ and $\bar{\xi} := \sup(\Omega)$.

Students also face costs to exerting effort, c_i, in the form of a convex cost function $c(a_i, e_i)$. In this general setting, effort can refer to, e.g., time spent studying, completing homework, or attending class. We parameterize the cost function according to

$$c(a_i, e_i) = \frac{\kappa e_i}{a_i}.$$

This functional form has the desired properties $\partial c(\cdot)/\partial a_i < 0$ and $\partial c(\cdot)/\partial e_i > 0$ for any $\kappa > 0$. While we proceed under the assumption that cost is linear in effort and production is concave in effort, our analyses would not substantively change if we instead imposed a convex cost function with a linear production function.

2.2 The Student's Problem

We specify a model in which students enjoy utility from their *grade point average* rather than their numeric grade/score, which departs from the groundwork established in Betts (1998). Conceptually, a student should not derive additional utility from earning a 95 versus a 94 in a class if both earn that student a grade of A, or 4.0 quality points.[3] We parameterize this mapping function $P(\cdot)$

[3] In high schools that rank students on the basis of GPA, GPA-dependent rank is one of the most important criteria for college admissions (Espenshade, Hale and Chung, 2005).

according to:

$$P(s_i) = \sum_{j \in \{A,B,C,D,F\}} \phi_j \mathbb{1}_{\{s_i \geq p_j\}}$$

If, for example, s_i falls in the B range, a student would derive utility $\phi_B + \phi_C + \phi_D + \phi_F$. The ϕ_j term then represents the marginal utility received by earning the next highest letter grade. Without loss of generality, we normalize the return to a grade of F by setting $\phi_F = 0$.[4] We further impose that $\phi_j > \phi_k$ for any grade $j > k$, meaning that students derive greater marginal utility from accessing higher grades.

Students make effort decisions at the beginning of each semester before ξ_i is realized. As a result, they seek to maximize their expected utility from earning a grade,

$$\mathbb{E}[u_i|a_i, e_i] = (\phi_A + \phi_B + \phi_C + \phi_D) \cdot \Pr(s_i \geq p_A) + ... + \phi_D \cdot \Pr(p_C > s_i \geq p_D) - \frac{\kappa e_i}{a_i}. \quad (1)$$

Because both high and low ability types incur shocks according to the same distribution, the difference in the levels of ability may generate differences in the *set of feasible grades*, which we denote by $\mathcal{G}^k$ for $k \in \{\ell, h\}$. We assume these sets are distinct, i.e., $\mathcal{G}^\ell \not\equiv \mathcal{G}^h$. In particular, we assume that the environment is defined such that low ability types experience

$$p_C > \underbrace{\mu + \beta a^\ell + \underline{\xi}}_{\text{lowest possible score}} \geq p_D \quad \text{and} \quad p_A > \underbrace{\mu + \beta a^\ell + \gamma \ln \bar{e} + \bar{\xi}}_{\text{highest possible score}},$$

while high ability types instead experience

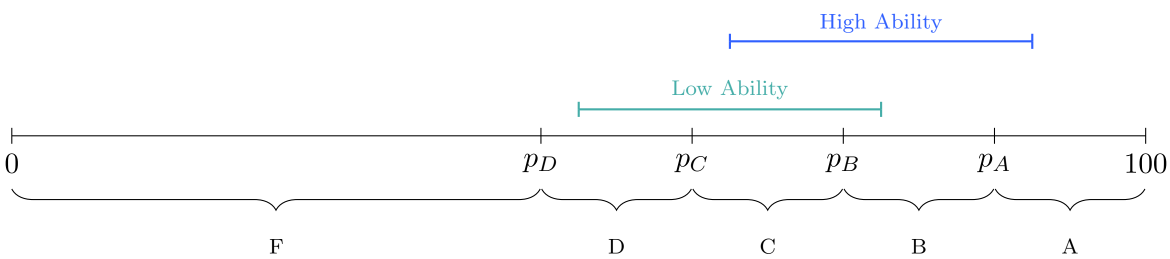
$$p_B > \underbrace{\mu + \beta a^h + \underline{\xi}}_{\text{lowest possible score}} \geq p_C \quad \text{and} \quad \underbrace{\mu + \beta a^h + \gamma \ln \bar{e} + \bar{\xi}}_{\text{highest possible score}} \geq p_A.$$

For either type, the first inequality denotes the infimum of their score set, which occurs when effort is minimized and the lowest production shock is realized; conversely, the second inequality denotes the supremum, obtained whenever effort is maximized and the highest productivity shock occurs. As a result, low ability types have a feasible choice set $\mathcal{G}^\ell = \{B, C, D\}$ and high types instead have $\mathcal{G}^h = \{A, B, C\}$. We display a graphical representation of these differences in Figure 1.[5]

[4] For completeness, we note that every grading policy will have $p_F = 0$.

[5] We recognize that feasible grade ranges can vary across subject domains and across school quality. However, we abstract from these factors to focus on short-term student response to changes in grading standards.

Figure 1: Feasible Grades for Students of Varying Ability Level



NOTES: This figure displays the difference in feasible scores for students of high and low ability type. The corresponding grade range is captured by the underbraces. For example, students earning a score between p_A and 100 earn a grade of A. The range for both types has identical length. The difference in the respective beginning or end point between the two types has a value of $\beta(a^h - a^\ell)$. This figure further demonstrates an example in which $\mathcal{G}^\ell = \{B, C, D\}$ and $\mathcal{G}^h = \{A, B, C\}$.

Following Equation 1, students then solve the following problem:

$$\max_{1 \leq e_i \leq \bar{e}} \quad (\phi_A + \phi_B + \phi_C + \phi_D) \cdot \Pr(s_i \geq p_A) + ... + \phi_D \cdot \Pr(p_C > s_i \geq p_D) - \frac{\kappa e_i}{a_i}.$$

Using the fact that $\xi_i \sim \mathcal{U}_{\{\bar{\xi}, \underline{\xi}\}}$, along with our aforementioned assumptions on feasible grades, it can be shown that low ability types choose an optimal level of effort

$$e_\ell^* = a^\ell \left(\frac{\phi_B + \phi_C}{\bar{\xi} - \underline{\xi}} \right) \frac{\gamma}{\kappa},$$

while high ability students instead choose

$$e_h^* = a^h \left(\frac{\phi_A + \phi_B}{\bar{\xi} - \underline{\xi}} \right) \frac{\gamma}{\kappa}.$$

We then find that $e_h^* > e_\ell^*$ and arrive at our first result:

Result 1: For any given grading policy P and collection of students $I \equiv \bigcup_i$ defined by their distinct ability levels a_i, a pair $\{j, k\}$ which satisfies $a_j > a_k$ and either $\mathcal{G}^j \not\equiv \mathcal{G}^k$ or $\mathcal{G}^j \equiv \mathcal{G}^k$ will also imply $e_j^* > e_k^*$; that is, students with higher ability levels optimally choose to exert a higher level of effort compared to lower ability students regardless of whether the difference in ability generates a difference in the set of feasible grades.

2.3 The Effects of a Policy Change

We now consider the effects of a state or district changing their grading policy P in favor of a new policy P' with corresponding cut points $\{p'_A, ..., p'_D, p'_F\}$. For tractability of our empirical setting, we assume in this discussion that policy P' is more lenient than P, meaning the value of each cut point is lower.[6] Suppose P' shifts p_A by $d > 0$. In other words, the district ends their use of an n-point grading scale in favor of an $(n+d)$-point grading scale, which yields the following relationships between new cut points and old ones:

$$p'_A = p_A - d, \qquad p'_B = p_B - 2d, \qquad p'_C = p_C - 3d, \qquad p'_D = p_D - 4d.$$

Perhaps unintentionally, this design generates larger changes for lower grades. This means that, e.g., students that strive for C's experience a relatively larger relaxation in standards than students that strive for A's.

We finally assume that $\xi_i \perp P$, which implies that $s_i \perp P$. This is equivalent to an assumption that scores are produced exogenously to grading policies, which allows us to directly compare the production of scores between policies. One possible violation to this would be if teachers differentially curved grades in response to policy changes.[7] We do not include the dimension of teacher response in our model because our research design is able to overcome this identification challenge.

Based on optimal effort decisions, we show the effects of a policy change are both (1) ambiguous for a given student and (2) potentially heterogeneous between types of students. To demonstrate this, we first solve the problem for low ability types in isolation and then introduce results including high ability types. Under the initial n-point policy P, low types experienced $\mathcal{G}^\ell \equiv \{B, C, D\}$. The policymaker's selection of d that forms the new $(n+d)$-point policy P' generates the new feasible grade set $\mathcal{G}^{\ell'}$. This new feasible set can be identical to $\mathcal{G}^\ell$ and generate no change in student effort, we call this a stationary policy. Alternatively, the policy can eliminate the lowest possible grade while maintaining the highest possible grade (i.e., $\mathcal{G}^{\ell'} = \{B, C\}$), which leads student to decrease their effort. We term this a contractionary policy. Finally, the policy can introduce a new highest possible grade (i.e., $\mathcal{G}^{\ell'} = \{A, B, C\}$ or $\mathcal{G}^{\ell'} = \{A, B, C, D\}$), leading students to increase their effort. We call this an expansionary policy.

[6] The opposite results will hold if instead P' is less lenient.

[7] While we do not model the role of teachers' discretion in assigning grades, further work could expand our model to incorporate this in the style of Diamond and Persson (2016).

Under a stationary policy, it will necessarily be the case that $\Delta e^*_\ell = 0$. However, whenever the lenient policy is contractionary, students will reduce their effort absolutely:

$$e^{*'}_\ell = a^\ell \left(\frac{\phi_B}{\bar{\xi} - \underline{\xi}} \right) \frac{\gamma}{\kappa} < a^\ell \left(\frac{\phi_B + \phi_C}{\bar{\xi} - \underline{\xi}} \right) \frac{\gamma}{\kappa} = e^*_\ell \quad \Longrightarrow \quad \Delta e^*_\ell < 0.$$

Conversely, an expansionary policy increases effort, regardless of whether the lower bound is changed.[8] Formally, for the case in which $\mathcal{G}^{\ell'} = \{A, B, C\}$,

$$e^{*'}_\ell = a^\ell \left(\frac{\phi_A + \phi_B}{\bar{\xi} - \underline{\xi}} \right) \frac{\gamma}{\kappa} > a^\ell \left(\frac{\phi_B + \phi_C}{\bar{\xi} - \underline{\xi}} \right) \frac{\gamma}{\kappa} = e^*_\ell \quad \Longrightarrow \quad \Delta e^*_\ell > 0.$$

Therefore, a student whose feasible grade set does not include either an A or an F has an entirely ambiguous response to a lenient grading policy. The effect will depend on both the size of the policy d and the relative span of their feasible score set.

In the case of the high ability student, the effects are less ambiguous. Due to the fact that $\mathcal{G}^h$ includes an A, no lenient policy P' can have an expansionary effect on high ability students. Following the previous analysis, if $\mathcal{G}^{h'} = \{A, B\}$, then $\Delta e^*_h < 0$; otherwise, high ability students will not change their effort in response to P'. This leads us to our next result:

> *Result 2:* A new grading policy P' which is more lenient than the currently enacted policy P will have a non-positive effect on the effort exerted by the highest ability students. The effect for lower ability students is ambiguous and depends on both the magnitude of the policy's leniency and the relative capability of these students.

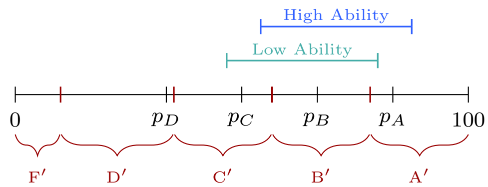
We conclude this section by graphically demonstrating the ambiguity of the effects of these types of policies, as depicted in Figure 2. In each of the six panels, we consider the same selection of d but vary the relative location of the score set for each ability type.[9]

Note that, by construction, P' induces a higher expected grade for all students. If effort embodies different measures of student engagement like attendance or study hours, then effort is a productive input in the accumulation of human capital. A reduction in effort which coincides with an increase in student GPA would suggest an inflation in grades which is artificially driven. However, lowering standards can result in a net increase in effort across students, as evident in Figure 2a. In this instance, high-achieving lower ability students are able to earn previously unobtainable grades. At

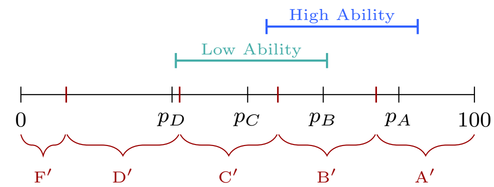
[8] In the most extreme case, which we ignore as unrealistic, a policy could make the set of feasible grades singular, which would actually decrease effort to its minimum.

[9] We can show the same six outcomes if we instead fix the distribution of students but allow the choice of d to vary.

