Growth-based school accountability rules and grade retention practices

(Latest version can be found here)

William Labadie

April 21, 2021

Abstract

Do accountability rules affect public school retention practices? Using a simple model of grade retention, I show that an administrator will retain students differently depending on the accountability ratings criteria he seeks to maximize. I show that an administrator whose school is rated based on student standardized exam passing rates has a strategic incentive to retain borderline students, while an administrator whose school is rated based on year-to-year growth in student standardized exam scores has incentive to retain only the lowest-scoring students. I further show that this effect is most pronounced in the final grade offered by a given school, when promotion of a student ensures her removal from the school's pool of test-takers . I test the predictions of my framework using a novel dataset of school-grade level retention rates for 7 states in the U.S. and an event study design. I find that about 18% fewer students are retained on average each year when a state adds student growth to the accountability criteria by which schools are evaluated. This number roughly corresponds to around 100,497 fewer retained students each year nationwide, and $1.4 billion saved in public school expenditures. I further find that administrators do retain significantly fewer students in the last grade offered by their schools, implying that administrators use retention strategically. This paper is the first to show evidence that school administrators are willing to use retention as a tool for optimizing their schools' accountability ratings, and demonstrates that the individual components of accountability systems alter administrator behaviors.

1 Introduction

Each year, public school administrators endeavor to guide their schools to satisfactory ratings as defined by their state's adopted accountability system. It is well-established that administrators value maximizing their schools' ratings, and that they respond to the specific criteria by which their schools are rated (Rockoff and Turner, 2010; Chakrabarti, 2013). Accountability systems do not include targeted goals around student retention, but an administrator's retention practices could have significant impacts on her school's overall rating. To the extent that administrators decide whether or not to retain a student with an eye towards optimizing their schools' ratings, different ratings criteria could lead to different retention decisions.

In this study, I evaluate the impact of a common accountability measure, year-to-year growth in student standardized test scores, on public school retention policy. Retention has been shown to have profound effects on students in the short and long term, but the school-side determinants of retention are not well-understood (Schwerdt, West and Winters, 2017; Eren, Lovenheim and Mocan, 2018). This paper sheds light on how administrator incentives could affect the retention decision.

I find that in the presence of growth-based accountability criteria, fewer public school students are held back, and that the effect grows over time. After 6 or more years under an accountability system that includes within-student score change, retention rates decrease by 18.3%. A back-of-the-envelope calculation using average enrollment counts from the Common Core of Data suggests that these estimates correspond to 100,497 fewer retained students each year nationwide. The reduction in years in the school system results in an average of \$1.4 billion less in student expenditures.[1] Back-of-the-envelope calculations based on other researchers' estimates of the effects of retention suggest that the reduction in retention also potentially avoids \$100-400 million in lost wages and \$91 million in crime-related costs incurred by retention of students, though it does potentially increase expenses associated with providing additional remedial math and reading courses by \$518 million.

I find that the effect of growth-based accountability criteria on retention is more pronounced in the final grade offered by a school. This result is consistent with the predictions of my statistical framework, and holds when the sample is restricted to elementary, middle, and high schools separately. This result suggests strategic retention on the part of school administrators, as the decision to retain or promote in the terminal grade either includes a student in the test-taking pool for an additional year, or removes him from the pool. That the differential effect of growth-based accountability criteria is negative suggests that administrators view retention as potentially harmful to year-over-year student score growth.

School administrators' incentives change with the accountability rules their schools are subject to. Through retention, an administrator has a degree of control over the difficulty of test faced by a given student. If the administrator expects that a grade-4 student is unlikely to pass the grade 5 exam, but likely to pass the grade 4 exam, retaining that student could have a non-negligible effect on his school's accountability rating. Under an accountability system based only on school passing rate, strategic retention of this sort could have an especially clear effect on the school's accountability score. However, under a system that rates schools on both passing rate and within-student score growth, the effect could be less clear. My statistical framework shows that different students stand to benefit the school's passing rate and sufficient score

[1] Based on current per-student expenditures of \$13,847 (National Center of Education Statistics, 2019).

growth rate upon retention. The higher the student's score, the more likely they are to pass the exam if retained. However, students with higher latent ability are less likely to pass a sufficient score growth standard if retained than those with lower latent ability. The change caused by the adoption of a growth-based accountability system is expected to be especially pronounced in the final grade offered by a school. Students that would increase the school's passing rate if retained are likely to have relatively high latent ability, and be less likely to pass the sufficient growth standard as a result. In the final grade offered by the school, these students can be removed from the pool of test-takers if promoted. In earlier grades, such students are likely to fail the sufficient growth standard whether promoted or retained, and will negatively affect the school's sufficient score growth rate equally.

On the whole, my results suggest that fewer students are retained under growth-based systems. To the extent that passing rate-based systems encouraged retention of students that otherwise would not have been retained for maximizing passing rates, this is a good thing. Prior research on retention has found substantial negative effects in the long term (Eren, Lovenheim and Mocan, 2018; Brodaty, Gary-Bobo and Prieto, 2013). This is especially true for older retained students (Jacob and Lefgren, 2009). I find some evidence of higher retention among younger students, however. It is possible that growth-based accountability systems encourage more targeted retention, since administrators would be harmed by retaining students that would be developmentally harmed by retention. This also would be a positive result, since the weight of evidence seems to suggest that retention is most beneficial when used as a highly individualized intervention (Jacob and Lefgren, 2004; Fruehwirth, Navarro and Takahashi, 2016).

To study the effects of growth-based accountability systems on retention practices, I develop a statistical framework in which students' test scores progress over time and administrators set score thresholds for each grade; students that score below the relevant score threshold are retained, while those that score above the threshold are promoted. This framework predicts different score threshold patterns for maximizing school-wide passing rate and school-wide score change, consistently predicting lower retention in the final grade offered by a school under a wide set of parameters. Because schools must satisfy both passing rate and score change-based criteria, the value of passing one is lower than it would be under a passing rate-only regime. The uncertainty of student ability translating to scores combined with the lower value of passing one criteria suppress retention rates in the final grade.

To estimate the impact of growth-based accountability criteria on retention, I exploit the adoption of within-student score growth into seven states' accountability criteria over time. I use a difference-in-differences strategy and a novel dataset assembled with the assistance of the states' education agencies to evaluate the effect of these accountability system changes. While the timing of states adopting the score change criteria is not random, it is exogenous to any school-level decision-making. Following Goodman-Bacon (2018), I employ an event-study design in my analysis. I study the differential effect in the final grade offered by a school with a triple-differences technique.

In this study, I show that administrators retain students differently depending on the accountability criteria they face. Almond, Lee and Schwartz (2016) show that public school administrators do exercise influence over the retention decision, though this study is the first to show that they do so strategically. A large literature exploring the unintended consequences of accountability systems has established already that administrators value satisfactory ratings (Rockoff and Turner, 2010). In this paper, I establish retention as a tool that administrators are willing to use in pursuit of better ratings for their schools. This paper contributes to a large literature exploring the tools administrators use to maximize their schools' ratings, including highly individualized tools (Figlio and Winicki, 2005; Reback, 2008). I show that administrators alter the body of students that contribute to their schools' ratings. Cullen and Reback (2006) find evidence of schools altering their test-taking pool by strategically exempting certain students from taking state exams to improve their ratings.

The rest of the paper is structured as follows. Section 2 will elaborate on school accountability and the incentives of administrators. Section 3 lays out my statistical framework. Section 4 discusses the data I use to test the predictions of the framework. Section 5 details the empirical strategies I use and the results of the analyses. Section 6 evaluates the costs and benefits of the change in retention practices under score change-based accountability based on estimates from the literature on the effects of retention. Section 7 concludes.

2 Background

Since the enactment of the No Child Left Behind (NCLB) Act of 2001, states have been required to evaluate public schools based on student performance on standardized math and English Language Arts (ELA) exams. Under NCLB, state education agencies were required to evaluate schools based on the rate at which students passed its standardized exam, and on the rate at which various subpopulations of students at the school passed the exam. However, NCLB allowed state education agencies a substantial degree of flexibility in designing their accountability systems; each agency could choose its own standardized exam to administer, and define what level of performance would be considered adequate, and accountability systems could and did evaluate schools on criteria beyond the minimum competency requirement. A common criteria, especially among states that had accountability systems in place prior to the enactment of NCLB, was year-to-year growth in individual students' scores.

The Every Student Succeeds Act (ESSA) of 2015 requires state education agencies to develop accountability systems that include multiple evaluation criteria, which must include "students' performance on the statewide assessment, high school graduation rates, and English language proficiency" as well as at least one additional measure of school quality that is left up to the individual agencies (Alliance for Excellent Education, 2016).

Figure 1 shows the percent of states whose accountability systems included a measure of student test score growth from the 2004-05 to 2017-18 school years. In the 2004-05 school year, 9 states evaluated schools partially based on the growth rate of students; by the 2017-18 school year, every state did. While states did not randomly choose whether or not to adopt growth into their accountability systems, the decision was exogenous to school-level decision-making, and affected the way in which schools were evaluated.

Clearly, a school administrator's incentives are affected when the measures by which his school is evaluated change. Once student score growth is included in a school's rating, the administrator must attempt to satisfy the state-defined criterion along with the other indicators the state education agency considers in order to maximize his school's overall rating. A deep body of research has shown that authorities at schools use a number of tools to increase their school's overall rating. Reback (2008) finds that Texas students whose scores are relatively important to their school's overall rating perform better than expected, and finds evidence of finely targeted resource allocation and instruction to these "high-leverage" students; Figlio and Winicki (2005) find that schools in Virginia at risk of accountability-based sanctions increase the caloric content of school lunches on test days; Craig, Imberman and Perdue (2013) find that district administrators increase instructional budgets after facing a negative rating shock. While Cullen and Reback (2006) find that Texas schools manipulate the pool of test-takers via exemptions, no study has examined whether or not administrators hold students back strategically as well.

