Included To Be Excluded: Accountability Pressure and Students in Special Education*

Gue Sung Choi†

October 14, 2023

Abstract

This study examines how schools respond to accountability pressures for students with disabilities and how such strategic responses affect their long-term outcomes. Using administrative data for all public schools in Texas, I find that in response to the mandate of incorporating students in special education into accountability measures, schools resorted to granting mass test exemptions to these students to protect their ratings. The weakest students were more likely to be removed, and schools with lower prior ratings were more likely to react. Furthermore, cohort-level analysis shows that such exclusion led to adverse long-term outcomes such as less high school graduation and employment in adulthood. These results indicate that poorly designed school incentives could lead to unintended school behaviors and, consequently negative impacts on students who were intended to be helped, even in the long run.

Keywords: School Accountability, high-stakes Exam, Special Education

JEL Codes: I12, I13

*I am grateful to Richard Murphy, Stephen Trejo, and Manuela Angelucci for their invaluable support and guidance. I also thank Scott Carrell, Cody Tuttle, Brendan Kline, Eric Chyn, Jinyeong Son, Bokyung Kim, Jori Barash, Ziyue Xu, Yumin Hong and Mu Yang Shin for their helpful comments. The research presented here utilizes confidential data from the State of Texas supplied by the Education Research Center (ERC) at The University of Texas at Austin. The views expressed are those of the authors and should not be attributed to the ERC or any of the funders or supporting organizations mentioned herein, including The University of Texas at Austin or the State of Texas. The conclusions of this research do not reflect the opinion or official position of the Texas Education Agency, the Texas Higher Education Coordinating Board, the Texas Workforce Commission, or the State of Texas. Any errors are my own.

†Department of Economics, University of Texas at Austin. Email: gschoi@utexas.edu

"We have the best interests of the child at heart, but now you are forcing us, because of our district rating, to protect ourselves instead of doing what's right for students." (*Nagle et al., 2007*)

1 Introduction

Expansion of school accountability systems, systems that evaluate schools based on student performance often with consequential sanctions or rewards, has been one of the most significant movements in U.S. education over the past few decades. By implementing a comprehensive assessment of student performance, and incentive designs built around it, policymakers aimed to encourage schools to provide the best education services. School accountability systems now have a long history in the U.S., ever since the No Child Left Behind (NCLB) Act in 2001 mandated all states to retain standardized test measures and accountability systems around them. Ample evidence suggests that such systems indeed have had positive effects on student test scores (Hanushek and Raymond, 2005; Rockoff and Turner, 2010; Dee and Jacob, 2011; Rouse et al., 2013; Chakrabarti, 2014; Reback et al., 2014) and some long-term outcomes (Deming et al., 2016; Eren and Ozturk, 2022), following significant efforts by schools (Chiang, 2009; Craig et al., 2013).

Not all students have benefited from them, however. Studies have suggested schools strategically responded to the accountability pressure, sometimes sacrificing some of their students who were less crucial to their ratings. For example, schools focused more on student-subject groups that matter most to their ratings (Reback, 2008; Neal and Schanzenbach, 2010), manipulated the testing pool through test exemptions or disciplinary actions (Jacob, 2005; Figlio, 2006; Figlio and Getzler, 2006; Cilliers et al., 2021), or even in some cases, engaged in blatant cheating (Jacob and Levitt, 2003). Such actions are primarily undesirable in the sense that they often exclude disadvantaged students[1] who most need help from schools. As a result, these reports led to several decades of fierce debates on the effects of school accountability on student achievements. Yet, no study has examined how such strategic school responses affect students in the long term.

This paper fills this gap by providing causal evidence on both schools' strategic responses to the school accountability pressure and their effects on long-term student outcomes. In the 1990s and early 2000s, the early-stage Texas school accountability system mainly relied on aggregate pass rates of particular student groups to impose rewards and penalties on schools.

[1] For example, Cullen and Reback (2006) reports increased test exemption for low-performing black and Hispanic students. Figlio and Getzler (2006) shows similar exclusions of low-income students with poor past test scores.

In 1999, Texas incorporated all test scores of special education (SE) students previously not considered in the rating calculation into its accountability system. Using comprehensive student-level panel data covering all public schools in Texas and later-life outcomes in colleges and labor markets, I examine how this accountability shock affected test scores and participation of students in special education and eventually their long-term outcomes. This strategy directly addresses various concerns about using test scores to evaluate the effects of accountability systems.