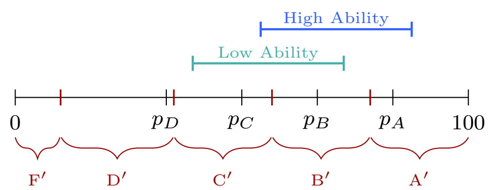
Figure 2: Ambiguity in Policy Effects for Different Ability Types



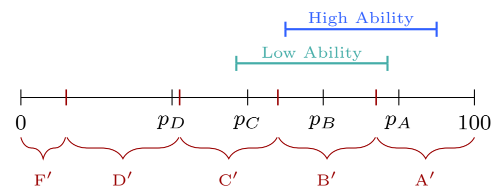
(a) $\Delta e_\ell^* > 0$ and $\Delta e_h^* = 0$



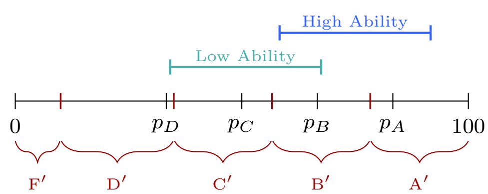
(b) $\Delta e_\ell^* = 0$ and $\Delta e_h^* = 0$



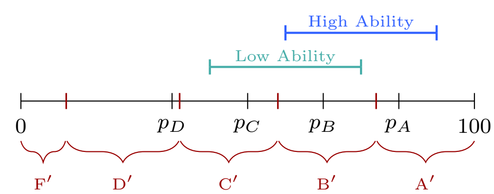
(c) $\Delta e_\ell^* < 0$ and $\Delta e_h^* = 0$



(d) $\Delta e_\ell^* > 0$ and $\Delta e_h^* < 0$



(e) $\Delta e_\ell^* = 0$ and $\Delta e_h^* < 0$



(f) $\Delta e_\ell^* < 0$ and $\Delta e_h^* < 0$

NOTES: The above figures illustrate different policies and ability distributions which could result in heterogeneous responses among students. In each panel, the original cut points for policy P are denoted in black. The new cut points corresponding to policy P' are denoted by the red lines, with the corresponding new grade regions outlined by the red braces and prime letters.

the same time, d is chosen to be small enough so that high ability students do not reduce their effort. Importantly, this policy represents a possible way to mitigate achievement gaps.[10]

Alternatively, lowering grading standards could instead exacerbate the achievement gap. For example, the policy enforced in Figure 2c results in a widening of the achievement gap (low ability students reduce their effort while high ability students maintain their effort) despite the fact that an identical policy reduced the gap in Figure 2a.

Overall, these disparate predictions point to the importance of policy design that is relevant to

[10] Figure 2e also depicts a possible way to reduce the achievement gap, albeit without improving the human capital accumulation of any group of students. Similarly, Figure 2d showcases a policy which reduces engagement among high ability students while boosting effort among low ability students.

the target student population. These predictions also help to explain why the literature on grade inflation has mixed—and at times incongruous—findings. The key takeaway of this model is that student response depends in part on the *magnitude of academic leniency* induced by the policy and the *discrepancy and spread* of the student score distribution—which is a function of their abilities and the return to their effort. As such, a relaxation of grading standards may lead students to decrease their investments in school in some contexts (Betts and Grogger, 2003; Figlio and Lucas, 2004; Babcock, 2010; Nordin, Heckley and Gerdtham, 2019; Hvidman and Sievertsen, 2021), while at the same time motivate and benefit students in other contexts (Dee et al., 2016; Ahn et al., 2019; Minaya, 2020). As educators and policymakers seek to change grading standards in their school districts, it is important that they understand that the heterogeneity and the direction of the effect will depend on both the score distribution of their student population and the magnitude of their grade change.

3 Institutional Background

3.1 Standardization of High School Grading Policies

Unlike most states in the U.S., the North Carolina State Board of Education explicitly outlines grading standards for all public high schools in the state. These standards include grading scales which reflect the correspondence of numeric scores to letter grades. In the Fall of 2014, the North Carolina State Board of Education voted to standardize high school grading policies to a 10-point scale in an effort to increase comparability between school districts and increase the competitive quality of students applying to colleges.[11] Table 1 outlines the specific changes associated with each letter grade.

As shown in Table 1, the change in letter grade standards created an additional 10-point buffer at the margin of passing a class. For example, a student taking 9th grade Math I in the 2014-2015 school year would need a 70 or higher to earn credit for the course, while a student in that same class the following year would instead need a minimum grade of 60. In addition to the 10-point decrease associated with a passing grade, the lowest numerical value to obtain any letter grade above an F also decreased with the new grading policy. Thus, this change represents a relaxation

[11] Two months prior, the state additionally approved the adjustment of quality point premiums associated with advanced courses (such as Honors, AP/IB, and college courses) to reflect a maximum GPA of 5.0. In this paper, we will focus on 9th grade student outcomes, therefore, the relevant policy change is the change in letter grade standards. See North Carolina's General Statute 116-11(10a) for more details.

Table 1: Changes in Academic Course Grades

Letter Grade (Grade Point)	Original	New
A (4.0)	93 – 100	90 – 100
B (3.0)	85 – 92	80 – 89
C (2.0)	77 – 84	70 – 79
D (1.0)	70 – 76	60 – 69
F (0.0)	0 – 69	0 – 59

NOTES: The above table displays the changes in grade thresholds as a result of the policy change. The first column displays the associated letter grade and points (used to calculate GPA) for a given threshold. "Original" refers to the standards mandated by the state prior to the 2015-2016 school year, and "New" refers to the updated ones. For example, a 91 corresponds to a B (3.0 grade points) in 2014 and an A (4.0 grade points) in 2016.

in standards at every letter grade value.

Students may also experience changes in grading standards through teacher grading practices. To assess the extent to which state-mandated changes in grading standards led to changes in grading practices, Figure 3 displays histograms of 9th grade math course grades for the first year of the policy (2016) and the year prior (2015). For all letter grade thresholds, these distributions show bunching just to the right of letter grade cutoffs, especially at the margin of passing a course. The comparison of 2015 to 2016 histograms shows how the distribution of numerical grades immediately shifted in response to the 2016 change in grading standards. We take this as evidence that assignment of grades is endogenous to statewide standards. Taken together, this shows how the new change in standards generated a real change in grading leniency across all school districts in the state.

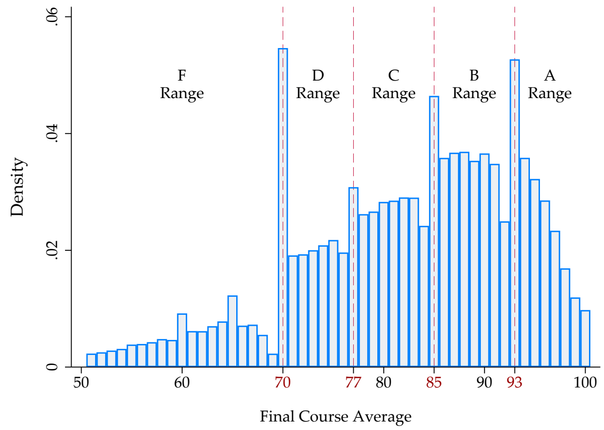
Next, we ask how the change in standards shifted students' grades earned in 9th grade. We do so by exploring the prevalence of letter grade frequencies for all core academic courses before and after the policy.[12] Figure 4 shows the percentage of letter grades for all core academic courses over time relative to 2015, the year prior to the North Carolina policy change. This figure shows the share of A's earned in core academic courses increased in 2016 by almost 20% and persisted through 2018. Furthermore, the share of F's obtained decreased by 20% in 2016 compared to the prior year. The shares of B's, C's, and D's changed substantially less, though all persistently decreased.

Given the stark shift at the top and bottom of the grade distribution shown in Figure 4, we explore changes in student GPA based on students' prior performance. We use student 8th grade math end-of-grade (EOG) exam scores to proxy for incoming academic preparedness. Figure 5 provides the change in GPA for 9th graders relative to 2015 across student academic preparedness. We report average GPA change for all 9th grade students, 9th graders belonging to the lowest decile

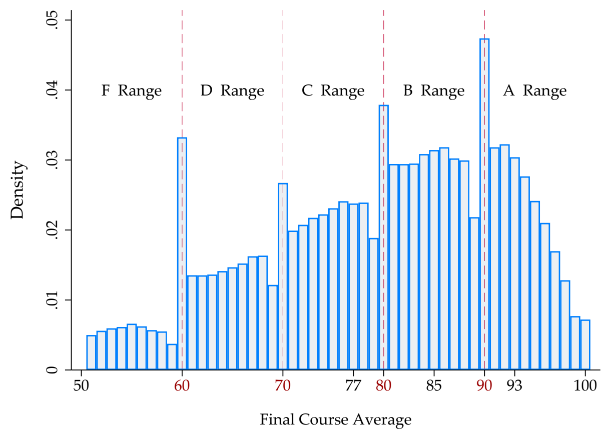
[12] We define core academic courses as courses in the subjects of math, English, science, and social studies.

Figure 3: Distribution of 9th Grade Math Final Course Grades

(a) 2015



(b) 2016



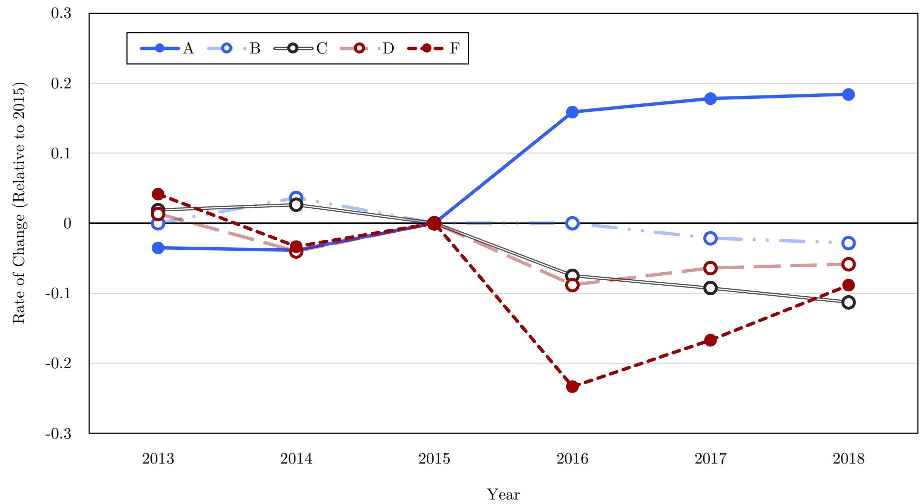
NOTES: These histograms plot NCERDC transcript-level math grades for 9th graders in 2015 (panel (a)) and 2016 (panel (b)). Course averages written in red denote the cutoff minimum for each corresponding letter grade; e.g., in the 2014-2015 school year, a 93 corresponds to the lowest numeric course average which earns a grade of A. We censor grades above 100 to have the value 100 and omit grades below 50 from the figures. The distribution for other course types follows a similar pattern. See Figure A.4 for equivalent histograms of an additional pre and post treatment year (2014 and 2017).

of 8th grade math EOG score distribution, and 9th graders belonging to the highest decile. Despite all students exhibiting immediate gains in GPA, lower achieving students return to pre-policy levels, while higher achieving students maintain GPA gains over time. Taken together, Figure 5 suggests that the primary driver of increases in GPA came from the newly relative ease in earning an A, which translated to GPA gains that were accrued mostly by higher achieving students.

3.2 Minimum Age Requirement for School Entry

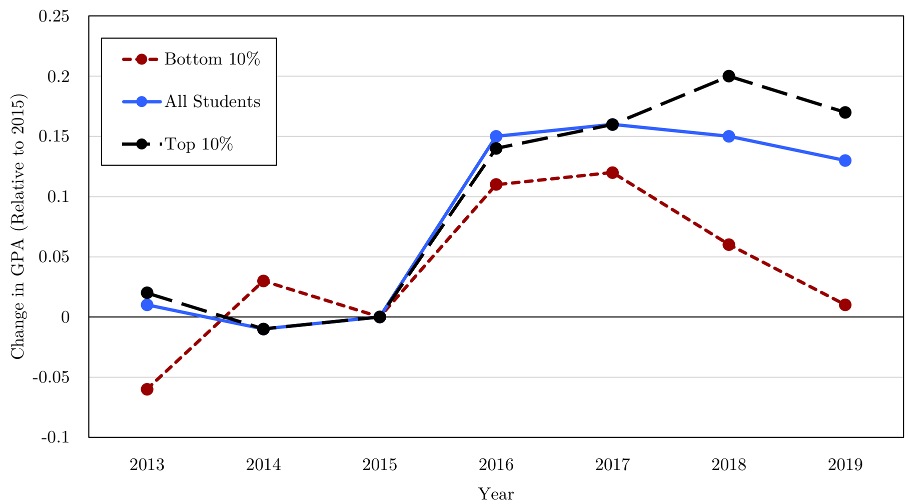
The second source of identifying variation used in our research design relies on North Carolina's Minimum Age Requirement for school entry. Under North Carolina's General Statute 115C-364,

Figure 4: Changes in Letter Grade Shares



NOTES: This figure plots the changes in the proportion of each letter grade earned in core academic courses. We standardize levels relative to 2015, the last year before the policy change occurred.