Under a passing rate-based accountability system, a ratings-motivated administrator might hold a student back for a number of reasons. If the student is in grade g and the administrator expects the student to pass the grade g exam in the coming school year, but not the grade $g+1$ exam, he might retain the student to bolster the school's passing rate. This is particularly true if the student is from one of the subpopulations whose passing rates are explicitly valued under NCLB. The administrator might also retain the student if he believes that retention will positively impact the student's development, and have a lasting and positive impact on the student's future scores. Some research has shown that retention has positive short-run effects on student scores, but negative long-run effects on student development (Schwerdt, West and Winters, 2017); if a school is very likely to fail and its administrator estimates that the short-run effect of the student passing after being retained exceeds the later drag the student might have on the school's passing rate. On the other hand, if the administrator estimates that the long-run drag exceeds the short-run benefit, he would be more likely to promote the student.

Under an accountability system that includes year-to-year growth, the incentives to retain change. On one hand, an administrator might be more likely to retain a student if he expects the short-run benefits to be hight enough; it would improve his school's likelihood of passing both the state's passing rate and score growth standards. On the other hand, if retention has persistent negative effects, holding a student back will make it less likely that his school passes the state's sufficient growth standard.

3 Conceptual framework

Because the overall effect of retaining a student is unclear, I build and simulate a basic one-school model of grade retention. The model relies on a factor model formulation of student skill accumulation and skill measurement by standardized exams, following the basic structure of Cunha and Heckman (2008). In each period t, students at the school earn a score on the state's standardized exam. Assume that there is only one subject exam, and only one score counts towards the school's rating. Assume that student scores are represented by the following dynamic factor structure:

$$s_{it} = \mu + \alpha(g_{it})\theta_{it} + \epsilon_{it} \tag{1}$$

where i represents the individual student, and θ_{it} is a dynamic factor for each student. $\alpha(g_{it})$ is the factor loading parameter, which translates latent ability θ_{it} into a measurable score, s_{it}. It is a function of the student's grade, g_{it}, reflecting the fact that students in different grades take tests of different difficulty levels; $\frac{\partial\alpha(g_{it})}{\partial g_{it}} < 0$, such that students with the same latent ability values will earn lower scores in expectation if they are placed in higher grades. ϵ_{it} represents a random shock to the measured scores, representing the noisiness of test scores in measuring ability.

I assume the following form for students' skills production:

$$\theta_{it} = \gamma_0 + \gamma_1\theta_{it-1} + \eta_{it} \tag{2}$$

where η_{it} is assumed independent across students and over time for the same students. I assume that students draw some initial value θ_{i0} prior to entering school. For simplicity, I assume that retention does not affect the skills production function. Research suggests that it would be more realistic to include retention in the skills production function, and this is a consideration for future work.

The administrator's objective is to allocate students across grades to maximize either the percent of students passing the state exam (passing rate) or both the passing rate and the percent of students exhibiting sufficient test score growth from one year to the next (sufficient growth rate). The school's passing rate in year t is given by

$$\Pi_t = \frac{\sum_{j=1}^{G}\sum_{i=1}^{N} \mathbb{1}(g_{it} = j) \times \mathbb{1}(s_{it} \geq \pi)}{N_t} \tag{3}$$

where G represents the final grade offered by the school and N represents the total number of students. π represents the externally set passing exam score. N_t represents the number of students whose scores count towards the school's rating in year t:

$$N_t = \sum_{j=1}^{G} \sum_{i=1}^{N} \mathbb{1}(g_{it} = j). \tag{4}$$

The school's sufficient growth rate is assumed to be measured as follows:

$$\Lambda_t = \frac{\sum_{j=1}^{G} \sum_{i=1}^{N} \mathbb{1}(g_{it} = j) \times \mathbb{1}(s_{it} - s_{it-1} \geq \lambda)}{N_t} \tag{5}$$

where λ represents the externally set target amount of score growth for a given student, and all other objects are as defined above. In reality, accountability measures of student growth vary across states; however, the percent of students demonstrating a sufficient level of score growth is a central component of all the growth-based accountability measures I analyze in this paper.

The school administrator is able to choose the difficulty of the exam that a student is exposed to in a given year through retention. The administrator can choose to retain a grade g student, and she will take the grade g exam rather than the more difficult grade $g+1$ exam. I assume that administrators make the retention decision by setting promotion thresholds δ_g for each grade. Grade g students that score at or above δ_g are promoted to grade $g+1$, while grade g students that score below δ_g are retained and repeat grade g. That is, administrators control students' grade level g based on the following:

$$g_{it} = g_{it-1} + \mathbb{1}(s_{it-1} \geq \delta_g). \tag{6}$$

I assume that retention is the only tool administrators have to influence their school's scores in this model; they can only affect Π_t and Λ_t through their δ_g choices. δ_g affects student scores, s_{it}, directly through $\alpha(g_{it}(\delta_g))$, but it also affects the population of grades g and $g+1$ through N_t, since g_{it} is a function of δ_g.

I assume that the school administrator has two important constraints. First, he must promote students that pass the standardized exam:

$$\delta_g \leq \pi_g \tag{7}$$

and he cannot retain a student in the same grade two years in a row

$$g_{it} \neq g_{it-2} \forall i. \tag{8}$$

These constraints are based on common rules adopted by school districts and state education agencies. The first has the additional feature of removing the option of retaining all passing students to maximize the school passing rate, which is appealing for the simulation of this model. I assume that these constraints hold for administrators regardless of the accountability system their schools are subject to.

An administrator of a school under a passing rate-based accountability system has the following optimization problem:

$$\max_{\delta_1,\delta_2,\delta_3} \Pi(\delta_1,\delta_2,\delta_3) \quad s.t. \quad \delta_g \leq \pi_g, \quad g_{it} \neq g_{it-2} \forall i. \tag{9}$$

while an administrator of a school under a passing rate and growth-based accountability system has the following optimization problem:

$$\max_{\delta_1,\delta_2,\delta_3} \Pi(\delta_1,\delta_2,\delta_3) + \Lambda(\delta_1,\delta_2,\delta_3) \quad s.t. \quad \delta_g \leq \pi_g, \quad g_{it} \neq g_{it-2} \forall i. \tag{10}$$

The mechanisms driving the administrator's decision are shown graphically in Figure 2. Figure 2a plots the relationship between a student's grade level, his latent ability, and his test score. Because the α function is decreasing in student grade level, the same level of latent ability will translate into lower scores for students in higher grades in expectation. A given student i with latent ability θ_{it} in year t earns a test score equal to $s_{it}(\theta_{it}, \alpha(g_{it} = j))$ if he is in grade j in year t. Figure 2a shows a student whose latent ability grows from θ_{it} in year t to θ_{it+1} in year $t+1$, and the scores he would earn in year $t+1$ if he is retained or promoted. $\bar{s}$ represents the maximum achievable score. If retained, the student will earn a score of $s_{it+1}(\theta_{it+1}, \alpha(g_{it+1} = j))$; if promoted, his score will be substantially lower, at $s_{it+1}(\theta_{it+1}, \alpha(g_{it+1} = j+1))$. If π, the externally-set passing standard, is between $s_{it}(\theta_{it}, \alpha(g_{it} = j))$ and $s_{it+1}(\theta_{it+1}, \alpha(g_{it+1} = j+1))$, the administrator's objective function is expected to be as well-served by promoting the student as by retaining him. In this case, an optimal δ_1 will be set below $s_{it}(\theta_{it}, \alpha(g_{it} = j))$. If π is between $s_{it+1}(\theta_{it+1}, \alpha(g_{it+1} = j+1))$ and $s_{it+1}(\theta_{it+1}, \alpha(g_{it+1} = j))$, however, the administrator would be best-served by retaining the student. If promoted to grade $j+1$, the student would not be expected to pass the grade $j+1$ exam; if retained, the student would be expected to pass the grade j exam.

Figure 2b shows the relationship between a student's grade level, his latent ability, and the rate of change in his scores. For the same values θ_{it} and θ_{it+1} plotted in Figure 2a, Figure 2b plots the partial derivative of s with respect to θ. The shape is determined by the functional form of α, since α is a function of grade level, which is determined by score - a function of θ. If λ, the externally-set sufficient growth standard, is below $\frac{\partial s_{it}(\theta_{it}, \alpha(g_{it}=j+1))}{\partial \theta}$, then the student's score is expected to grow enough to pass the standard if he is promoted from grade 1 to grade 2. If λ is above $\frac{\partial s_{it}(\theta_{it}, \alpha(g_{it}=j+1))}{\partial \theta}$, the student's score growth is expected to satisfy the standard if he is retained, but not if he is promoted to grade $j+1$. To maximize the sufficient score growth standard only, administrators should set δ_j such that students with $\frac{\partial s_{it}(\theta_{it}, \alpha(g_{it}=j+1))}{\partial \theta} < \lambda$ are retained; this occurs for students with relatively low θ_{it}. As a result, students with relatively low scores are those more likely to satisfy the sufficient growth standard when retained.