Figure 5: Changes in 9th Grade GPA



NOTES: This figure reports relative unweighted 9th grade GPAs in core classes only for students who took the 8th grade math EOG exam, the most-taken EOG for 8th graders. Approximately 90% of all 9th graders have valid scores. The subsamples of students listed above respectively refer to those scoring at or below the 10th percentile in the distribution of EOG scores, all students taking the EOG, and those scoring at or above the 90th percentile. We standardize levels relative to 2015, the last year before the policy change occurred.

children aged 5 years old before October 17 of that school year are entitled to entry into the public school system. In practice however, the timing of school entry can be influenced by parents or caregivers. As a result, not all students are perfectly assigned to cohorts based on North Carolina's minimum age requirement rule. Our empirical specification flexibly accounts for this non-compliance of the kindergarten entry rule to recover a local average treatment effect (LATE) estimate of the effect of academic leniency on student outcomes.

4 Data

We rely on rich administrative data on the universe of North Carolina public school students from the 2013-2014 to 2018-2019 school years provided by the North Carolina Education Research Data Center (NCERDC).[13] These data allow us to capture key information on student learning and engagement.

To examine the effects of grade leniency on student learning, we draw from student performance on Math I standardized tests for two reasons. First, Math I is a required course for graduation that most students take in 9th grade, the first grade for which there are transcripts available. Second, the state of North Carolina assesses student learning in Math I every school year via the End-of-Course (EOC) state assessments. Thus, focusing on Math I allows us to observe both student performance in the class (numeric marks) and well as student learning (test performance).

While prior literature (Hastings, Neilson and Zimmerman, 2012; Jackson, 2018) has used GPA and absences as proxies for student engagement, our policy mechanically shifts GPA in a way that may obscure student response. Thus, we analyze student GPA and absences separately, and we rely primarily on student absences and chronic absenteeism to proxy for student engagement.[14] We recover annual GPA from administrative records of high school transcripts from core courses.[15] Transcript data include the courses students take, their code, subject, and final numeric mark obtained. One notable limitation to the data provided by NCERDC is that student birth dates are anonymized at the month level (see Cook and Kang (2016) for a greater discussion). Given that our research design relies on the quasi-random assignment of students to cohorts based on their date of birth, we supplement NCERDC data with restricted-access data on students' exact birth dates provided by the North Carolina Department of Public Instruction.

We make several restrictions to our sample for the purposes of precise empirical estimation. First, we omit students belonging to charter high schools. Charter schools generally have different courses and can grade students differently. Thus, we omit them from the estimation to ensure comparability of GPAs across students. Second, we drop students with disabilities from the analysis. We do this primarily because these students may not participate in standard 9th grade courses

[13] Throughout this paper we use the spring term year to refer to each school year.

[14] We follow the state's definition of chronic absenteeism as being absent 10% or more of the days in the school, which translates to 18 days.

[15] We define core academic coursework as ELA, Math, Science, and Social Studies

Table 2: Descriptive Statistics

	Regression Discontinuity Sample		Full Sample	
	(1) 2015 9th graders −2 mo. Oct. 17	(2) 2016 9th graders +2 mo. Oct. 17	(3) 2015 9th graders All	(4) 2016 9th graders All
Demographic				
Female	0.521	0.503	0.517	0.513
	(0.500)	(0.500)	(0.500)	(0.500)
White	0.470	0.458	0.477	0.467
	(0.499)	(0.498)	(0.499)	(0.499)
Asian	0.020	0.019	0.019	0.020
	(0.140)	(0.136)	(0.137)	(0.138)
Black	0.274	0.290	0.287	0.285
	(0.446)	(0.454)	(0.452)	(0.451)
Hispanic	0.177	0.174	0.157	0.171
	(0.382)	(0.380)	(0.364)	(0.377)
Other	0.059	0.058	0.060	0.058
	(0.236)	(0.234)	(0.237)	(0.234)
Socioeconomic				
EDS	0.719	0.726	0.714	0.715
	(0.450)	(0.446)	(0.452)	(0.452)
Rural	0.540	0.528	0.539	0.529
	(0.498)	(0.499)	(0.498)	(0.499)
Academic				
LEP	0.067	0.053	0.057	0.054
	(0.250)	(0.224)	(0.232)	(0.225)
Gifted (Math)	0.038	0.045	0.040	0.040
	(0.191)	(0.208)	(0.196)	(0.196)
Gifted (Reading)	0.038	0.049	0.039	0.044
	(0.192)	(0.216)	(0.194)	(0.205)
N	10,479	10,140	57,458	61,589

NOTES: This table shows descriptive statistics of student demographic, socioeconomic, and academic characteristics. Columns (1) and (2) present descriptive statistics of the sample used in our regression discontinuity estimation. Columns (3) and (4) present descriptive statistics for all 2015 and 2016 9th graders in the state.

required for graduation, such as Math I. Next, we remove students retained in the 9th grade but keep students retained in any other grade. Doing so eliminates the possibility for students close to the birth date cutoff to be considered both non-treated (9th grader in year t) and treated (9th grader in year $t+1$). Lastly, we restrict the analysis to those students with a valid Math I test score.[16]

Table 2 presents an overview of the data used in our analysis. Columns (1) and (2) present

[16] Despite testing the concepts covered in a traditional 9th grade math course, only 66% of Math I EOC test takers belong to the 9th grade. In 2015 and 2016, 27% of test takers were in the 8th grade, while 6% were in the 10th grade. As a result, this restriction drops nearly 40% of 9th graders, most of whom took the exam in the 8th grade.

descriptive statistics of the sample used in our regression discontinuity estimation. Students in column (1) make up our counterfactual group. They have birth dates just to the left of the cutoff and are therefore assigned to the 2015 9th grade cohort. Conversely, students in column (2) have birth dates just to the right of the cutoff and are assigned to the 2016 9th grade cohort, the first to experience more lenient grading. Columns (3) and (4) present descriptive statistics for all 2015 and 2016 9th graders in the state. Overall, statistics shown in Table 2 demonstrate a balance of student characteristics between treatment and comparison groups of our regression discontinuity sample.[17] When comparing students in our regression discontinuity sample to those of the whole state (columns (3) and (4)), we see that students close to the bandwidth are similar in their observable characteristics to the state average. While we still maintain a local interpretation of our treatment effects, we emphasize that the local sample is highly characteristic of the population of interest. Finally, a comparison of statistics between columns (3) and (4) suggest strong balance across the two 9th grade cohorts of interest.

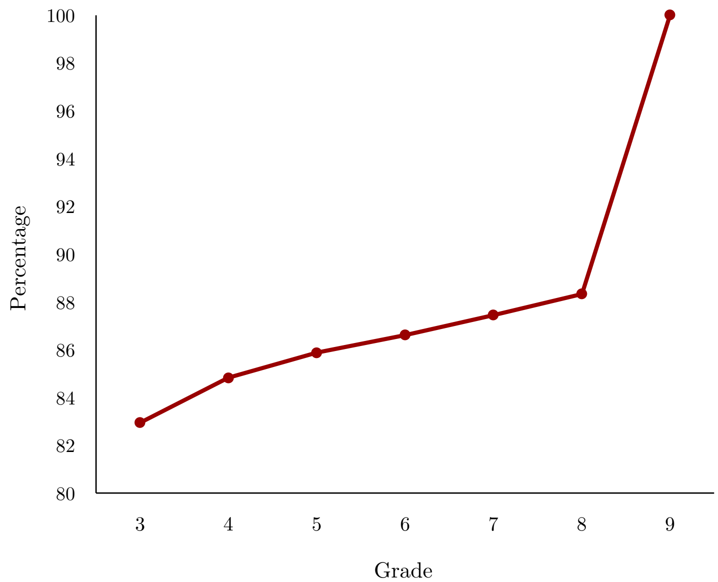
5 Research Design

Identifying the causal effect of academic leniency is challenging for several reasons. First, the relaxation of stringent standards often occurs gradually over time. Year-to-year changes in how grades are allocated tend to be small, if not completely indiscernible. This means that testing for sizeable effects related to grade inflation requires a large enough panel to capture observable differences in the distribution of grades and student achievement. A simple comparison of cohorts at either end of a sufficiently large panel would exhibit bias induced by unobserved time trends. These trends may also induce a response in student behavior unrelated to—and misattributed to—changes in academic standards. To circumvent this issue of an incremental change in *de facto* leniency, we exploit the roll-out of the new grading policy in North Carolina. This policy explicitly outlined how schools would distribute letter grades and did so abruptly from one year to the next. Importantly, the immediacy in which the policy was announced, approved, and implemented provides us with two comparison groups that are likely unaffected by spurious time trends.

A second important issue to consider when analyzing the effects of academic standards is the fact that districts might differ in their explicit grading schemes (e.g., the relative course average needed to earn a letter grade of A) or their leniency *conditional* on the scheme (e.g., the share

[17] See Section 7 for formal tests of covariate smoothness across the threshold.

Figure 6: First Instance of Non-Compliance Among Non-Compliers



NOTES: We condition on being a delayed student (relative one's age and North Carolina law) in 9th grade. Here, we include only students with North Carolina administrative records for every year available. The rates do not change considerably when focusing instead on students who, e.g., first appear in the data in the sixth grade. In fact, 86.2% of delayed ninth graders who first enter the panel in grade six exhibit delay in grade six, compared to 86.9% in the figure above.

of A's given). The universality of the North Carolina policy addresses this first concern directly. Each district in North Carolina was mandated to follow the 10-point scale, regardless of the scale they utilized during the previous academic year. However, we find no evidence to suggest that any district operated on a grading scale other than the 7-point scale in the year prior. Put together, this provides us with state-wide variation in explicit academic standards that all students experience in the same way. While this policy design does not account for differences in between-district levels of leniency, the inclusion of district-level fixed effects in an estimating regression does.

The third and final challenge, which the policy alone cannot address, is selection into academic cohorts. Deming and Dynarski (2008) show that parents' incentives to "red-shirt" their children has risen over time. This process delays a child's entry into kindergarten with the intended goal of providing them with age advantages during their time in the public school system. The selection on gains induced by this decision has the potential to bias our estimates, especially if parents' strategic incentives differed substantially between the two cohorts of interest. Related to this, our focus on 9th grade students requires consistent matriculation across grades, especially during the transition from middle school to high school. We find evidence of the contrary. In fact, as evident in Figure 6, a large, discontinuous mass of students are held back in 8th grade relative to the rate in any earlier grade. While we cannot distinguish why students experience an additional year in the 8th grade, this selection into delayed 9th grade matriculation introduces a potential non-compliance problem when estimating the effect of the policy. For this reason, a direct comparison of outcomes between the treated (the first cohort to experience the 10-point scale) and the non-treated (the last cohort to experience the 7-point scale) will be distorted by endogenous selection bias. To overcome this, we construct a mechanism which randomly allocates students either to the lenient or the strict policy by combining the timing of the implementation with North Carolina's minimum age requirements for school entry.

The two cohorts we compare faced the same kindergarten entry rule: if a child turns 5 before October 17, they may enter the public school system that year; otherwise, they are delayed entry and must enroll in the subsequent year. This assignment rule can be extended to high school entry, adjusting for probabilistic differences attributable to retention. Conditional on a smooth distribution of birthdays, assignment to high school entry year is then as good as random for a small bandwidth of birthdays around the cutoff. The selection into academic cohorts outlined previously leads us to estimate a fuzzy regression discontinuity design, or two-stage least squares (2SLS) model, which adjusts for the level of non-compliance and recovers estimates which carry a causal interpretation (Hahn, Todd and Van der Klaauw, 2001).

Formally, we estimate the following 2SLS specification,

$$d_{ij} = \delta_0 + \delta_1 \mathbb{1}_{\{b_i \geq 0\}} + g(b_i) + \boldsymbol{\delta}' \boldsymbol{x}_i + \alpha_j + \eta_{ij} \tag{2}$$

$$y_{ij} = \gamma_0 + \tau \hat{d}_{ij} + g(b_i) + \boldsymbol{\gamma}' \boldsymbol{x}_i + \alpha_j + \varepsilon_{ij}, \tag{3}$$

where i indexes students and j indexes school districts. The variable d_{ij} is a treatment indicator that takes on the value 1 if student i belongs to the first cohort to experience the 10-point scale, 0 if they belong to the last cohort to experience the 7-point scale. The running variable b_i represents the relative distance from student i's birth date to the North Carolina entry cutoff of October 17, which we standardize to 0. As such, $\mathbb{1}_{\{b_i \geq 0\}}$ is an indicator for whether the birthday of student i is to the right of the kindergarten entry rule and $g(\cdot)$ is a flexible polynomial in the birth date of students at both sides of the cutoff. The vector of student-level covariates, $\boldsymbol{x}_i$, includes student demographic characteristics and socioeconomic characteristics. To control for any unobserved district-level effects, we include district fixed effects in α_j. The variable y_{ij} denotes any outcome of interest, such as GPA, test scores, and attendance. The residualized decision rule is captured by $\hat{d}_{ij}$ and *iid* idiosyncratic error in each stage is denoted, respectively, by η_{ij} and ε_{ij}. Given this framework, our parameter of interest is τ, which captures the local average treatment effect of being randomly assigned to the lenient academic standard. Due to our use of a discrete running variable with a small number of possible values, we follow Kolesár and Rothe (2018) and Armstrong and Kolesár (2020) in estimating honest confidence intervals from a fuzzy regression discontinuity design without covariates.[18] We then construct the same estimator under a 2SLS framework and introduce covariates and fixed effects to obtain our fully-specified model. Our results are robust to specification level. In each specification, we estimate local linear regressions with a triangular kernel density as outlined in Fan and Gijbels (1992) and report robust standard errors.

Several studies in the economics of education utilize date of birth in a similar fashion, beginning with Angrist and Krueger (1991) and extending to include Elder (2010), Navarro-Palau (2017), Dee and Sievertsen (2018), Ordine, Rose and Sposato (2018), and Persson, Qiu and Rossin-Slater (2021). Outside of the context of education, exact-date-of-birth RDDs have been used to study the effects of vaccine eligibility (Humlum, Morthorst and Thingholm, 2022), access to Medicare (Goldsmith-Pinkham, Pinkovskiy and Wallace, 2021), and retirement reform on household retirement decisions (Stancanelli, 2017), to name a few. Our design most closely matches Cook and Kang (2016), who utilize kindergarten entry cutoffs to measure longer-run outcomes like middle school test score performance and the propensity to commit crime. We build on this earlier work by combining the birthday RD design with the exogenous timing of a policy change, which allows us to identify the causal effect of the policy directly.

6 Results

In this section, we present regression discontinuity estimates of Equation 3 for North Carolina 9th graders. First, we present evidence of treatment in Section 6.1. Then, we present causal effect estimates of increased grade leniency on student academic and engagement outcomes in Section 6.2. Finally, we presents longer-run effects in Section 6.3.

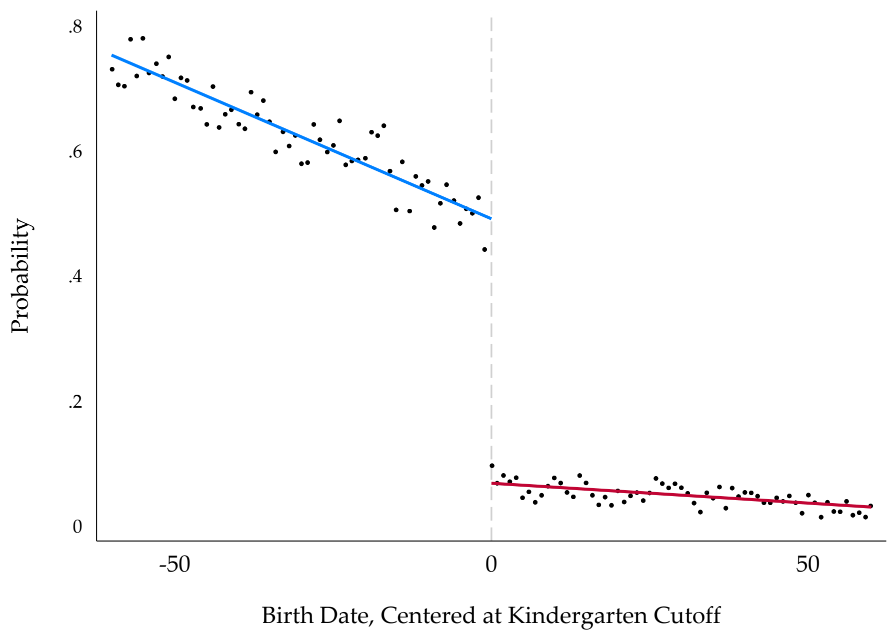
[18] For completeness, we report the rest of the Armstrong and Kolesár (2020) procedure and our related assumptions. We assume the parameter space corresponds to a second-order Hölder smoothness class, bounding second derivatives globally by a constant K_s for stage s. In determining the value of $\boldsymbol{K} := (K_1, K_2)$, we run quadratic regressions on the left-hand side of the cutoff for a large, viable window of the running variable, which we establish as 70 days. We then estimate $d_i = \lambda_0 + \lambda_1 b_i + \lambda_2 b_i^2 + e_i^1$ under the restriction $b_i \in \{-70, ..., -1\}$ and set $K_1 = 4|\hat{\lambda}_2|$. Similarly, we estimate $y_i = \mu_0 + \mu_1 b_i + \mu_2 b_i^2 + e_i^2$ and then set $K_2 = 4|\hat{\mu}_2|$. Our results are not sensitive to the selection of $\boldsymbol{K}$, as shown in Table A.1.

6.1 Evidence of Treatment

Before presenting the main estimates, we provide evidence to establish how birth date affects assignment to the first treated cohort that experienced the new grading policy in North Carolina. Under our regression discontinuity design, we argue that kindergarten entry rules randomly assigned students to the first cohort to experience explicit academic leniency. Crucial to our identification strategy is the power of these entry rules in the assignment of students into cohorts.

Figure 7 plots the likelihood of entering high school in 2015 based on student distance to the birth date cutoff. This figure shows a strong discontinuity at the cutoff, of approximately 45 percentage points, on the probability of assignment to the first treatment cohort. This result establishes the relevance of our research design.

Figure 7: Discontinuity in High School Entry



NOTES: The above figure plots the discontinuity in the probability of entering high school in 2015 for students born in the fall of 2000. The threshold corresponds to October 17, the kindergarten entry cutoff in North Carolina. This discontinuity is consistent with discontinuities of cohort assignment in other comparison years (see Table A.3).

The discontinuity presented in Figure 7 is not sharp, which suggests that some students either do not comply with state-mandated kindergarten entry rules or are retained prior to 9th grade. We interpret non-compliance as a combination of both parents requesting waivers to delay their child's entry date for kindergarten and students undergoing retention.[19] Among 9th grade students

[19] School entry and progression is ultimately a choice of parents or caregivers. Figure A.1 presents a decision tree that illustrates how non-compliance can come about.

who did not comply with the kindergarten entry rule, 86% are non-compliant by 3rd grade, which indicates that student non-compliance is not primarily driven by 9th grade policy. However, the discontinuous jump in retention, coupled with the evidence of non-compliance in Figure 7 validates our need to instrument for student selection into academic cohorts.

6.2 Causal Effects on 9th Grade Student Outcomes

We first examine the effects of grading leniency on student GPA, which is a direct outcome of the policy change. However, interpreting effects on student GPA is challenging because any change in GPA could reflect both the mechanical increase driven by the policy and/or true student response. Thus, we further explore effects on student learning and student engagement through student Math I EOC scores and absences. We provide visual evidence of discontinuities around the cutoff for each of these outcomes in Figure 8. These plots suggest that more lenient grading standards led to higher student GPAs and higher absences.

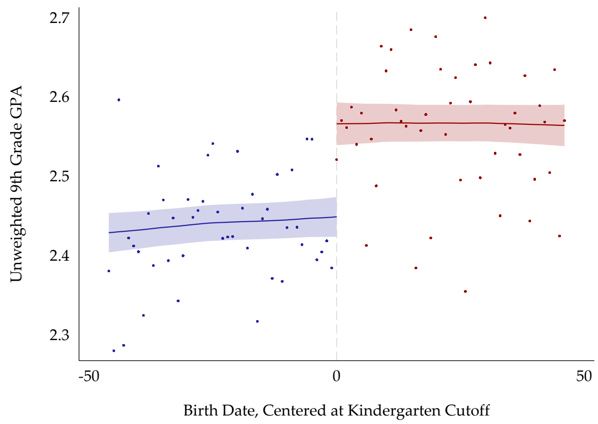
Table 3 presents estimates of Equation 3 for these student outcomes. These estimates indicate that increased grade leniency resulted in an increase of approximately 0.266 GPA points. This boost in GPA corresponds to an 11% increase. Changes in student GPA can reflect changes in student effort as well as mechanical increases driven by the policy. To understand the extent to which this mechanical effect likely increased GPAs, we draw from pre-treatment (2015) student numeric grade data and recover a *simulated* GPA using the new grading scale (2016). The difference between actual and simulated GPA measures helps us understand how GPA would change in the absence of student/teacher response.[20] This difference is estimated to be 0.298 points.[21] Given how close our regression discontinuity estimate is to this simulated change, we interpret the regression discontinuity effects as mechanical, rather than a reflection of true student response. Estimates in column (2) of Table 3 further support this interpretation, as they indicate that increases in GPA were not accompanied by greater student achievement in Math I EOC scores.

Finally, estimates for student absences in columns (3) and (4) indicate that increased grade leniency led students to become more absent in school. Students to the right side of the birth date cutoff are, on average, 1.3 days more absent in school, than students just to the left side of the cutoff. This corresponds to a 20% increase in school absences. Estimates of the effects on chronic

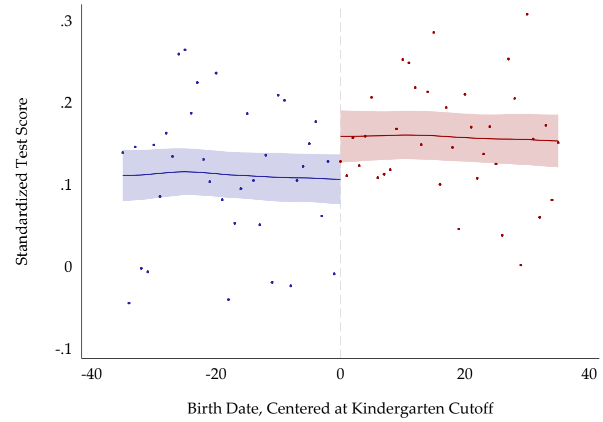
[20] Under the assumption that, in the absence of the new grading policy, post-treatment numeric grade distributions would have been similar to pre-treatment ones.

[21] This is likely an underestimate, with the true mechanical value exceeding 0.3 points. A non-negligible portion of transcripts in the data report course grades as letters instead of the numeric average, meaning we cannot re-scale these grades in the simulation.

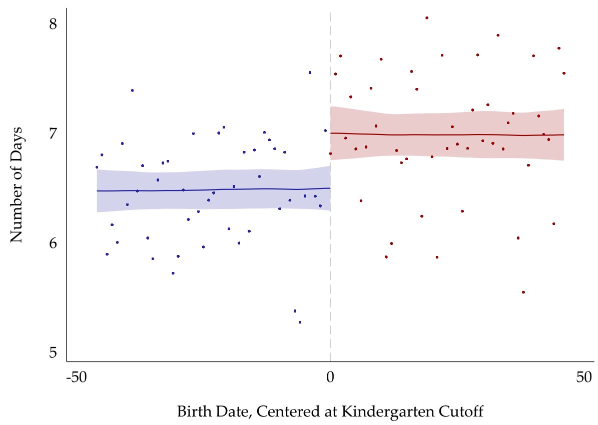
Figure 8: Changes in 9th Grade Outcome Across Birth Date Cutoff



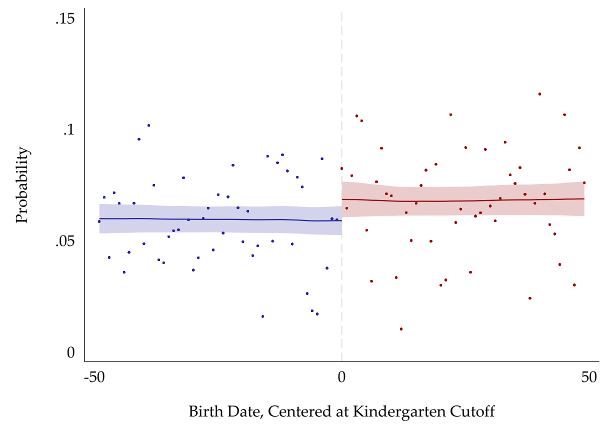
(a) 9th Grade GPA



(b) Math I EOC Score



(c) 9th Grade Absences



(d) 9th Grade Chronic Absences

NOTES: This figure shows discontinuities in 9th grade unweighted GPA, Math I EOC scores, 9th grade absences, and the likelihood of 9th grade chronic absence (panels (a), (b), (c), and (d)). All panels are created by plotting the average outcomes across each birth date and fitting the data using a polynomial of degree zero on either side of the cutoff, which yields mean smoothing local to the cut point side. For each outcome, we fix the bandwidth to match the optimal one estimated using the procedure outlined in Section 5. In an effort to match the honest confidence interval detailed in Armstrong and Kolesár (2020), we set the confidence level to 97.07% for the bands displayed.

absenteeism rates indicate that increases in school absences seem to be driven by students on the upper tail of the absences distribution, as threshold crossing increases student chronic absenteeism rates by 0.048 points.