To gain some insight into the retention decision under different accountability systems, I perform a Monte Carlo exercise, in which I perform 10,000 simulations of the model. Each iteration simulates one school with three grades for ten years, with 100 new students entering the first grade each year. The parameter assumptions I employ are given in Table 1. I define the factor loading $\alpha_{it} = \frac{1}{\sqrt{g_{it}}}$, which is the functional form used in Figure 2. I assume that measurement is relatively noisy while the noise in skill accumulation is relatively small: $\epsilon_{it} \sim$ N(0,20). and $\eta_{it} \sim$ N(0,10). I also assume $\theta_{i0} \sim$ N(50,25), and that raw ability grows by 35 points per year in expectation, and thus assign $\gamma_0 = 35$. This means that the average student entering the school will score a 50 on the first grade exam in expectation, and if promoted each year, she will score a 60 on the second grade exam and a 69 on the third grade exam. Finally, I assume that the state passing standard, π, is equal to 50, such that the average student would be expected to pass the exam each year without needing to be retained. I assume that the state sufficient growth standard, λ, is equal to 10, about the rate at which the average student's scores are expected to naturally increase in the absence of retention. An administrator of a school with students scores centered substantially below or above the passing standard, or one whose student population's scores are expected to grow substantially slower or faster than the sufficient growth standard, could behave differently than the one I simulate.

The simulation shows that an administrator will use different retention strategies to maximize schoolwide passing rate than he will to maximize both passing rate and sufficient score growth rate. Figure 3 summarizes the results of the simulation. To maximize both passing and sufficient score growth rate, an administrator should set a higher δ_1 on average than he would to maximize passing rate only, and a much lower δ_3. δ_2 should be set similarly as under a passing rate-based system. The optimal δ_1, δ_2, and δ_3 values result in 2.925 percentage-point higher retention rates in the first grade on average (a 37.5% increase), 0.201 percentage-point lower retention rates in grade 2 (a 3% decrease), and 6.485 percentage-point lower retention rates in the third grade on average (a 77% decrease). These results are not surprising given the mechanisms shown in Figure 2. Students in grade 1 are expected to satisfy the sufficient growth standard whether promoted or retained I focus on these predictions of the model - that adoption of growth-based accountability leads to lower overall retention rates, lower retention rates in the final grade offered by a school, and higher retention rates in early grades offered by a school - in my empirical analysis.

4 Data

I use a novel source of administrative data on retention rates at the school-grade level from the 2011-12 to the 2017-18 school year. I obtained this data from the education agencies of 7 states. It includes retention rates and school identifying information for every school in each state over this 7-year span. There are numerous types of schools in each state's public school system, which typically are either exempt from accountability rules, or the rules apply differently to them than they do to typical public schools. Some of the most common of these are charter schools, magnet schools, career and technical education (CTE) schools, and disciplinary alternative education schools. I omit these schools from my analysis, and focus on standard public schools, where accountability rules are likely to be most salient.

I combine this data with data on the components of states' accountability systems over the same period. I constructed this data set primarily using publicly available state statutes and administrative code regulations. I used data on state accountability system components from Education Week's Education Counts Research Center from the 2001-02 to 2011-12 school year to establish a baseline. This data was collected by Education Week via survey; state education agencies self-reported the components in their accountability systems each year, with documentation to support their responses. Their dataset includes indicators for whether or not a state "assigns ratings to all schools based on state-developed criteria", "uses measures of individual student growth to rate schools", and "uses measures of individual student growth for state ratings". The data I collected from the state statutes and agencies fills in the same indicators for 2012-13 through 2017-18. Every state in my data had a criteria evaluating schools based on average student performance on standardized tests prior to adding a measure for individual student growth. Table 2 shows the timing of adoption of growth-based accountability for each state in my sample.

Different states define measures that rate schools on individual student growth differently. Most states that include a measure of individual student growth in school ratings measure the percent of students in each school whose test score increased by a sufficient amount from one year to the next (as defined by the state agency); some states define growth as the difference between a student's score and a comparable group of students' scores from the previous year, to attempt to average out idiosyncracies in test performance. However, all definitions emphasize year-to-year score increases.

I also rely on the Common Core of Data (CCD) in conjunction with this data for enrollment counts for various subpopulations in the school in each grade, the grades offered by each school, Title I status, charter status, and magnet status. The CCD variable defining a school's grade span is crucial to my analyses on differential behavior by schools towards students in terminal grades. I use this data to construct a terminal grade indicator, which equals one if a given grade is the last offered by the school. Overall, my data covers around 12,400 public schools across 7 years in which states gradually adopt measures based on individual student score growth in their accountability ratings systems.

I use this cross-state panel to analyze the effect that growth-based accountability systems have on public school retention practices. By using a cross-state panel, I am able to exploit the staggered adoption of growth-based accountability across states, comparing the difference in retention practices in states that adopt growth-based accountability to the difference in those that don't, across the time period spanned by my data.

Table 3 displays the mean and standard deviations of some variables of interest in my data for schools subject to growth-based accountability rules and for those not subject to growth-based accountability rules. The average grade in the sample is around 4, and the terminal grade in the sample is near 6, reflecting the fact that my sample is restricted to elementary and middle schools. In general, schools under both systems look similar, and there is no clear reason to think that the schools are not comparable. In schools that are subject to growth-based accountability rules, the shares of enrolled students that are white and Black are higher than in schools that are not subject to growth-based accountability rules, and the share of enrolled students that are Hispanic is lower. These small differences are not threats to identification, however, as my empirical strategy focuses on within-school differences.

5 Empirical analysis

5.1 Measuring the Impact of Growth-based Accountability Systems on Retention Practices

I exploit the gradual adoption of growth-based accountability systems over time across states to study its effect on retention practices. The empirical strategy I use in this paper builds on a large literature in applied economics, which uses variation in timing of treatment to estimate a difference-in-differences effect.

My empirical strategy compares the difference in retention rates in states that adopt growth-based accountability systems before and after the switch to the difference for those that don't over the same time period. Since all states had adopted some form of growth-based accountability by the 2017-18 school year, a basic difference-in-difference design, in which some states are part of the "treatment" group and some are part of the "control" group, is not possible. Instead, I use an event-study design, which allows for analysis even with no untreated units (Abraham and Sun, 2018). In addition, the event-study design allows for varying treatment effects over time, while the standard difference-in-differences framework is biased in the presence of treatment effects that vary over time.

My empirical approach relies on the exogeneity of the adoption of growth-based accountability systems. While the timing of a state's adoption of a new accountability system and the content of that system are not likely to be random, it is likely to be exogenous to school administrators and their behavior. Accountability systems are adopted by state legislators, and retention practices are not mentioned in any of the accountability systems analyzed in this paper.

I use an event-study design of the following form:

$$\begin{aligned} r_{gcdst} =& \alpha + \beta_1 \mathbb{1}(t - T_s^* \leq -2)_{st} + \beta_2 \mathbb{1}(t - T_s^* \in [0,2])_{st} \\ &+ \beta_3 \mathbb{1}(t - T_s^* \in [3,5])_{st} + \beta_4 \mathbb{1}(t - T_s^* \in [6,8])_{st} \\ &+ \gamma_c + \eta_g + \epsilon_{gcdst}. \end{aligned} \tag{11}$$

Here, r_{gcdst} represents the retention rate in grade g of campus c in district d of state s, year t, and T_s^*: year in which state s adopted a student score growth component in school ratings criteria. γ_c and η_g represent school and grade fixed effects respectively. In all specifications, I cluster my standard errors at the state-year level to allow for serial correlation, and because treatment is assigned at that level (Abadie et al., 2017).

Rather than estimate a separate coefficient for each post-adoption year, I group the event-study coefficients to gain precision. I cluster standard errors at the state-year level, to allow for serial correlation and correlation in the residuals of schools in the same state. The results of a placebo test show no pre-treatment effect and can be found in the first row of Table 4.

The administrator's decision to retain is most impactful in the last grade offered at his school. If he expects a student to perform poorly if held back, promoting the student removes her from the pool of students whose test scores determine the school's rating. To test the effect of growth-based accountability systems on retention practices in the last grade offered by a school, I estimate equation 2 separately for each grade, restricting attention to the first grade offered by a school, the middle grades offered, and the last grade offered. I also include a dummy variable identifying a grade as the last offered by a school as a third difference, estimating the following:

$$\begin{aligned} r_{gcdst} =& \alpha + \beta_1 \mathbb{1}(t - T_s^* \geq 0)_{st} \times \mathbb{1}(g = G_{ct}^T) + \beta_2 \mathbb{1}(t - T_s^* \geq 0)_{st} \\ &+ \beta_3 \mathbb{1}(g = G_{ct}^T) + \gamma_c + \epsilon_{gcdst}, \end{aligned} \tag{12}$$

where G_{ct}^T represents the final grade offered at school c in year t. The results of a placebo test show no statistically significant pre-treatment effect, and can be found in row 1, column (2) of Table 4.

5.2 Results

Table 4 presents the results of my baseline empirical analyses. The results of the estimation of equation (11) with and without grade fixed effects can be found in columns (1) and (2) respectively. I prefer the specification estimated in column (2), as retention strategy is likely different in different grades in similar ways across schools. For example, parent-initiated retention in early grades is common for matching a student's age to his cohort in a more preferable way. Both specifications, however, indicate that growth-based accountability policies have negative effects on overall retention rate within-school. The effect grows more pronounced over time, and after at least 6 years of growth-based accountability, it is statistically significant. I find that within school and grade, growth-based accountability reduces retention rates by 0.620 percentage points on average - a 30% decrease from the pre-growth average. Without the grade fixed effects, the effect is 0.589 percentage points, an 18.3% drop. Importantly, both columns (1) and (2) show no evidence of significant pre-trends. Though the estimates for the pre-growth accountability period are positive, they are not significant. Even without the statistical significance, the magnitude of the estimates may cause concern;

however, a pre-trend of the opposite sign of the estimated effect would bias my results towards zero, if anything.