The model presented in Section 2 shows how differences in costs and returns to effort across students of different levels of academic preparedness can generate differential response to changes in grading standards. Descriptive statistics presented in Section 3 further support this claim by showing important differences in the ways that increased grading leniency affected high and low achieving student groups. We unpack the main results of this paper by exploring heterogeneous

Table 3: Regression Discontinuity Estimates for 9th Grade Outcomes

	(1) Core Academic Grade Point Average	(2) Math I EOC Score (Standardized)	(3) Total Number of Days Absent	(4) Probability of Chronic Absence
10-Point Scale	0.266	0.063	1.324	0.048
	(0.071)	(0.089)	(0.616)	(0.020)
97.07% CI	(0.111, 0.420)	(-0.130, 0.257)	(-0.019, 2.667)	(0.003, 0.093)
1st Stage Estimate	0.434	0.431	0.434	0.434
Mean (Left)	2.436	0.108	6.479	0.059
Bandwidth	46.27	35.65	46.24	49.96
Observations	13,304	10,122	13,304	14,189

Notes: Means to the left are calculated manually over a bandwidth range $\{-h^*, ..., -1\}$, where h^* is the MSE-optimal bandwidth. We report 97.07% confidence intervals to align with the honest critical value of 2.18 outlined in Armstrong and Kolesár (2020). Results come from regressions which include controls and district fixed effects. The covariates used include indicators for gender, race/ethnicity, socioeconomic status, and gifted status in reading and math.

response to treatment across students with different levels of prior academic preparedness in Table 4.[22] To approximate prior academic preparedness, we split our sample of 9th graders into two groups based on their 8th grade Math EOG performance in panel A of Table 4.[23] Differences across these two groups are statistically significant for all outcomes. These estimates suggest that gains in GPA are driven mostly by gains from students on the upper end of the test score distribution. Students with 8th grade test scores above the median see an increase in GPA of 0.296 points, while students below the median see no gains in GPA.

This finding is striking for two reasons. First, the policy mechanically increases GPA for all students regardless of their position in the numeric grade distribution. Second, since the change in letter grade standards created an additional 10-point buffer at the margin of passing a class (see Table 1), *a priori*, we would expect this policy to boost GPA for lower achieving students the most. In fact, estimates of simulated changes in GPA for students on the lower end of the distribution yield a mechanical increase of 0.341 points. This change is estimated to be 0.250 points for students on the upper end of the distribution. Thus, taken together, null GPA effects for students on the lower end of the test score distribution imply that student effort response for this group is such that it completely offsets a mechanical increase that was likely to benefit them the most.

Estimates for absences further support this claim. These results suggest that estimated increases in absences recovered in Table 3 are driven by students on the lower end of the test score distribution.

[22] We present results across student gender, race, and SES in Table A.4. Differences across these subgroups are not statistically significant for any outcome.

[23] 91% of 9th graders in our sample have a valid 8th grade Math End-of-Grade (EOG) exam.

Table 4: Effects Across Prior 8th Grade Math EOG

	(1) Core Academic Grade Point Average	(2) Math I EOC Score (Standardized)	(3) Total Number of Days Absent	(4) Probability of Chronic Absence
Low Ability	0.065	-0.126	2.584	0.092
	(0.093)	(0.103)	(0.980)	(0.032)
High Ability	0.296	-0.111	0.373	0.024
	(0.090)	(0.112)	(0.773)	(0.026)

NOTES: We fix the bandwidth for each specification to be the same as in the full sample version. We present results from regressions which include covariates and district fixed effects. Heteroskedasticity-robust standard errors are reported in parentheses beneath each point estimate.

Students with 8th grade test scores below the median see an increase in absences of 2.6 days, while students above the mean experience no change. Thus, greater grade leniency leads lower achieving students to decrease their engagement in school, undoing any mechanical gains in GPA introduced by the policy, while higher achieving students see all the gains of the policy without significant changes in their effort allocation.

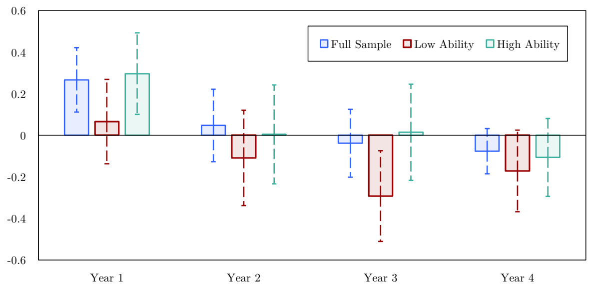
6.3 Longer-Run Effects

We provided evidence that grade leniency differentially impacted students of varying levels of prior academic preparedness during the first year of the policy's implementation. This section explores whether these differences compound over time. On one hand, the initial exposure to this policy could have caused gaps which subsequently disappeared as students adjusted to the new grading standard. On the other hand, gaps may widen each year if the effects of the policy generated lasting changes in the education trajectories of students. If the latter proves to be true, it would suggest that academic leniency and grade inflation can exacerbate achievement gaps and lessen human capital accumulation for students already at a deficit. We explore this by following our cohort of 9th graders for three additional years through high school completion.[24]

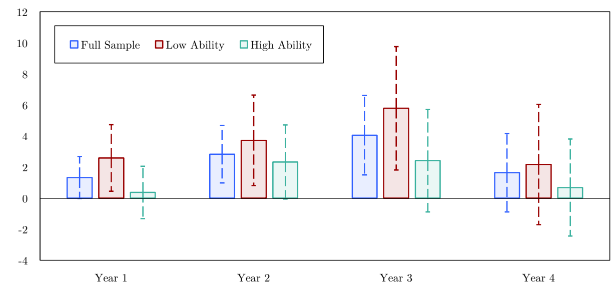
Figure 9 presents estimates of effects for years 1 through 4 for student GPA, absences, and chronic absenteeism. Year 1 estimates correspond to the estimates presented in Table 3. We do not include estimates of student test performance because Math I is only tested in 9th grade (year 1). There are two main takeaways from Figure 9. First, initial increases in GPA driven by the policy immediately fade away after year 1 for all students. This is in line with the interpretation that

[24] Notice that for this analysis we focus on *years* of high school after 9th grade rather than *grades*. We do so to account for remedial students who must repeat grades. Focusing on years rather than grades allows us to include these students in the estimation of longer-run results of exposure to more lenient grading standards.

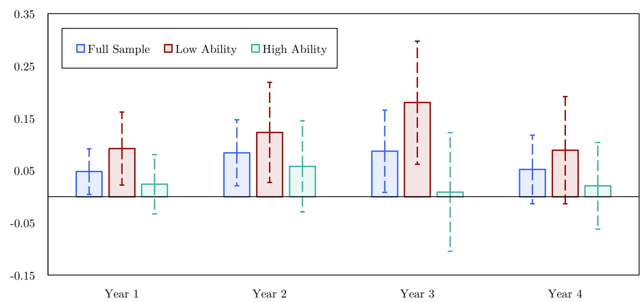
Figure 9: Effects Throughout High School, by Ability



(a) Unweighted GPA



(b) Number of Days Absent



(c) Chronic Absence

NOTES: Each bar represents the coefficient from a fully-specified 2SLS model for the listed outcome of interest. Bars capture the 95% honest confidence intervals outlined in previous sections. See Table A.5 for the table version.

the GPA boost recovered in Table 3 is driven mostly by a mechanical policy change rather than a reflection of student response in terms of effort and engagement in school. Second, the effect of increased student absences and chronic absenteeism persist and compound over time for students on the lower end of the test score distribution.

Given this compounded effect on absenteeism for students on the lower end of the test score distribution, we explore potential repercussions on economically relevant student outcomes such as ACT scores, dropping out of high school, graduating high school on time, and intention to enroll in college in Table 5. While we see no effects on any outcome for the full sample (column (1)), heterogeneous effects by pre-treatment achievement (columns (2) and (3)) suggest that grade leniency decreased ACT achievement for students on the lower end of the test score distribution. At the same time, it also led to marginal increases in high school graduation, which is consistent with grade inflation.

We highlight an important feature of these long-term comparisons that arises from treatment dosage. Recall that our research design compares students who happen to belong to the first cohort of 9th graders exposed to more lenient grading because their birth date was just to the right of the

Table 5: Estimates for Longer-Run Outcomes of Academic Achievement

		Grade 8 Math EOG	
	(1) Full Sample	(2) Below Median	(3) Above Median
ACT Score	-0.075	-0.857	-0.335
	(0.370)	(0.404)	(0.559)
Drop Out	0.002	-0.003	0.007
	(0.005)	(0.008)	(0.006)
Graduate	0.014	0.044	-0.006
	(0.019)	(0.025)	(0.029)
College Intent	0.012	-0.023	0.003
	(0.033)	(0.049)	(0.045)

NOTES: We present results from regressions which include covariates and district fixed effects. Heteroskedasticity-robust standard errors are reported in parentheses beneath each point estimate. The respective optimal bandwidth for each outcome is 31, 40, 41, and 48 days.

kindergarten entry cutoff, to students whose birth date was just to the left. For 9th grade outcomes (year 1), this means that our treatment group received lenient grading while our counterfactual group received more stringent grading in 9th grade. Comparisons for years 2 and above, however, do not recover the effect of 2 or more years of lenient grading. Instead, because the policy impacts all students in the school system once implemented, these comparisons estimate the causal effect of an additional year of exposure to academic leniency. Given this, they are are likely a conservative estimate for the underlying long-term effects.[25]

7 Internal Validity

The causal interpretation of our results depend on several assumptions related to the research design's validity. In this section, we provide four tests to validate this design and provide assurance of both the point estimates we recover and their interpretation.

7.1 Manipulation of the Running Variable

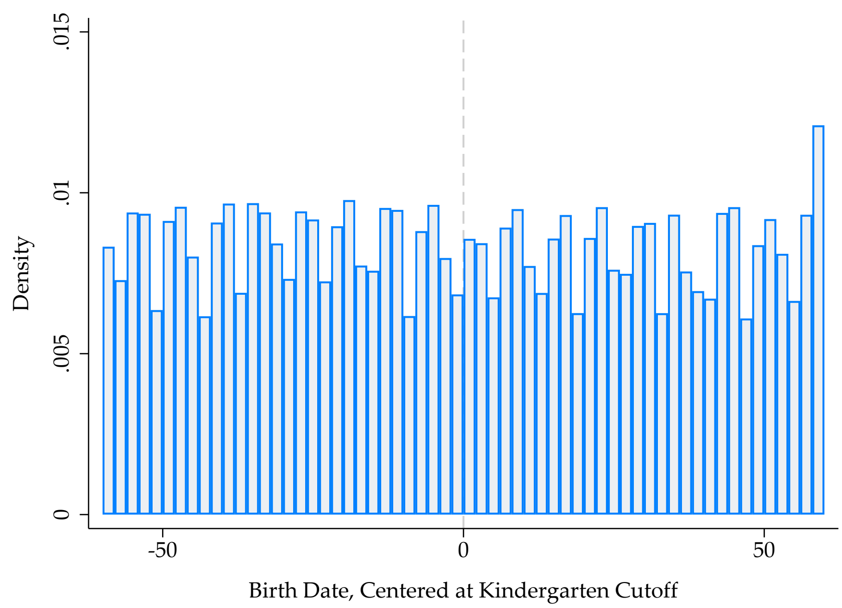
The validity of a regression discontinuity design fails whenever agents can plausibly manipulate the running variable and their likelihood of receiving treatment (McCrary, 2008). In our setting, this could take two forms. First, parents may strategically "red-shirt" their children in kindergarten

[25] For example, our year 2 estimates effectively compare 10th grade outcomes for 10th graders who received 2 academic years of lenient grading with 10th graders that received just one.

to guarantee they begin school as one of the oldest in their class (Bassok and Reardon, 2013). Second, parents could strategically retain their students if they had prior knowledge of the policy and wanted their student to experience lenient academic standards sooner.

We formally test for manipulation using the procedure outlined in Cattaneo, Jansson and Ma (2018), which estimates a local polynomial density on either side of the cutoff and tests for the null hypothesis that the limit of both functions as they approach the cutoff from either side are equal. Using a 60-day bandwidth, the test yields an associated p-value of 0.165. Thus, we fail to reject the null hypothesis that the density functions are equal at the cutoff and find no evidence of systematic manipulation of the running variable. Visual evidence presented in Figure 10 shows a smooth distribution of birthdays around the kindergarten entry cutoff, which further supports this finding.

Figure 10: Distribution of Birth Dates Around the Cut Point



NOTES: For visual ease, we bin birth dates in groups of two. Given that we plot 60 days on either side of the cutoff, we then display 30 bars in the figure above on both the left and right of 0.

With these results in mind, we proceed under the assumption that birth dates exogenously sort students to a kindergarten entry year, which maps almost directly to when they begin high school. As a result, student birth date is a valid instrument for our setting.

7.2 Covariate Smoothness

We next test for covariate smoothness across the threshold. Our empirical analysis functions as a quasi-experiment only if observable attributes of students trend smoothly at the boundary of kindergarten entry assignment. As an example, our assumption of random assignment fails if economically disadvantaged students disproportionately comply with the assignment mechanism relative to non-economically disadvantaged students. If this were the case, the comparison of students just to the left and just to the right of the cutoff would additionally capture socioeconomic effects, biasing and eliminating any causal interpretation of our results.

Table 6 presents the regression discontinuity results for student characteristics under three separate specifications.[26] In Panel A, we conduct the same two-part estimation procedure outlined by Kolesár and Rothe (2018) and Armstrong and Kolesár (2020) to obtain optimal bandwidths and honest confidence intervals. Unlike our main analysis, these tests contain no other covariates. The average optimal bandwidth among the covariates using this procedure is about 40 days. To test for bandwidth sensitivity, we repeat the analyses with a smaller fixed bandwidth ($h = 20$, presented in Panel B) and a larger fixed bandwidth ($h = 60$, presented in Panel C). For nearly every covariate, we find no evidence to suggest discontinuity around the threshold. Panel A finds small discontinuous jumps at the 5% level for female students and black students, but these discontinuities disappear for small adjustments to the bandwidth. Put together, we take these results to suggest that our assignment mechanism operates identically for different groups of North Carolina students.