Column (3) of Table 4 presents the results of estimating equation (12). The results suggest negative effects of a similar magnitude to the overall effect found in column (1) in each post-treatment period; perhaps because the impact of promoting a student out of school is certain, the impact of the policy was more immediate in the terminal grades offered by schools. Students in terminal grades are clearly held back at a much lower rate under both growth- and level-based accountability systems; however, they are held back at an even more reduced rate after a change to a growth-based system.

5.3 Sensitivity analysis

An emerging body of research has recently highlighted some of the weaknesses of difference-in-differences estimators in the presence of time-varying treatment. In particular, recent advances have shown that estimates are likely to be biased in the presence of heterogeneous treatment effects (Goodman-Bacon, 2018; Imai and Kim, 2020; Sun and Abraham, 2020). While using an event study design mitigates such concerns to a degree, the possibility of the presence of treated states in the control group causing biased estimates persists. Given the finding of only long-term significant effects in my baseline estimation, the concern is salient in this context. If a significant negative treatment effect does in fact exist in the years after adoption of growth-based accountability, previously-treated states will experience a decline in retention rates, and the difference in differences between those states and recently-treated states will be smaller than it would be if compared to a true control group. To investigate the possibility of bias, I drop all states but Texas and Colorado, the earliest adopters of growth-based accountability I observe, and Virginia, which did not include a growth component in its accountability system until 2018 (see Table 2 for timing of adoption), from my sample. By restricting my sample to these three states, I eliminate the possibility of the two treated states being used in the control group for any estimator; however, the strategy does reduce my estimating power somewhat and, importantly, uses only Virginia as a control state. Virginia is appealing as a control since I do not observe its schools under growth-based accountability; however, it is only one state, and estimates will be based on differences between Colorado and Texas and Virginia alone. In short, this strategy avoids some of the pitfalls of the weighted estimates used when treatment is staggered, but also forgoes its benefits. I estimate equation (11) with this limited sample as well as a design with a more flexible set of treatment dummies:

$$r_{gcdst} = \alpha + \beta_1 \mathbb{1}(t - T_s^* \leq -2)_{st} + \sum_{i=0}^{6} \beta_{i+2} \mathbb{1}(t - T_s^* = i)_{st} + \gamma_c + \eta_g + \epsilon_{gcdst}. \qquad (13)$$

The results of this regression and the "binned" version of the regression can be found in Table 5, column (2). The results of the two regressions are broadly similar, though there are some notable differences. In particular, I find negative effects of larger magnitudes in years 3, 4, and 5, and I find no significant effect in year 6. The differences in magnitudes are not particularly surprising due to the difference in control group, and the similarity in the overall pattern of effects is encouraging. However, the significance of effects in years earlier than those found in my main regression suggests the possibility of downward bias in the year 3-5 estimate of the main regression. I next estimate equations (11) and (13) using all states but Texas and Colorado. These two states are not the only potential sources of bias, but they did adopt growth-based accountability earliest. The results of these regressions can be found in Column (3) of Table 5. Interestingly, I do find significant negative effects in years 3 and 5, and on the years 3-5 estimate of the binned regression. These results, taken together, suggest to me that the negative effects of growth-based accountability on retention may set in earlier than my main regression suggests.

Another concern is with my definition of treatment. In all estimations up to this point, I define treatment as the year after adoption of growth-based accountability. That is, if a state adopts a new accountability system including a growth-based component ahead of the 2013-14 school year (Year 0), I define schools in the state as treated in all future years starting with that year. However, it is common among states in this sample to issue ratings to schools in the year of adoption of a new accountability system, but not implement any of the consequences of those ratings. School administrators are typically aware of the lack of consequences when this is done. As a result, it's possible that treatment should not "turn on" until Year 1. While administrators do receive ratings in Year 0 in these cases, they may not change their strategic behavior until the consequences are put in place, and use Year 0 as a learning year. Such behavior has been documented in Texas schools by Craig, Imberman and Perdue (2013). I re-estimate equations (11) and (13), using the full sample as well as the two restricted samples discussed above, using Year 0 as the reference year. The results of this estimation can be found in Table 6. Interestingly, the significance of the results changes slightly. I find effects of similar magnitude in all cases, and I find effects significant at the 95% level in Years 2, 5 and 6 of treatment. The results in columns (2) and (3) are similar in shape to those presented in Table 5, but also show significance starting in year 2. Again, the lack of effect in years 3 and 4 in particular in the full-sample regression in contrast with the results from the 3-state and 4-state regressions in columns (2) and (3) suggest possible downward bias in the full-sample regression. I also estimate the same two regressions using Years -1 and 0 as the reference group, and the results are quite similar to those presented in Table 6. The results of the regressions using Years -1 and 0 can be found in Table 7.

5.4 Effect heterogeneity

If a school administrator has certain beliefs about how different subpopulations might perform after being retained, they may be differentially impacted by the switch to growth-based accountability. In addition, since the performances of some subpopulations are heavily weighted in many accountability systems, the policy impact on retention practices may differ by subpopulation.

The Texas Education Agency (TEA) makes retention rates available at the subpopulation-grade-year level; as a result, I am able to test for the effect of Texas' switch to a growth-based accountability system on subgroup-specific retention rates. No other state in my sample publishes retention rates by subgroup/grade, so I must use a new estimation strategy. I use a simple difference-in-differences design. Assignment to treatment depends on the year of observation - Texas school ratings included a measure of student growth starting in the 2012-13 school year - and I use an indicator for whether or not a given grade is the last offered by a school as a measure of treatment intensity. Thus, I am comparing the retention rates of various subgroups in the last grade offered by a school to those of the same groups in all other grades offered by the school before and after exposure to a growth-based accountability system. I estimate regressions of the following form:

$$\begin{aligned} r_{pgct} =& \alpha + \beta_1 \mathbb{1}(t \geq 2013)_t \times \mathbb{1}(g = g_c^T)_{gct} + \beta_2 \mathbb{1}(t \geq 2013)_t \\ &+ \beta_3 \mathbb{1}(g = g_c^T)_{gct} + n_{pct} + \eta_c + \epsilon_{gct} \end{aligned} \tag{14}$$

where r_{pgct} represents the retention rate of subgroup p students in grade g at campus c in year t, g_c^T represents the last grade offered at campus c and n_{pct} represents the total number of sudents of subgroup p enrolled in school c in year t. I estimate this equation separately for elementary and middle schools.[2] The validity of this approach requires that retention rates in terminal and non-terminal grades exhibiting parallel trends in the pre-treatment period. Most subgroups of interest fail an informal parallel trend check; for this reason, I focus only on male, Black, Hispanic, and white students in this analysis. Figure A.1 in Appendix A includes plots of the retention rate for each of these subgroups over time, from 2004-05 to 2017-18, in terminal and non-terminal grades.

5.5 Results

Table 8 presents the effects of growth-based accountability on the retention rates of male, Black, Hispanic, and white students in Texas. I present the results for all four subgroups, but only male students pass a placebo test; as such, the results for Black, Hispanic, and white students should be considered non-causal. I estimate equation 4 separately for elementary, middle, and high schools. The results presented in Table 8, panel A show that the switch to growth-based accountability had a relatively small but significant effect on the retention rate of elementary school boys. The retention rate among boys in terminal grades decreases by .137 percentage points after the switch to growth-based accountability - a 4% decrease relative to the pre-growth average retention rate. This is largely fueled by a decrease in the retention rate of high school boys in terminal grades (i.e. high school graduates), and partly by a decrease in the retention rate of elementary school boys. The rate at which elementary school boys are held back decreases by .056 percentage points after the switch to growth-based accountability -

[2]High schoolers that repeat a course are counted as retained by TEA; for this reason, I choose to exclude them from my analysis.

a 1.6% decrease relative to the pre-growth average retention rate. The results of Panels B, C, and D show similar patterns - modest decreases in terminal-grade retention rates after the shift to growth-based accountability. Notably, the decreases for Black and Hispanic students seem to be driven primarily by decreases in the terminal-grade retention rate in middle school rather than in elementary or high school. Overall, I do not find compelling evidence that any of these subgroups are more affected by the policy change than others.

6 Costs and benefits of Growth-based accountability systems

In this section, I use estimates of the effects of retention from the literature to calculate some back-of-the envelope costs and benefits associated with the decrease in retention rates caused by growth-based accountability systems. The results of my analyses suggest that retention rates drop on average by around 18% after a state implements a growth-based accountability system, and that the change is more immediate in terminal grades. The total enrollment across grades 3-8 in the U.S. was 22,890,943 for the 2018-19 school year. The average retention rate in a school under a growth-based system was 2.020 in my data. Using these numbers, I calculate that around 457,819 students are retained each year, and that growth-based accountability systems lead to around 100,497 fewer retained students each year. National Center of Education Statistics (2019) reports that 2019 average per-student expenditures in the US were \$13,847. Assuming that a retained student spends an additional year in school relative to the case in which he is not retained, these numbers taken together suggest that the reduction in retentions caused by growth-based accountability lead to around \$1.392 billion saved in student expenditures.

I combine estimates from the literature with various sources to calculate per-student costs and benefits of retention. These calculations and sources can be found in Table 9. Prior research has shown that retention has substantial negative long-term effects on students, especially those retained in later grades.

Based on estimates from Brodaty, Gary-Bobo and Prieto (2013), I calculate that retention decreases the beginning-of-career wage of a student by \$2,682 (in 2019 dollars) due to delayed entry into the job market. If retained students accept lower-paying starting jobs in the future, this could affect their future wages, as well. Deveraux (2002) estimates that about 60% of the wage differential between two individuals that started jobs at the same time could be explained by the difference in the starting wage. Through their effect on retention practices, growth-based accountability systems stop the retention of 100,500 students, and thus around \$270 million in lost beginning-of-career wages, and potentially even more in future wages.