7.3 Confounding Age Effects

Our third validity test examines whether any underlying age effects drive our results. The use of birth date as a running variable has the potential to introduce age effects from the fact that students to the right of the cutoff are older than students to the left by construction of the design. In primary school, this age advantage has meaningful effects (Dobkin and Ferreira, 2010; McCrary and Royer, 2011). If this effect persists through secondary schooling, then our parameter of interest, τ, may take the form $\tau := \tau^P + \tau^{\text{age}}$, where τ^P is the isolated effect of the policy and τ^{age} is the effect induced by age advantage.

[26] Figure A.3 presents analogous visual evidence using a simple linear fit over a binscatter of the relevant covariates across a 60-day bandwidth.

Table 6: Fuzzy RD Results for Demographic Characteristics with Varying Bandwidths

	Female	White	Asian	Black	Hispanic	EDS	Rural	LEP	AIGM	AIGR
Panel A: $h = h^*$										
Born after Oct. 16	-0.136	0.086	-0.000	-0.127	0.015	-0.041	-0.022	0.008	0.032	0.029
	(0.044)	(0.047)	(0.012)	(0.053)	(0.037)	(0.042)	(0.039)	(0.021)	(0.018)	(0.017)
95% Honest CI	(-0.233, -0.039)	(-0.016, 0.189)	(-0.027, 0.026)	(-0.243, -0.011)	(-0.065, 0.094)	(-0.133, 0.052)	(-0.108, 0.064)	(-0.039, 0.054)	(-0.007, 0.071)	(-0.008, 0.067)
1^{st} Stage Estimate	0.439	0.439	0.437	0.425	0.437	0.440	0.442	0.436	0.441	0.442
Mean (Left)	0.534	0.481	0.021	0.273	0.166	0.700	0.539	0.057	0.041	0.042
Bandwidth	41.89	37.59	34.52	25.96	34.60	39.36	55.29	31.38	40.00	53.73
Panel B: $h = 20$										
Born after Oct. 16	-0.127	0.051	0.007	-0.112	0.039	-0.051	-0.026	0.022	0.047	0.026
	(0.067)	(0.067)	(0.016)	(0.061)	(0.049)	(0.062)	(0.068)	(0.027)	(0.024)	(0.030)
95% Honest CI	(-0.258, 0.004)	(-0.081, 0.183)	(-0.024, 0.038)	(-0.236, 0.011)	(-0.058, 0.136)	(-0.174, 0.071)	(-0.160, 0.109)	(-0.032, 0.075)	(-0.001, 0.095)	(-0.032, 0.085)
1^{st} Stage Estimate	0.425	0.425	0.425	0.425	0.425	0.426	0.429	0.434	0.428	0.428
Panel C: $h = 60$										
Born after Oct. 16	-0.101	0.070	-0.009	-0.058	0.007	-0.017	-0.023	-0.025	0.026	0.030
	(0.037)	(0.037)	(0.009)	(0.033)	(0.028)	(0.034)	(0.038)	(0.015)	(0.015)	(0.016)
95% Honest CI	(-0.206, 0.003)	(-0.050, 0.190)	(-0.045, 0.027)	(-0.244, 0.127)	(-0.092, 0.107)	(-0.121, 0.087)	(-0.109, 0.063)	(-0.092, 0.041)	(-0.020, 0.071)	(-0.008, 0.067)
1^{st} Stage Estimate	0.439	0.439	0.439	0.439	0.439	0.440	0.442	0.441	0.441	0.441

NOTES: We restrict the sample to students with valid measures for the Math I EOC. Estimates presented follow the fuzzy regression discontinuity methods outlined in Kolesár and Rothe (2018). We report robust standard errors in parentheses. 95% honest confidence intervals are given beneath these estimates, utilizing a critical value of 2.18. Students classified by North Carolina as economically disadvantaged fall under the category "EDS". The column "Rural" captures rurality measured at the school level and "LEP" refers to whether a student is limited English proficient. The AIG variables denote whether the student is academically gifted in elementary math and reading, respectively.

Table 7: Comparison in Estimates for Placebo Sample

	(1) Core Academic Grade Point Average	(2) Math I EOC Score (Standardized)	(3) Total Number of Days Absent	(4) Probability of Chronic Absence
Panel A: Reproduced Main Analysis				
10-Point Scale	0.266	0.063	1.324	0.048
	(0.071)	(0.089)	(0.616)	(0.020)
97.07% CI	(0.111, 0.420)	(-0.130, 0.257)	(-0.019, 2.667)	(0.003, 0.093)
Observations	13,304	10,122	13,304	14,189
Panel B: Placebo Analysis				
10-Point Scale	0.025	0.025	0.838	0.039
	(0.080)	(0.082)	(0.447)	(0.019)
97.07% CI	(-0.148, 0.199)	(-0.155, 0.204)	(-0.137, 1.812)	(-0.004, 0.081)
Observations	10,151	11,092	18,930	13,139

NOTES: Means to the left are calculated manually over a bandwidth range $\{-h^*, ..., -1\}$, where h^* is the MSE-optimal bandwidth. We report 97.07% confidence intervals to align with the honest critical value of 2.18 outlined in Armstrong and Kolesár (2020). The covariates used in panels C and D include indicators for gender, race/ethnicity, socioeconomic status, and gifted status in reading and math.

Our estimates for τ should be unbiased measures of τ^P if $\tau^{\text{age}} = 0$. While we cannot disentangle this possible effect from our desired effect in the analysis sample, we circumvent this identification issue by applying an identical analysis on a placebo sample of North Carolina students. Formally, for a placebo group of students, we estimate the effect of delayed kindergarten (subsequent high school entry) on our four outcomes of interest: GPA, test scores, number of days absent, and the likelihood of chronic absence. We minimize the potential for unobserved heterogeneity to bias our placebo estimates by choosing the closest two-year cohort to our actual cohort of interest. In practice, we repeat the estimation procedure for the sample of students in North Carolina born one year prior to our main analysis birth cohort. These students, if delayed, begin high school one year *before* the policy was introduced. The "control" group in this analysis begins high school two years before the introduction of the lenient grade standards. Given that both treatment and control groups in this placebo comparison are exposed the the same grading standards, any estimate $\tau \neq 0$ would suggest that $\tau^{\text{age}} \neq 0$.

Table 7 displays the results from this placebo analysis. In Panel A, we reproduce the main results for ease of comparability. Our results suggest that age effects do not bias our estimates. For each outcome, we find null results at the 5% level using the honest confidence interval method used for the main results of the paper (Kolesár and Rothe, 2018). Given that this comparison group enters high school one year prior to our analysis group, we do not expect any other possible effects attributable to our RD design to bias our estimates.

7.4 Changes in Education Production Inputs

Our final internal validity check explores the role of external time-varying factors that may enter at the school level and potentially bias our student-level estimates. We focus on one of the most important inputs to the production of education which meaningfully impact gains in student achievement and engagement: teacher quality (Chetty, Friedman and Rockoff, 2014; Jackson, 2018). A shock to teacher quality at the same time as the introduction of the policy would distort the estimated effects associated with academic leniency.

To obtain information on teacher characteristics, we link students in our analytical sample to Math I teachers using course membership files. NCERDC course membership files contain records of the teacher who taught a particular course and the students that took that course in any given school year which allows for a straightforward match of students to teachers.[27] Overall, we are able to match 96% of the students in our analytical sample to a teacher record from the course membership files.[28]

Table 8 presents regression discontinuity estimates of teacher characteristics, namely education and experience. Overall, we do not recover any meaningful difference in teacher characteristics for students above and below the birth date cutoff. Given that we find no evidence to suggest that the estimated effects of grade leniency recovered in Section 6 are driven by teacher quality, we conclude this section by extending the interpretation as one which rules out time varying changes in educational inputs as a general source of contamination in our estimation procedure. With these validity tests in mind, we take our results as capturing the true underlying effect of the lenient policy for the local sample of students born near the relevant cut date.

[27] Course membership files are available starting in the 2005-06 school year. Before 2006, researchers relied on the School Activity Report (SAR) files to link students to teachers. The SAR files contain information on school personnel that may be linked to student test score data via the information on the person monitoring the exam. One of the main limitations of this linking procedure is that it is not possible to establish whether the teacher listed in a student's test score record was actually that student's teacher for that test subject. Course membership records greatly improve student-to-teacher linking as they contain information on the teacher who taught each particular course. However, the only information identifying teachers in the course membership data is a single teacher name variable, unlike the SAR personnel data, which includes formatted first, last names, and social security numbers. As such, approximately 80% of teachers in course membership files can be matched to SAR data files. For more information see NCERDC Technical Report 1B and Technical Report #5.

[28] Some students (10%) take multiple Math I courses with more than 1 teacher in the same school year. This is mostly driven by students taking Math I in the Fall and Spring terms with different teachers. For these students, we use the characteristics of the Fall term teacher in our estimation.

Table 8: Testing for Differences in Educational Inputs

	(1) Bachelor's Degree	(2) Advanced Degree	(3) High Experience
10-Point Scale	-0.039	0.037	0.057
	(0.044)	(0.044)	(0.037)
97.07% CI	(-0.135, 0.056)	(-0.059, 0.133)	(-0.023, 0.136)
1st Stage Estimate	0.440	0.440	0.438
Mean (Left)	0.706	0.289	0.492
Bandwidth	38.58	37.94	62.70
Observations	10,507	10,298	17,272

NOTES: High experience denotes whether a teacher has more than 9 years of teaching experience (the median value of teaching experience in the data). Means to the left are calculated manually over a bandwidth range $\{-h^*, ..., -1\}$, where h^* is the MSE-optimal bandwidth. We report 97.07% confidence intervals to align with the honest critical value of 2.18 outlined in Armstrong and Kolesár (2020). Results come from regressions which include controls and district fixed effects. The covariates used include indicators for gender, race/ethnicity, socioeconomic status, and gifted status in reading and math.

8 Conclusion

The economic benefits of high school graduation are large and persistent, which has led local governments across the country to enact policy with the explicit goal of increasing graduation rates. One popular mechanism designed to boost the performance of low-achieving students is to relax stringent academic standards to help students meet graduation requirements and obtain a high school degree. As a result, the last decade has seen steady increases in graduation rates and grade point averages in US high schools without any measured increase in student achievement. At face value, a decline in academic standards can naturally lead to more equitable outcomes since a larger share of students meet graduation requirements and obtain a high school diploma. However, this simplified view ignores the possible endogenous role that academic standards have on student investment in school, which directly impacts their accumulation of human capital.

We show, theoretically and empirically, that the decision to lower standards may result in the unintended consequence of expanding gaps in academic achievement if student effort response is such that it offsets designed gains for low ability students. Under a fuzzy regression discontinuity research design that leverages variation from statewide grading policy and school entry rules, we find that increased academic leniency led to mechanical GPA gains among high ability students, but large effort reductions among low ability students. These heterogeneous effects compound over time and exacerbate gaps in student ACT performance. Therefore, the short-run gains of artificially raising high school completion rates may result in a permanent widening of long-run welfare gains, especially when lowered standards are not associated with any relative increase to human capital accumulation.

Our analysis is policy relevant for educators and policymakers seeking to change academic standards in their school districts. The findings of this paper highlight the importance of understanding how policy can shift incentives for effort of students, with the potential for heterogeneous effects across student populations. Our analysis shows how student response to incentives generated by shifts in education policy, not only can undo designed gains, but can exacerbate gaps if response is heterogeneous across student groups.

References

Ahn, Thomas, Peter Arcidiacono, Amy Hopson, and James R Thomas. 2019. "Equilibrium grade inflation with implications for female interest in stem majors." National Bureau of Economic Research.

Angrist, Joshua D, and Alan B Krueger. 1991. "Does compulsory school attendance affect schooling and earnings?" *The Quarterly Journal of Economics*, 106(4): 979–1014.

Armstrong, Timothy B, and Michal Kolesár. 2020. "Simple and honest confidence intervals in nonparametric regression." *Quantitative Economics*, 11(1): 1–39.

Babcock, Philip. 2010. "Real costs of nominal grade inflation? New evidence from student course evaluations." *Economic inquiry*, 48(4): 983–996.

Bar, Talia, Vrinda Kadiyali, and Asaf Zussman. 2009. "Grade information and grade inflation: The Cornell experiment." *Journal of Economic Perspectives*, 23(3): 93–108.

Bassok, Daphna, and Sean F Reardon. 2013. ""Academic redshirting" in kindergarten: Prevalence, patterns, and implications." *Educational Evaluation and Policy Analysis*, 35(3): 283–297.

Becker, William E, and Sherwin Rosen. 1992. "The learning effect of assessment and evaluation in high school." *Economics of Education Review*, 11(2): 107–118.

Betts, Julian R. 1998. "The impact of educational standards on the level and distribution of earnings." *The American Economic Review*, 88(1): 266–275.

Betts, Julian R, and Jeff Grogger. 2003. "The impact of grading standards on student achievement, educational attainment, and entry-level earnings." *Economics of Education Review*, 22(4): 343–352.

Blagg, Kristin, and Matthew Chingos. 2016. "Varsity Blues: Are High School Students Being Left Behind?"