Based on estimates from Eren, Lovenheim and Mocan (2018), I calculate that a student retained in the 8th grade will cause \$338 more in violent crime costs in expectation, and \$564 more in drug offense costs. In total, growth-based accountability systems avoid around 58,000 violent crimes based on figures from Federal Bureau of Investigation (2019) and \$91 million in crime-related costs. Beyond these costs, violent crimes cause long-term socio-emotional problems for their victims, reduce physical activity of neighbors, and lower house values (Langton, 2014; Janke et al., 2016; Taylor, 1995).

Based on estimates from Manacorda (2012) and estimates on the returns to education from Kolesnikova (2010), I calculate that an individual retained in middle school will earn $1,000-$4,000 less in yearly wages than he would in the absence of the retention. Other researchers have found a host of negative effects associated with decreased wages, including increased chances of divorce and decrease chances of marrying, decreased access to childcare, and increased levels of obesity and hypertension (Fremstad and Boteach, 2015; Census Bureau, 2011; Leigh, 2013). Through their effect on retention practices, growth-based accountability systems avoid $101-404 million in lost annual wages, as well as additional costs associated with low wages.

On the other hand, some research has found retention to have academic benefits, particularly in the short-term. Schwerdt, West and Winters (2017) find that retained students enrolled in fewer remedial courses later in life; based on their findings, I calculate that retained students spend around $4,034 less on remedial reading courses and $1,121 less on remedial math courses because they were retained. Jacob and Lefgren (2004) find evidence of gains to standardized tests scores among retained 3rd graders, and Schwerdt, West and Winters (2017) find that retention increases high school GPA's. While both of these benefits are real and non-negligible, the financial implications of them are less clear-cut, and I was unable to calculate a concrete number based on these estimates. Through their effect on retention practices, growth-based accountability systems increase spending on remedial reading and math courses by around $518 million, and may incur additional costs related to reduced GPA and foregone early test score bumps for students.

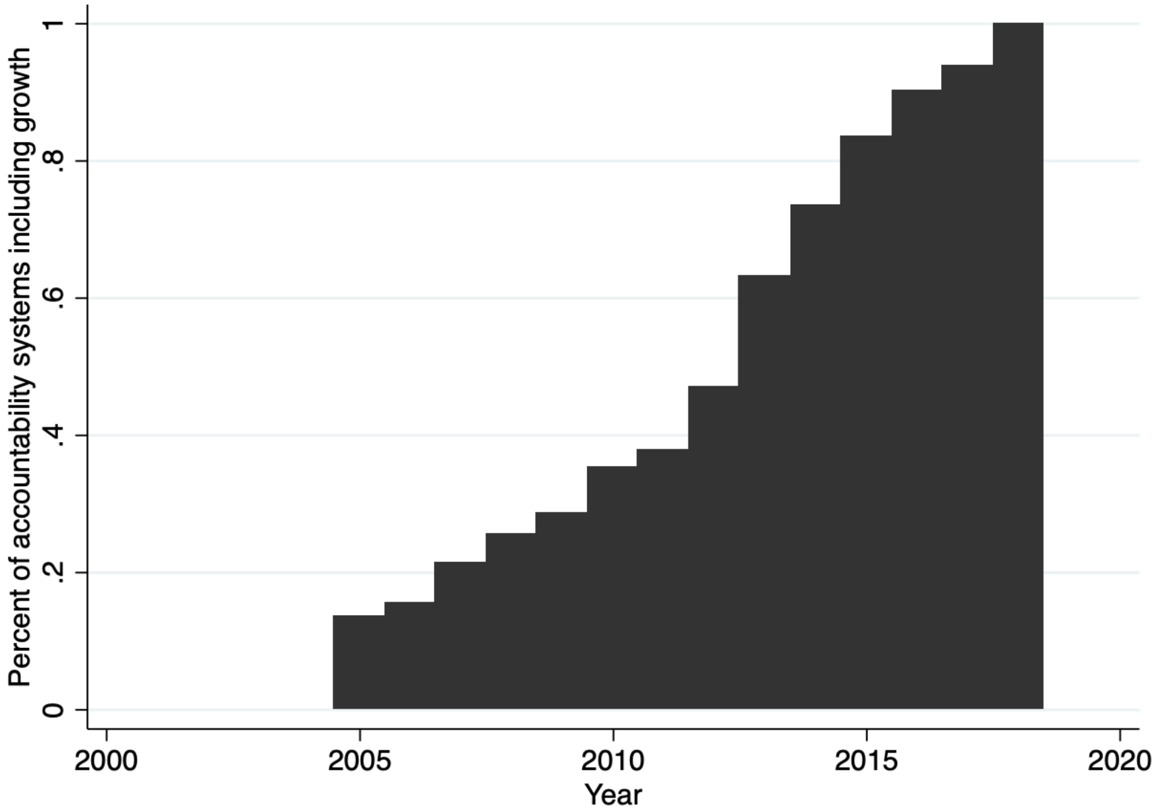
By suppressing retention, growth-based accountability systems avoid many of the long-term negative consequences of retention that may have otherwise been experienced by around 2.7 million students every year. This effect reduces student expenses by the public school system by around $1.4 billion, increases annual wages of the would-be-retained by $100-400 million in total, and avoids $91 million in crime-related costs. Additional negative externalities may exist beyond these. The change does incur $518 million in remedial course enrollment fees for students not retained that otherwise would have been.

7 Conclusion

This paper provides evidence that public school administrators use retention as a tool in optimizing their school's rating. They respond to changes in the way the school is evaluated - particularly a switch to school ratings that incorporate year-to-year individual student score growth - by changing the rate at which they retain students. My estimates, based on a 7-state sample, suggest that schools retain 18% fewer students after operating under a growth-based system for at least 6 years, and that the effect is more pronounced in the final grade offered by a school, where promotion of a student evicts them from the school's pool of test-takers. I find little evidence of effect heterogeneity by gender or ethnicity. Overall, administrators choose to retain less when student score growth matters to their school ratings, particularly when they are able to remove the student from the test-taking pool via promotion, suggesting that administrators use retention strategically.

8 Tables & Figures

Figure 1: Presence of growth in accountability systems over time



Notes: Includes all 50 states and Washington, D.C.
Sources: Education Week Education Counts Research Center; individual states' education agencies.

Figure 2: Level and year-to-year change in student score are affected by grade level

Figure 2a: Student scores for two different grades

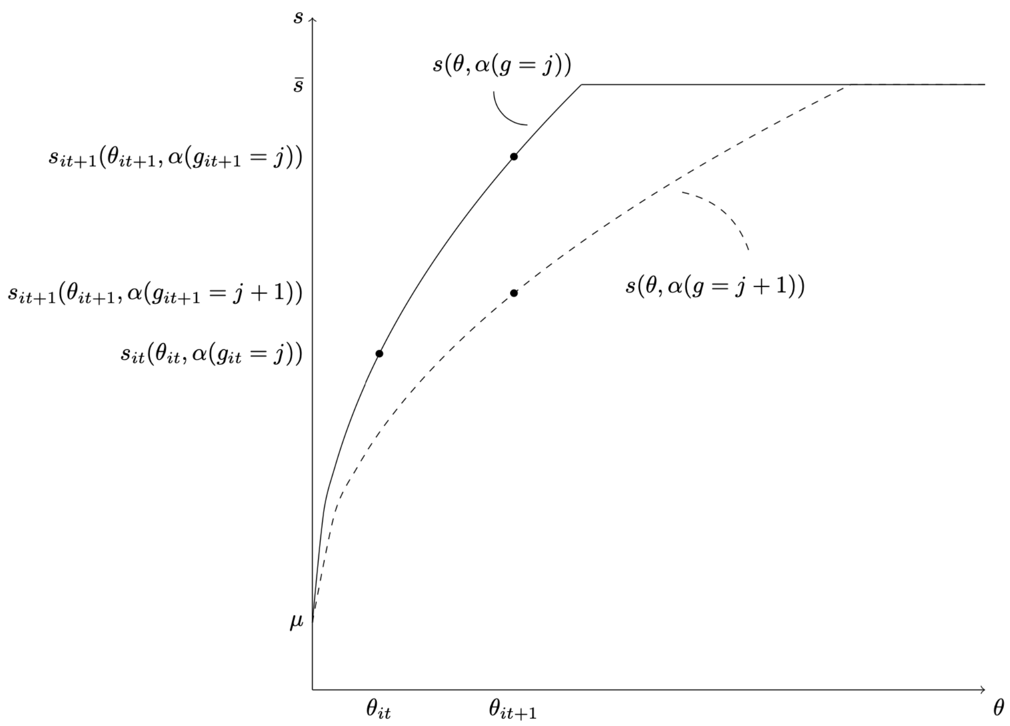


Figure 2b: Rate of change in student scores for two different grades

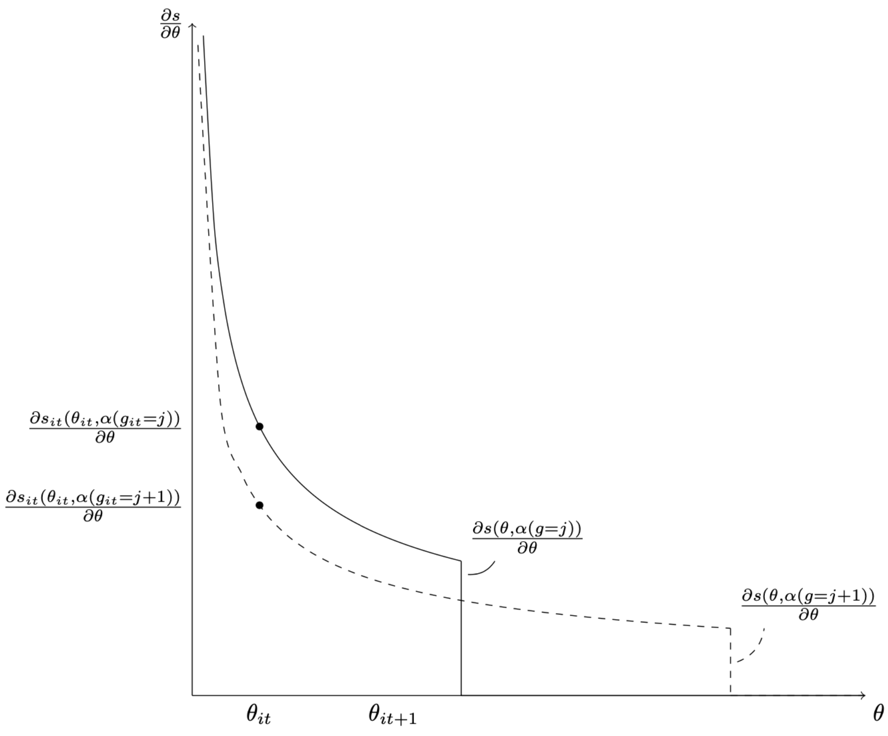


Figure 3: Passing and growth rate-maximizing retention rates

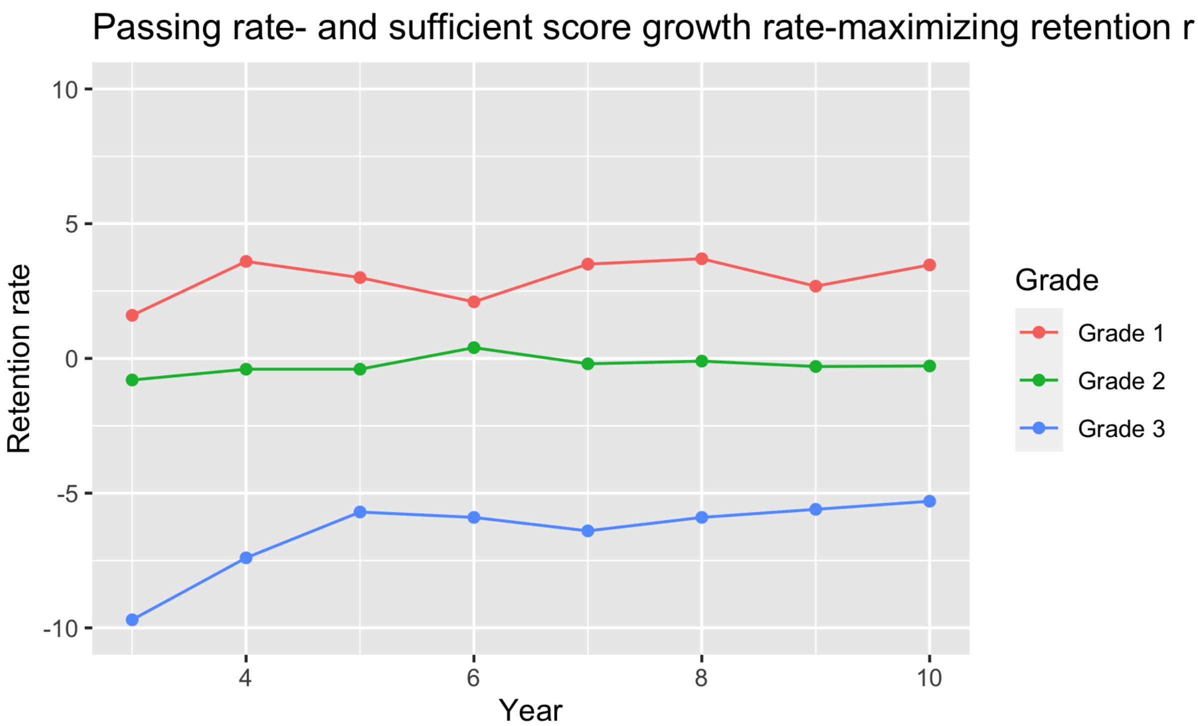


Table 1: Timing of adoption of growth-based accountability

	CO	TX	RI	MA	MI	LA	VA
2011-12							
2012-13	✓	✓	✓				
2013-14	✓	✓	✓				
2014-15	✓	✓	✓	✓			
2015-16	✓	✓	✓	✓			
2016-17	✓	✓	✓	✓	✓		
2017-18	✓	✓	✓	✓	✓	✓	

Notes: ✓indicates that state accountability system included a measure of student growth and that the system had been in place for at least one school year prior.

Table 2: Parameter values for simulation

	Mean	S.D.
μ	0	0
ϵ_{it}	0	20
γ_0	35	0
γ_1	1	0
θ_{i0}	50	25[1]
η_{it}	0	10
π	50	0
λ	10	0
Number of students entering grade 1 each year	100	0
Number of years simulated	10	0
Number of simulations	10000	0

Notes:
[1]: Initial θ_{it} draw for students entering school.

Table 3: Descriptive statistics

	Whole sample	No growth-based accountability	Growth-based accountability
Growth component	0.899	0.000	1.000
	(0.301)	(0.000)	(0.000)
Grade	3.831	3.879	3.826
	(2.223)	(2.152)	(2.231)
Terminal grade	5.973	5.882	5.984
	(1.511)	(1.456)	(1.516)
Grade-level male pop. share	0.481	0.482	0.481
	(0.102)	(0.098)	(0.103)
Grade-level white pop. share	0.424	0.372	0.430
	(0.347)	(0.323)	(0.349)
Grade-level Black pop. share	0.148	0.132	0.150
	(0.220)	(0.188)	(0.223)
Grade-level Hispanic pop. share	0.428	0.496	0.420
	(0.334)	(0.319)	(0.335)
Grade-level enrollment	95.158	99.210	94.689
	(82.378)	(84.322)	(82.137)
% of students retained in grade	2.017	2.092	2.009
	(4.405)	(4.621)	(4.380)
N	298102	29980	268122

Notes: Standard deviations in parentheses.

Table 4: The Effect of Growth-based Accountability on Retention

	(1)	(2)	(3)
Years -5 to -2	0.097	0.136	0.033
	(0.061)	(0.096)	(0.063)
Years 0 to 2	-0.060	-0.087	0.053
	(0.080)	(0.085)	(0.079)
Years 3 to 5	-0.092	-0.122	0.002
	(0.087)	(0.092)	(0.087)
Year 6	-0.589***	-0.620***	-0.517***
	(0.094)	(0.098)	(0.103)
$\mathbb{1}(g = G_c^T)$			-0.625***
			(0.008)
Years -5 to -2 $\times\mathbb{1}(g = G_c^T)$			0.369***
			(0.011)
Years 0 to 2 $\times\mathbb{1}(g = G_c^T)$			-0.687***
			(0.059)
Years 3 to 5 $\times\mathbb{1}(g = G_c^T)$			-0.614***
			(0.041)
Year 6 $\times\mathbb{1}(g = G_c^T)$			-0.524***
			(0.089)
Grade FE		✓	
N	289820	289820	289820

Notes: Dependent variable is grade-level retention rate. Standard errors, clustered at the state-year level, are reported in parentheses. * denotes significance at 10%, ** denotes 5%, and *** denotes 1%. All regressions include school fixed effects.

Table 5: Testing for potential bias stemming from effects on early adopters

	(1) Full sample	(2) TX, CO, VA only	(3) TX, CO removed
Years -5 to -2	0.096	0.093	0.071
	(0.061)	(0.108)	(0.093)
Year 0	0.035	0.037	0.116
	(0.080)	(0.098)	(0.102)
Year 1	-0.023	0.001	0.081
	(0.082)	(0.099)	(0.100)
Year 2	-0.176	-0.127	-0.048
	(0.085)	(0.097)	(0.083)
Year 3	0.016	-0.217*	-0.137*
	(0.103)	(0.097)	(0.082)
Year 4	-0.084	-0.208*	-0.128
	(0.085)	(0.100)	(0.089)
Year 5	-0.252**	-0.308**	-0.227**
	(0.077)	(0.100)	(0.075)
Year 6	-0.597***	-0.190	
	(0.088)	(0.143)	
Years -5 to -2	0.097	0.094	0.171
	(0.061)	(0.108)	(0.093)
Years 0 to 2	-0.060	-0.042	0.037
	(0.080)	(0.092)	(0.101)
Years 3 to 5	-0.092	-0.226*	-0.148*
	(0.087)	(0.092)	(0.084)
Years 6 to 8	-0.589***	-0.188	
	(0.094)	(0.143)	
N	289820	65167	238973

Notes: Dependent variable is grade-level retention rate. Standard errors are reported in parentheses. Errors are clustered at the state-year level in the regressions corresponding to columns (1) and (3), and are heteroskedasticity-robust in the regression corresponding to column (2). * denotes significance at 10%, ** denotes 5%, and *** denotes 1%. All regressions include school fixed effects.