Butcher, Kristin F, Patrick J McEwan, and Akila Weerapana. 2014. "The Effects of an Anti-Grade Inflation Policy at Wellesley College." *Journal of Economic Perspectives*, 28(3): 189–204.

Cattaneo, Matias D, Michael Jansson, and Xinwei Ma. 2018. "Manipulation testing based on density discontinuity." *The Stata Journal*, 18(1): 234–261.

Chetty, Raj, John N Friedman, and Jonah E Rockoff. 2014. "Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood." *American economic review*, 104(9): 2633–2679.

Cook, Philip J, and Songman Kang. 2016. "Birthdays, schooling, and crime: Regression-discontinuity analysis of school performance, delinquency, dropout, and crime initiation." *American Economic Journal: Applied Economics*, 8(1): 33–57.

Cullen, Julie Berry, and Randall Reback. 2006. "Tinkering Toward Accolades: School Gaming Under a Performance Accountability System." *Advances in Applied Microeconomics*, 14(1): 1–34.

Dee, Thomas S, and Hans Henrik Sievertsen. 2018. "The gift of time? School starting age and mental health." *Health economics*, 27(5): 781–802.

Dee, Thomas S, Will Dobbie, Brian A Jacob, and Jonah Rockoff. 2016. "The causes and consequences of test score manipulation: Evidence from the new york regents examinations." National Bureau of Economic Research.

Deming, David, and Susan Dynarski. 2008. "The lengthening of childhood." *Journal of economic perspectives*, 22(3): 71–92.

Diamond, Rebecca, and Petra Persson. 2016. "The long-term consequences of teacher discretion in grading of high-stakes tests." National Bureau of Economic Research.

Dobkin, Carlos, and Fernando Ferreira. 2010. "Do school entry laws affect educational attainment and labor market outcomes?" *Economics of education review*, 29(1): 40–54.

Durden, Garey C, and Larry V Ellis. 1995. "The effects of attendance on student learning in principles of economics." *The American Economic Review*, 85(2): 343–346.

Elder, Todd E. 2010. "The importance of relative standards in ADHD diagnoses: evidence based on exact birth dates." *Journal of health economics*, 29(5): 641–656.

Espenshade, Thomas J, Lauren E Hale, and Chang Y Chung. 2005. "The frog pond revisited: High school academic context, class rank, and elite college admission." *Sociology of Education*, 78(4): 269–293.

Fan, Jianqing, and Irene Gijbels. 1992. "Variable bandwidth and local linear regression smoothers." *The Annals of Statistics*, 2008–2036.

Figlio, David N. 2006. "Testing, Crime and Punishment." *Journal of Public Economics*, 90(4-5): 837–851.

Figlio, David N, and Maurice E Lucas. 2004. "Do high grading standards affect student performance?" *Journal of Public Economics*, 88(9-10): 1815–1834.

Glewwe, Paul, Nauman Ilias, and Michael Kremer. 2010. "Teacher Incentives." *American Economic Journal: Applied Economics*, 2(3): 205–27.

Goldsmith-Pinkham, Paul, Maxim Pinkovskiy, and Jacob Wallace. 2021. "The Great Equalizer: Medicare and the Geography of Consumer Financial Strain." *arXiv preprint arXiv:2102.02142*.

Hahn, Jinyong, Petra Todd, and Wilbert Van der Klaauw. 2001. "Identification and estimation of treatment effects with a regression-discontinuity design." *Econometrica*, 69(1): 201–209.

Hansen, Anne Toft, Ulrik Hvidman, and Hans Henrik Sievertsen. 2023. "Grades and employer learning." *Journal of Labor Economics*.

Harris, Douglas N, Lihan Liu, Nathan Barrett, and Ruoxi Li. 2023. "Is the rise in high school graduation rates real? High-stakes school accountability and strategic behavior." *Labour Economics*, 82: 102355.

Hastings, Justine S, Christopher A Neilson, and Seth D Zimmerman. 2012. "The effect of school choice on intrinsic motivation and academic outcomes." National Bureau of Economic Research Working Paper 18324.

Heckman, James J. 2006. "Skill formation and the economics of investing in disadvantaged children." *Science*, 312(5782): 1900–1902.

Humlum, Maria Knoth, Marius Morthorst, and Peter Thingholm. 2022. "Sibling Spillovers and the Choice to Get Vaccinated: Evidence from a Regression Discontinuity Design."

Hurwitz, Michael, and Jason Lee. 2018. "Grade inflation and the role of standardized testing." *Measuring success: Testing, grades, and the future of college admissions*, 64–93.

Hvidman, Ulrik, and Hans Henrik Sievertsen. 2021. "High-stakes grades and student behavior." *Journal of Human Resources*, 56(3): 821–849.

Jackson, C Kirabo. 2018. "What do test scores miss? The importance of teacher effects on non–test score outcomes." *Journal of Political Economy*, 126(5): 2072–2107.

Jacob, Brian A. 2005. "Accountability, Incentives and Behavior: The Impact of High-Stakes Testing in the Chicago Public Schools." *Journal of Public Economics*, 89(5-6): 761–796.

Kolesár, Michal, and Christoph Rothe. 2018. "Inference in regression discontinuity designs with a discrete running variable." *American Economic Review*, 108(8): 2277–2304.

Koretz, Daniel M. 2002. "Limitations in the Use of Achievement Tests as Measures of Educators' Productivity." *Journal of Human Resources*, 37(4): 752–777.

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

McCrary, Justin, and Heather Royer. 2011. "The effect of female education on fertility and infant health: evidence from school entry policies using exact date of birth." *American Economic Eeview*, 101(1): 158–195.

Metcalfe, Robert, Simon Burgess, and Steven Proud. 2019. "Students' Effort and Educational Achievement: Using the Timing of the World Cup to Vary the Value of Leisure." *Journal of Public Economics*, 172: 111–126.

Minaya, Veronica. 2020. "Do differential grading standards across fields matter for major choice? Evidence from a policy change in Florida." *Research in Higher Education*, 61(8): 943–965.

Murnane, Richard J. 2013. "US high school graduation rates: Patterns and explanations." *Journal of Economic Literature*, 51(2): 370–422.

Navarro-Palau, Patricia. 2017. "Effects of differentiated school vouchers: Evidence from a policy change and date of birth cutoffs." *Economics of Education Review*, 58: 86–107.

Nordin, Martin, Gawain Heckley, and Ulf Gerdtham. 2019. "The impact of grade inflation on higher education enrolment and earnings." *Economics of Education Review*, 73: 101936.

Ordine, Patrizia, Giuseppe Rose, and Daniela Sposato. 2018. "Parents Know Them Better: The Effect of Optional Early Entry on Pupils' Schooling Attainment." *Economic Inquiry*, 56(3): 1678–1705.

Persson, Petra, Xinyao Qiu, and Maya Rossin-Slater. 2021. "Family spillover effects of marginal diagnoses: The case of ADHD." National Bureau of Economic Research.

Stancanelli, Elena. 2017. "Couples' retirement under individual pension design: A regression discontinuity study for France." *Labour Economics*, 49: 14–26.

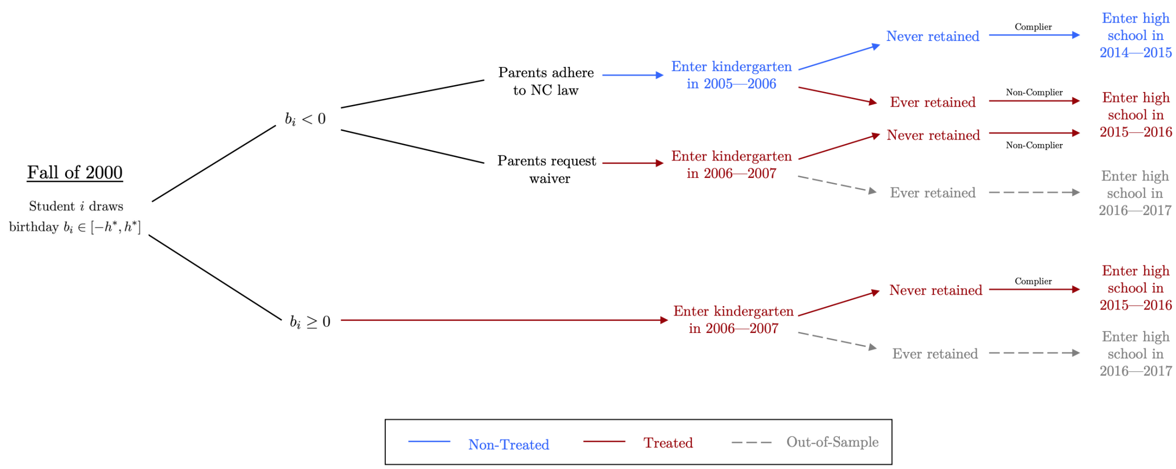
Stinebrickner, Ralph, and Todd R. Stinebrickner. 2008. "The Causal Effect of Studying on Academic Performance." *The B.E. Journal of Economic Analysis & Policy*, 8(1).

Todd, Petra E, and Kenneth I Wolpin. 2003. "On the specification and estimation of the production function for cognitive achievement." *The Economic Journal*, 113(485): F3–F33.

A Appendix

A.1 Timeline and Compliance to Kindergarten Entry Rules

Figure A.1: Pathway to Treatment Status for Analysis Group



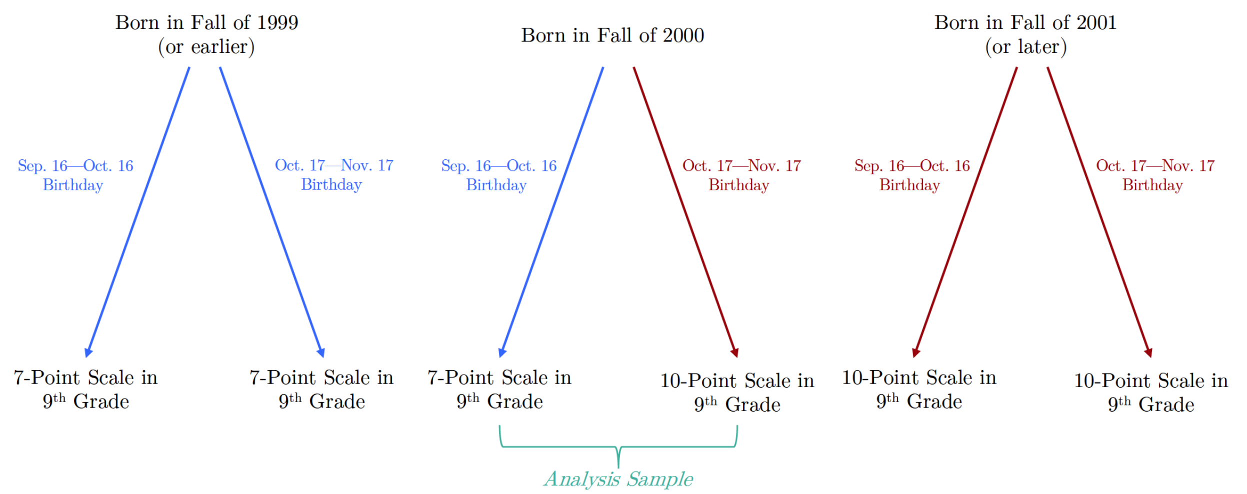
NOTES: In the above, h^* denotes the MSE-optimal bandwidth generated by the honest RD method outlined in Kolesár and Rothe (2018).

We present the timeline of some analysis sample student i's life cycle in public school in Figure A.1. We denote non-treated students (in blue) as those who enter high school in year 2015 and treated students (in red) as those instead entering in 2016. As evident by the figure, there are numerous opportunities for students to become non-compliers. In North Carolina, parents can request waivers for their child to start kindergarten one year after the law requires, given sufficient reasoning. Some districts also require that students perform sufficiently well on a kindergarten pre-test before granting the waiver.

However, students fully complying with the kindergarten entry rule may still become non-compliers in our analysis. Because treatment status refers to the year a student begins the 9th grade, any student retained between kindergarten and eighth grade incurs a shifted treatment status.

We find that 29.42% of students born at most six months before October 17th do not comply with the state's cutoff rule. Ideally, we would differentiate between those not complying because their parents chose to delay their entry and those held back in early grades. The former group self-selects into having an age advantage in their cohort while the latter is retained, presumably by their teachers.

Figure A.2: The Impact of Birthdays on Grading Scale Exposure Between Years



NOTES: Students born to the left of the cutoff in our analysis sample begin high school in the 2014-2015 school year; the treated students instead begin in the 2015-2016 school year. We provide a 30-day bandwidth as an illustrative example; in reality, the bandwidth depends on the outcome of interest but never exceeds two months.

A.2 Design

Under our regression discontinuity research design, treated students close to the birth date cutoff unknowingly begin high school under a more relaxed grading standard compared to those born just before, who instead face the usual 7-point scale. We argue that, even in the face of student retention in primary schooling, assignment to treatment status is quasi-random around the birth date cutoff because the policy became known *during* our treated sample's eighth grade year. Figure A.2 illustrates how we select our treatment (in red) and control (in blue) groups to created our analysis sample.