Table 6: Testing sensitivity of results to definition of treatment variable: reference year is 0

	(1) Full sample	(2) TX, CO, VA only	(3) TX, CO removed
Years -5 to -2	0.061	0.011	0.093
	(0.113)	(0.084)	(0.112)
Year -1	0.035	0.037	0.000
	(0.080)	(0.098)	(0.071)
Year 1	-0.058	-0.035	-0.026
	(0.041)	(0.051)	(0.019)
Year 2	-0.210***	-0.164**	-0.154**
	(0.049)	(0.050)	(0.019)
Year 3	-0.018	-0.253***	-0.244***
	(0.078)	(0.051)	(0.020)
Year 4	-0.118*	-0.244***	-0.235***
	(0.043)	(0.052)	(0.049)
Year 5	-0.287***	-0.345***	-0.307***
	(0.029)	(0.080)	(0.061)
Year 6	-0.631***	-0.227	
	(0.051)	(0.149)	
Years -5 to -1	-0.003	0.011	0.094
	(0.093)	(0.084)	(0.099)
Years 1 to 2	-0.134*	-0.098*	-0.077
	(0.056)	(0.044)	(0.053)
Years 3 to 4	-0.067	-0.248***	-0.226***
	(0.059)	(0.045)	(0.037)
Years 5 to 6	-0.512***	-0.290**	
	(0.106)	(0.090)	
N	289820	65167	238973

Notes: Dependent variable is grade-level retention rate. Standard errors are reported in parentheses. Errors are clustered at the state-year level in the regressions corresponding to columns (1) and (3), and are heteroskedasticity-robust in the regression corresponding to column (2). * denotes significance at 10%, ** denotes 5%, and *** denotes 1%. All regressions include school fixed effects.

Table 7: Testing sensitivity of results to definition of treatment variable: reference years are -1 and 0

	(1) Full sample	(2) TX, CO, VA only	(3) TX, CO removed
Years -5 to -2	0.067	0.065	0.093
	(0.101)	(0.094)	(0.100)
Year 1	-0.057	-0.033	-0.026
	(0.040)	(0.050)	(0.019)
Year 2	-0.210***	-0.162**	-0.154***
	(0.049)	(0.049)	(0.019)
Year 3	-0.018	-0.251***	-0.244***
	(0.078)	(0.050)	(0.021)
Year 4	-0.118*	-0.242***	-0.235***
	(0.042)	(0.051)	(0.051)
Year 5	-0.286***	-0.337***	-0.307***
	(0.028)	(0.074)	(0.051)
Year 6	-0.631***	-0.219	
	(0.051)	(0.142)	
Years -5 to -2	0.053	0.066	0.133
	(0.093)	(0.094)	(0.100)
Years 1 to 2	-0.133*	-0.096*	-0.079
	(0.056)	(0.043)	(0.052)
Years 3 to 4	-0.066	-0.246***	-0.228***
	(0.059)	(0.044)	(0.037)
Years 5 to 6	-0.511***	-0.282***	
	(0.105)	(0.083)	
N	289820	65167	238973

Notes: Dependent variable is grade-level retention rate. Standard errors are reported in parentheses. Errors are clustered at the state-year level in the regressions corresponding to columns (1) and (3), and are heteroskedasticity-robust in the regression corresponding to column (2). * denotes significance at 10%, ** denotes 5%, and *** denotes 1%. All regressions include school fixed effects.

Table 8: Effect of Growth-based Accountability on Retention Rate: Texas

	Full sample	Elementary school	Middle school	High school
Panel A: Male students				
Post-growth	-0.014	-0.009	-0.003	-0.130
	(0.016)	(0.009)	(0.045)	(0.121)
$\mathbb{1}(g = G_c^T)$	0.065**	0.102***	0.075*	-0.597***
	(0.024)	(0.019)	(0.031)	(0.142)
Post-growth $\times$ $\mathbb{1}(g = G_c^T)$	-0.137***	-0.056*	-0.075	-0.660*
	(0.037)	(0.023)	(0.051)	(0.286)
N	245702	181322	41247	23133
Panel B: Black students				
Post-growth	0.002	0.002	0.004	-0.001
	(0.002)	(0.001)	(0.003)	(0.011)
$\mathbb{1}(g = G_c^T)$	0.025***	0.008*	0.018	0.039
	(0.007)	(0.004)	(0.011)	(0.023)
Post-growth $\times$ $\mathbb{1}(g = G_c^T)$	-0.020**	-0.005	-0.011	-0.040
	(0.007)	(0.004)	(0.010)	(0.026)
N	309871	232045	52542	25284
Panel C: Hispanic students				
Post-growth	-0.005	0.009*	0.018	-0.056
	(0.006)	(0.005)	(0.021)	(0.043)
$\mathbb{1}(g = G_c^T)$	0.025*	0.035**	0.056**	0.082
	(0.011)	(0.013)	(0.018)	(0.055)
Post-growth $\times$ $\mathbb{1}(g = G_c^T)$	-0.028*	-0.020	-0.060*	-0.056
	(0.012)	(0.011)	(0.027)	(0.074)
N	261830	197226	45884	18720
Panel D: White students				
Post-growth	0.003	-0.001	0.006	0.018
	(0.001)	(0.001)	(0.004)	(0.010)
$\mathbb{1}(g = G_c^T)$	0.021***	0.015*	0.012	0.083*
	(0.006)	(0.007)	(0.008)	(0.034)
Post-growth $\times$ $\mathbb{1}(g = G_c^T)$	-0.016**	-0.010	-0.004	-0.094**
	(0.006)	(0.006)	(0.010)	(0.034)
N	309790	237318	51782	20690

Notes: Dependent variable is grade-level retention rate. Standard errors, clustered at the district-year level, are reported in parentheses. * denotes significance at 10%, ** denotes 5%, and *** denotes 1%. All regressions include school fixed effects and control for the total enrolled white students at the school.

Table 9: Estimated effects of retention

Mechanism	Per-student yearly cost/benefit of retention	Source for effect of retention
Decreased beginning-of-career wages	$2,682 decrease in beginning-of-career wage	Brodaty, Gary-Bobo and Prieto (2013)[1]
Increased probability of committing violent crime as an adult	$338 more in expected violent crime costs	Eren, Lovenheim and Mocan (2018)[2]
Increased probability of drug conviction	$564 more in expected drug offense costs	Eren, Lovenheim and Mocan (2018)[3]
Decreased educational attainment	$1,000-$4,000 decrease in earnings	Manacorda (2012)[4]
Lower remedial reading course enrollment	$4,034 less in remedial course expenses	Schwerdt, West and Winters (2017)[5]
Lower remedial math course enrollment	$1,121 less in remedial course expenses	Schwerdt, West and Winters (2017)[6]

[1] Brodaty, Gary-Bobo and Prieto (2013) calculate that a one-year delay into the job market caused by retention decreases beginning-of-career wages by 9%. The number given combines this estimate with the average 2019 entry-level salary of $32,592 (ZipRecruiter, 2020).

[2] Eren, Lovenheim and Mocan (2018) estimate a 58.44% increase in probability of committing violent crime as an adult when retained in eighth grade. Federal Bureau of Investigation (2019) report 1,203,808 violent crimes in the U.S. in 2019, and Miller, Cohen and Wiersema (1996) estimate that violent crime imposed annual costs of $426 million in the US ($694.49 million in 2019 dollars). These numbers suggest that each violent crime costs $577 in 2019 dollars. The number given in row two combines this per-crime cost with the estimate given in Eren, Lovenheim and Mocan (2018).

[3] Eren, Lovenheim and Mocan (2018) estimate a 10.02% increase in probability of a drug conviction when retained in eighth grade. OIson and Stout (1991) estimate that the cost of investigating and arresting a drug offender in 1989 was $2,711 ($5,635 in 2019 dollars). The number given in row three combines this per-offfense cost with the estimate given in Eren, Lovenheim and Mocan (2018).

[4] Kolesnikova (2010) suggests a 10% return to a year of education for practical purposes. Social Security Administration (2020) report the average wage of an American in 2019 was $51,916, suggesting a rough return of $5,000 per year of education. The number given in row 4 combines this with the estimate given in Manacorda (2012) of 0.2-0.8-year decrease in educational attainment caused by retention.

[5] Douglas-Gabriel (2016) reports that the average cost of a remedial course at a four-year institution was around $3,000 in 2016 ($3,201.57 in 2019 dollars). The number given in row 5 combines this with the estimate of a 1.26-course decrease in remedial reading course enrollment due to retention given in Schwerdt, West and Winters (2017).

[6] Douglas-Gabriel (2016) reports that the average cost of a remedial course at a four-year institution was around $3,000 in 2016 ($3,201.57 in 2019 dollars). The number given in row 6 combines this with the estimate of a 0.35-course decrease in remedial math course enrollment given in Schwerdt, West and Winters (2017).

References

Abadie, Alberto, Susan Athey, Guido Imbens, and Jeffrey Wooldridge. 2017. "When Should You Adjust Standard Errors for Clustering?"

Abraham, Sarah, and Liyang Sun. 2018. "Estimating Dynamic Treatment Effects in Event Studies With Heterogeneous Treatment Effects." *Journal of Econometrics*.

Alliance for Excellent Education. 2016. "Every Student Succeeds Act Primer: Personalized Learning." Alliance for Excellent Education April, Washington, D.C.

Almond, Douglas, Ajin Lee, and Amy Ellen Schwartz. 2016. "Retention Heterogeneity in New York City Schools."

Balingit, Moriah. 2017. "Virginia drafts new plan to hold struggling schools accountable."

Brodaty, Thomas O, Robert J Gary-Bobo, and Ana Prieto. 2013. "Does Speed Signal Ability? The Impact of Grade Retention on Wages."

Chakrabarti, Rajashri. 2013. "Vouchers, public school response, and the role of incentives: Evidence from florida." *Economic Inquiry*, 51(1): 500–526.

Craig, Steven G., Scott A. Imberman, and Adam Perdue. 2013. "Does it pay to get an A? School resource allocations in response to accountability ratings." *Journal of Urban Economics*, 73(1): 30–42.

Cullen, Julie Berry, and Randall Reback. 2006. "Tinkering Toward Accolades: School Gaming under a Performance Accountability System."

Cunha, Flavio, and James J Heckman. 2008. "Formulating , Identifying and Estimating the Technology of Cognitive and Noncognitive Skill Formation." *The Journal of Human Resources*, 43(4): 738–782.

Douglas-Gabriel, Danielle. 2016. "GradePoint: Remedial classes have become a hidden cost of college." *Washington Post*, , (11/27/2017).

Eren, Ozkan, Michael F. Lovenheim, and Naci H. Mocan. 2018. "The effect of grade retention on adult crime: Evidence from a test-based promotion policy."

Federal Bureau of Investigation. 2019. "Violent Crime."

Figlio, David N., and Joshua Winicki. 2005. "Food for thought: The effects of school accountability plans on school nutrition." *Journal of Public Economics*, 89(2-3): 381–394.

Fruehwirth, Jane Cooley, Salvador Navarro, and Yuya Takahashi. 2016. "How the Timing of Grade Retention Affects Outcomes : Identification and Estimation of Time-Varying Treatment Effects." *Jornal of Labor Economics*, 34(4): 979–1021.

Goodman-Bacon, Andrew. 2018. "Difference-in-Differences With Variation in Treatment Timing."

Imai, Kosuke, and In Song Kim. 2020. "On the Use of Two-Way Fixed Effects Regression Models for Causal Inference with Panel Data." *Political Analysis*, 1–11.

Jacob, Brian A., and Lars Lefgren. 2004. "Remedial education and student achievement: A regression-discontinuity analysis." *Review of Economics and Statistics*, 86(1): 226–244.

Jacob, Brian A., and Lars Lefgren. 2009. "The Effect of Grade Retention on High School Completion." *American Economic Journal: Applied Economics*, 1(3): 33–58.

Kolesnikova, Natalia. 2010. "The return to education isn't calculated easily." *Regional Economist*, , (January): 1–12.

Manacorda, Marco. 2012. "The cost of grade retention." *Review of Economics and Statistics*, 94(2): 596–606.

Martin, Carmel, Scott Sargrad, and Samantha Batel. 2016. "Making the grade: A 50 state analysis of school accountability systems." Center for American Progress May.

Miller, Ted R., Mark A. Cohen, and Brian Wiersema. 1996. "Victim Costs and Consequences: A New Look."

National Center of Education Statistics. 2019. "Public School Expenditures."

OIson, D.E., and L.S. Stout. 1991. "Cost of Processing a Drug Offender Through the Criminal Justice System." Chicago.

Reback, Randall. 2008. "Teaching to the rating : School accountability and the distribution of student achievement." *Journal of Public Economics*, 92: 1394–1415.

Rockoff, Jonah, and Lesley J. Turner. 2010. "Short-run impacts of accountability on school quality." *American Economic Journal: Economic Policy*, 2(4): 119–147.

Schwerdt, Guido, Martin R. West, and Marcus A. Winters. 2017. "The effects of test-based retention on student outcomes over time: Regression discontinuity evidence from Florida." *Journal of Public Economics*, 152: 154–169.

Sentell, Will. 2017. "Growth over grades: Key panel agrees on changes in public school rating system."

Social Security Administration. 2020. "Measures of Central Tendency for Wage Data."

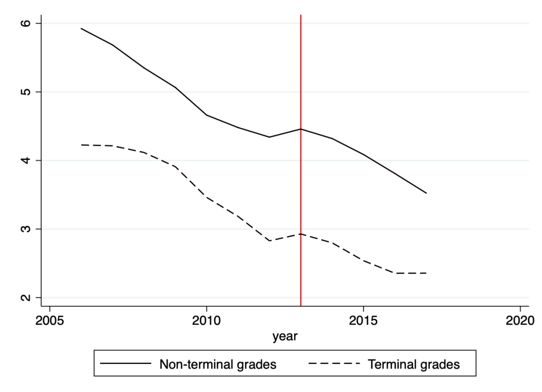
Sun, Liyang, and Sarah Abraham. 2020. "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects." *Journal of Econometrics*, , (xxxx): 1–25.

Zaniewski, Ann, and Lori Higgins. 2016. "House battle awaits $720M DPS plan."

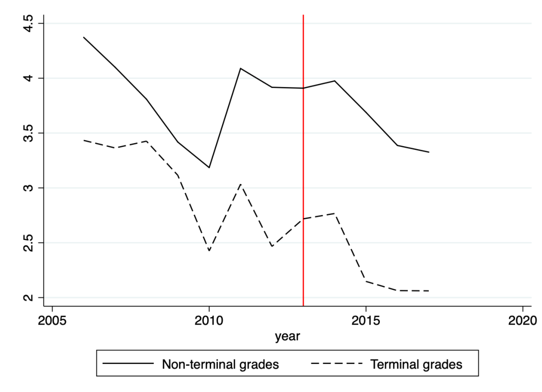
ZipRecruiter. 2020. "What is the Average Entry Level Salary by State."

A Parallel trends figures for effect heterogeneity analysis

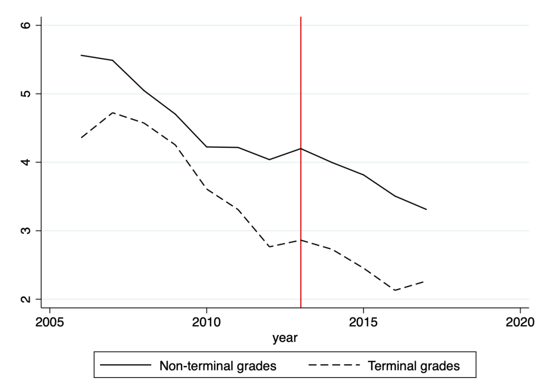
Figure A.1: Parallel trend checks



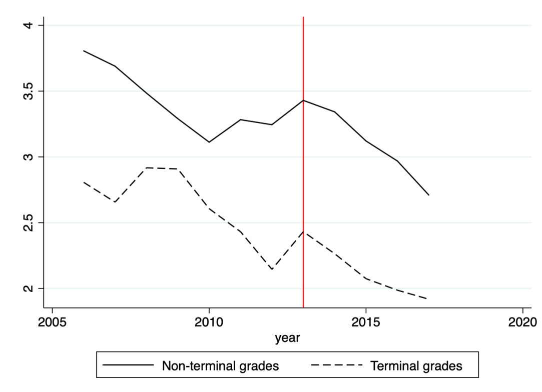
(a) Retention rate among male students



(b) Retention rate among Black students



(c) Retention rate among Hispanic students



(d) Retention rate among white students

B Data appendix

The following provides greater detail on my data sources.

Accountability data

I relied on Education Week's Education Counts Research Center for identifying variables on whether or not states "assigns ratings to all schools based on state-developed criteria", "uses measures of individual student growth to rate schools", and "uses measures of individual student growth for state ratings" for the 2011-12 school year. The latter two variables both were coded as treatment variables, as they were used in different years of the survey. For the 2015-16 school years, I relied on a May 2016 report from the Center for American Progress analyzing the accountability components of each state's accountability system (Martin, Sargrad and Batel, 2016). For all other years, I ascertained the timing of growth-based accountability criteria adoption through various means including state statutes and regulations, state education agency reports and records, and local reporting. I relied on newpapers.com for access to newspaper articles, on the National Council of State Legislatures' bill tracking tool to access legislation, and state education agency sites in conjunction with the Internet Archive's Wayback Machine to access historical agency reports.

Texas' accountability system standards are presented for each year 2005 on the Texas Education Agency website. Colorado's Department of Education published a report detailing the evolution of the state's accountability system in 2019; it can be found here. Massachusetts began tracking growth in 2009, but did not include it in accountability ratings until the 2013-14 school year. A 2014 memo from the Commissioner of Massachusetts' Board of Elementary and Secondary Education describes the incorporation of student growth into the state's accountability system, and can be found here. Rhode Island adopted its current accountability system, which includes student growth, in 2012, and first implemented it in the 2012-13 school year. The status of growth in its accountability system was checked using the Wayback Machine; the archived webpage showing school classifications and components by which they are classified can be found here. Louisiana included a growth-based component in its evaluation of schools for the first time in the 2017-18 school year, according to a 2017 Baton Rouge newspaper article (Sentell, 2017). Virginia included a growth-based component in its evaluation of schools for the first time in the 2018-19 school year, according to a 2017 Washington Post article (Balingit, 2017). Michigan implemented a new student growth indicator which was included in school report cards for the first time in the 2016-17 school year, according to a 2016 Detroit Free Press article (Zaniewski and Higgins, 2016).

Retention data

My data on retention rates in Texas was taken from the Texas Academic Performing Reports. The reports can be downloaded from the Texas Education Agency's website. Data from Colorado was provided by Colorado's Department of Education by request.

Data from Massachusetts is available through the Massachusetts Department of Elementary and Secondary Education at https://www.doe.mass.edu/DataAccountability.html. Data from Rhode Island can be requested through the Frequently Requested Data portal at the Rhode Island Department of Elementary and Secondary Education's website. Data from Michigan can be requested from MI School Data here. Data from Louisiana is available through the Louisiana Department of Education here. Data from Virginia was provided by the Virginia Department of Education at request. These data sets sometimes use their own conventions for identifying schools, such that two schools in Michigan and Texas might both be identified as School 1 in County 1 in District 1. To avoid such a scenario, I relied on the Common Core of Data, which identifies every public school in the country with an NCES ID number. It also includes the relevant state education agency's school identification numbers. Using this, I was able to merge the Common Core of Data with each state's data separately, and link schools to their NCES ID numbers, which are unique.