A.3 Regression Discontinuity Specifications

Table A.1: Results Under Various Changes to RD Specifications

	GPA in 9th Grade	Math I EOC Score (Standardized)	Number of Days Absent in 9th Grade	Prob. of Chronic Absence in 9th Grade
Panel A: Analysis Results (Reproduced)				
10-Point Scale	0.266	0.063	1.324	0.048
	(0.071)	(0.089)	(0.616)	(0.020)
97.07% CI	(0.111, 0.420)	(-0.130, 0.257)	(-0.019, 2.667)	(0.003, 0.093)
Panel B: 30-Day Bandwidth				
10-Point Scale	0.260	0.057	1.634	0.076
	(0.090)	(0.097)	(0.791)	(0.027)
97.07% CI	(0.063, 0.456)	(-0.154, 0.269)	(-0.090, 3.359)	(0.017, 0.135)
Panel C: Quadratic Fit of Running Variable				
10-Point Scale	0.253	0.007	1.697	0.085
	(0.108)	(0.136)	(0.946)	(0.031)
97.07% CI	(0.019, 0.488)	(-0.289, 0.302)	(-0.366, 3.759)	(0.017, 0.153)
Panel D: $\boldsymbol{K}$ Reduced by 20%				
10-Point Scale	0.269	0.065	1.274	0.042
	(0.069)	(0.084)	(0.587)	(0.019)
97.07% CI	(0.121, 0.417)	(-0.118, 0.247)	(-0.006, 2.554)	(0.000, 0.084)
Panel E: $\boldsymbol{K}$ Increased by 20%				
10-Point Scale	0.264	0.064	1.339	0.049
	(0.074)	(0.092)	(0.624)	(0.021)
97.07% CI	(0.104, 0.425)	(-0.136, 0.264)	(-0.022, 2.699)	(0.004, 0.094)

NOTES: We report 97.07% confidence intervals to align with the honest critical value of 2.18 outlined in Armstrong and Kolesár (2020). Panel A reproduces the main results. Panel B fixes all outcomes to a 30-day bandwidth. Panel C considers $g(b_i)$ as a quadratic as opposed to a linear fit. Panels D and E respectively decrease and increase the values of (K_1, K_2) in the first stage estimation procedure for Kolesár and Rothe (2018).

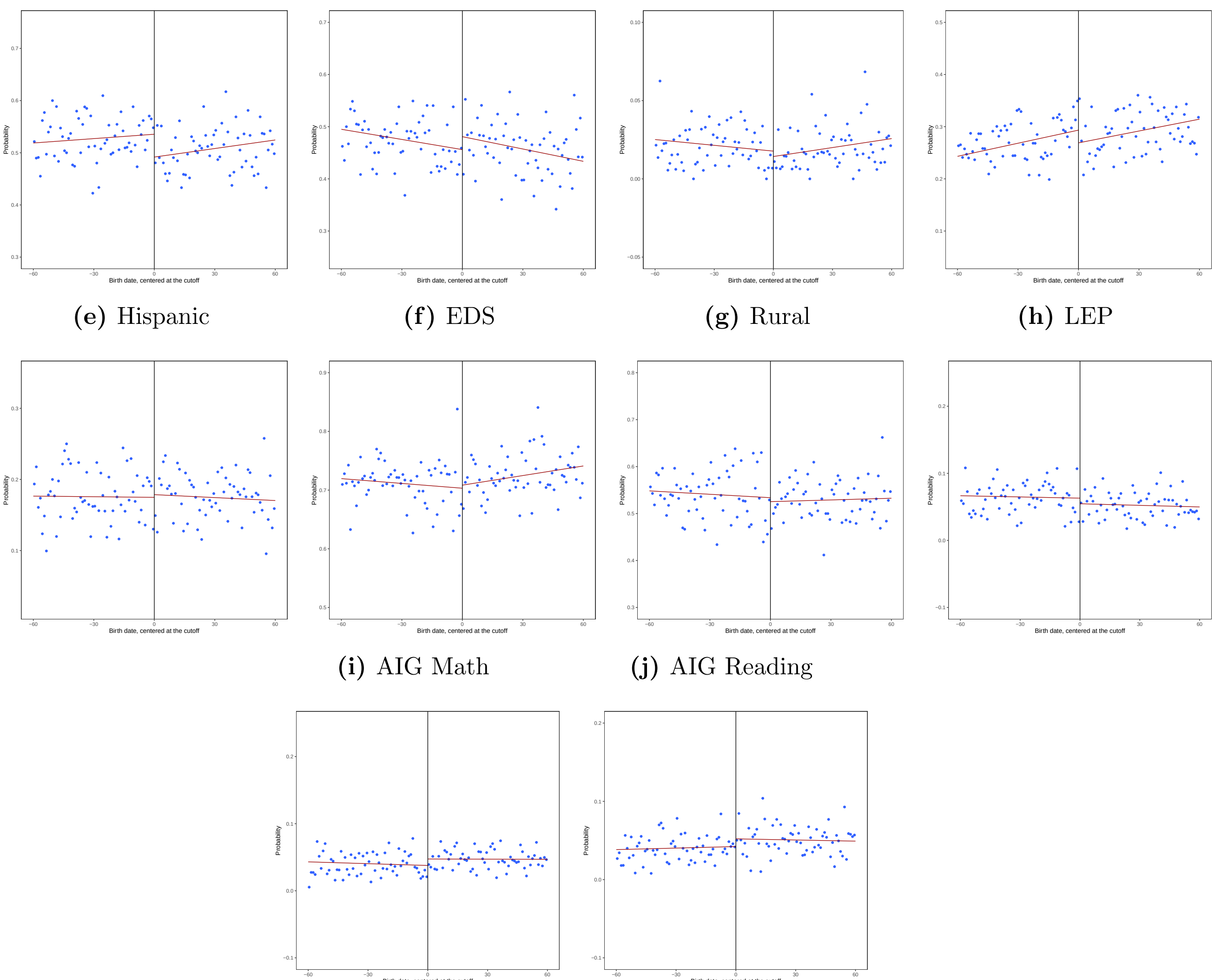
Table A.2: Student Outcomes Under Additional Specifications

	GPA in 9th Grade	Math I EOC Score (Standardized)	Number of Days Absent	Probability of Chronic Absence
Panel A: Kolesár & Rothe				
10-Point Scale	0.253	0.171	1.232	0.040
	(0.074)	(0.092)	(0.614)	(0.020)
95% Honest CI	(0.091, 0.414)	(-0.031, 0.373)	(-0.095, 2.558)	(-0.003, 0.084)
1^{st} Stage Estimate	0.439	0.437	0.439	0.440
Mean (Left)	2.436	0.108	6.479	0.059
Bandwidth	46.27	35.65	46.24	49.96
Panel B: 2SLS Without Covariates				
10-Point Scale	0.253	0.173	1.233	0.041
	(0.074)	(0.093)	(0.616)	(0.020)
97.07% CI	(0.091, 0.415)	(-0.029, 0.375)	(-0.110, 2.575)	(-0.003, 0.085)
1^{st} Stage Estimate	0.439	0.437	0.439	0.440
Observations	13,901	10,593	13,901	14,816
Panel C: 2SLS With Covariates				
10-Point Scale	0.263	0.095	1.189	0.042
	(0.072)	(0.092)	(0.626)	(0.021)
97.07% CI	(0.107, 0.419)	(-0.104, 0.295)	(-0.178, 2.551)	(-0.003, 0.089)

NOTES: Means to the left are calculated manually over a bandwidth range $\{-h^*, ..., -1\}$, where h^* is the MSE-optimal bandwidth. We report 97.07% confidence intervals to align with the honest critical value of 2.18 outlined in Armstrong and Kolesár (2020). The covariates used in panels C and D include indicators for gender, race/ethnicity, socioeconomic status, and gifted status in reading and math.

A.4 Covariate Smoothness Plots

Figure A.3: Smoothness of Student Pre-Treatment Covariates



(a) Female **(b)** White **(c)** Asian **(d)** Black

(e) Hispanic **(f)** EDS **(g)** Rural **(h)** LEP

(i) AIG Math **(j)** AIG Reading

Notes: This figure shows discontinuities in student pre-treatment covariates. All panels are created by plotting the average outcomes across each birth date fitting the data using a linear regression without controls on either side of the cutoff.

A.5 Additional Tables

Table A.3: Student Sorting Across Years

	2014	2015	2016	2017	2018
Born after Oct. 16	-0.395	-0.411	-0.398	-0.421	-0.461
	(0.015)	(0.014)	(0.015)	(0.015)	(0.013)
95% Honest CI	(-0.427, -0.363)	(-0.441, -0.381)	(-0.432, -0.365)	(-0.453, -0.389)	(-0.489, -0.432)
Mean (Left)	0.545	0.568	0.541	0.575	0.594
Bandwidth	30.49	33.36	27.08	30.65	35.17

Notes: Each effect in the first row comes from estimates produced by the procedure for a sharp regression discontinuity outlined in Kolesár and Rothe (2018) using a data-driven approach to calculate the bounds on the second derivative.

Table A.4: Effects Across Prior Academic Preparedness and Demographics

	(1) Core Academic Grade Point Average	(2) Math I EOC Score (Standardized)	(3) Total Number of Days Absent	(4) Probability of Chronic Absence
Panel A: Gender				
Female	0.268	-0.001	0.947	0.054
	(0.089)	(0.109)	(0.799)	(0.026)
Male	0.244	0.064	1.550	0.037
	(0.115)	(0.143)	(0.952)	(0.033)
Panel B: Race				
White	0.337	0.040	1.696	0.063
	(0.112)	(0.147)	(0.992)	(0.033)
Non-White	0.202	0.049	1.078	0.036
	(0.092)	(0.110)	(0.795)	(0.026)
Panel C: EDS Status				
EDS	0.257	0.025	1.458	0.059
	(0.084)	(0.104)	(0.759)	(0.026)
Non-EDS	0.263	0.091	0.773	0.010
	(0.132)	(0.176)	(1.033)	(0.029)

Notes: We fix the bandwidth for each specification to be the same as in the full sample version. We present results from regressions which include covariates and district fixed effects. Heteroskedasticity-robust standard errors are reported in parentheses beneath each point estimate.

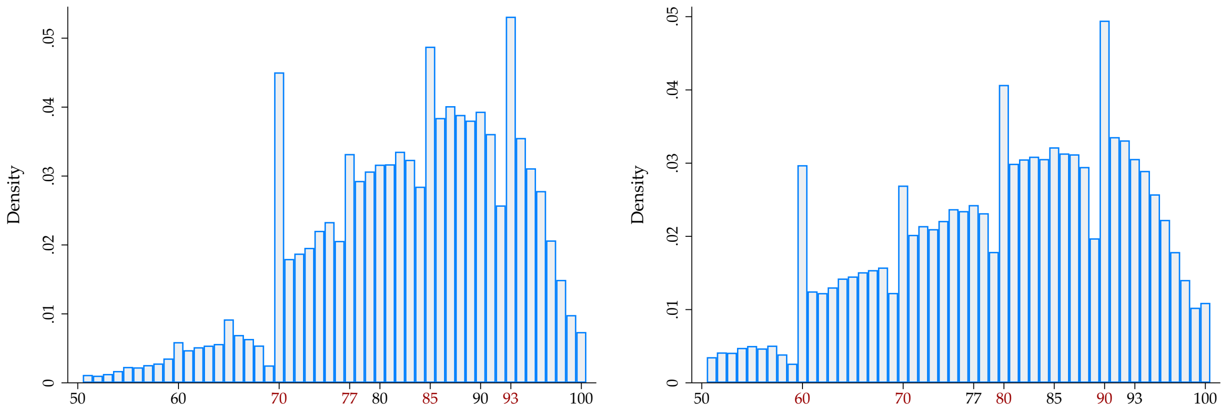
Table A.5: Effects on Longer-Run Outcomes

	Unweighted GPA	Days Absent	Chronic Absence
Panel A: Year 2			
Full Sample	0.131	1.114	0.026
	(0.055)	(0.475)	(0.016)
Low Ability	0.115	1.801	0.058
	(0.075)	(0.761)	(0.026)
High Ability	-0.001	1.039	0.003
	(0.062)	(0.524)	(0.017)
Panel B: Year 3			
Full Sample	-0.000	3.598	0.082
	(0.048)	(0.651)	(0.020)
Low Ability	-0.077	5.949	0.162
	(0.067)	(0.293)	(0.033)
High Ability	-0.098	2.143	0.040
	(0.056)	(0.688)	(0.023)
Panel C: Year 4			
Full Sample	0.023	2.334	0.053
	(0.050)	(0.912)	(0.023)
Low Ability	-0.097	3.444	0.104
	(0.072)	(1.481)	(0.038)
High Ability	-0.006	1.223	0.019
	(0.064)	(1.011)	(0.028)

NOTES: We present robust standard errors in parentheses. Each panel refers to a different year in high school. We abstain from writing, e.g., "10th Grade" because some students may not matriculate. In an effort to maintain the same sample size across years, we do not condition on grade level.

A.6 Additional Figures

Figure A.4: Additional Distributions of Final Course Averages in 9th Grade Math



(a) 2014 **(b)** 2017

NOTES: As before, red labels denote cut grades. We fix both distributions to display only final grades at or above 50, accounting for the vast majority of transcript grades.