The 1999 reform in Texas targeting SE students is significantly distinguished from other accountability variations in previous literature. It generated greater accountability pressure that schools found much more challenging to cope with. The higher education costs and the poor academic performance of SE students made them "unwanted" by schools held accountable in Texas, thus creating incentives to exclude SE students from testing and education. The crudeness of the early-stage Texas school accountability system and the existence of exemption provisions allowed for SE students further reinforced such incentives. This policy context provides an excellent empirical setting to reveal how schools strategically exclude disadvantaged students and how such actions ultimately affect student outcomes. My study also has important implications for understanding the consequences of other state accountability programs since Texas served as a benchmark state for the national mandate of NCLB, being one of the earliest adopters of a full-scale school accountability system.

For my empirical strategy, I leverage the school-level variations in prior shares of students in special education – targets of the 1999 reform –, using difference-in-differences frameworks. This approach is based on the idea that the accountability pressure that a new influx of underperforming students into the rating system generates is proportional to the number of SE students within each school. Therefore, I compare individual-level student outcomes between schools that had many SE students previously not considered by the system and schools that did not, across years before and after the reform. I show how such differences in outcomes from the prior shares evolve to support the critical assumption of this framework.

I find that accountability pressure on SE students caused schools to give mass test exemptions to these students, likely to protect their accountability ratings. 10 percentage points additional prior shares of SE students led to a 7–10 percentage points drop in standardized test participation rates. A heterogeneous effect model using students' past performances reveals that this exclusion was highly selective: students with lower past test scores were removed first. On the other hand, there was no evidence of improvement in the test scores of SE students, given that they were kept being tested. I also provide suggestive evidence that district leadership could have been making these decisions rather than local schools.

By comparing the outcomes of different cohorts around 1999, I also show that accountability pressure negatively impacted students' long-term outcomes, contrary to the original intention of the 1999 reform. My estimates indicate that 10 percentage points additional SE student shares at high schools lead to 0.7 percentage points lower high school graduation rates and 1.2 percentage points lower employment rates between age 25 and 29. My results show that these negative impacts were likely due to exclusions in high schools, where low-performing SE students became less likely to take high school exit exams and more likely to drop out before reaching grade 10[2]. Heterogeneous analyses based on students' past test scores show that similar to the exclusion from testing, negative impacts on long-term outcomes were largely driven by low-performing SE students.

This paper contributes to an extensive literature on schools' strategic responses to accountability pressure. Studies reveal that schools concentrate resources on the student-subject groups that matter most to their ratings (Reback, 2008; Neal and Schanzenbach, 2010), reshape their testing pools by test exemptions (Cullen and Reback, 2006; Figlio and Getzler, 2006; Jennings and Beveridge, 2009), dropouts (Heilig and Darling-Hammond, 2008; Cilliers et al., 2021), and disciplinary actions (Figlio, 2006). I add to this literature by providing more concrete evidence using a sharp policy variation of accountability pressure, which most of the studies lacked[3]. My empirical setting is distinguished from the literature with a policy variation specifically targeting SE students, arguably the most vulnerable student group in many school accountability systems. This advantage enables me to isolate the effects on SE students without concerns about spillover effects from general education students.

Finally, I contribute to another growing literature on the long-term effects of school accountability systems. Unlike a broad literature that examines the short-run impacts of the accountability pressure, only two more recent studies analyzed this important question. Deming et al. (2016) focuses on its effects on long-term educational and labor market outcomes, and Eren and Ozturk (2022) studies the effects on criminal activity and self-sufficiency, both showing positive effects on general students. My study focuses on students at the bottom of the distribution of their academic performances, closely aligned with the results of Deming et al. (2016) that use similar data from Texas and show, unlike overall net effects, lowest-performing students experience large negative long-term impacts from accountability pressure. I complement this finding by examining the effects of a policy dedicated to low-performing students with disabilities and showing which manipulative behaviors were associated with such impacts.

[2]Before 2003, high school students in Texas had to take and pass exit-level exams in grade 10. This exam was the only standardized test for which high schools were held accountable.

[3]Richardson (2015) uses a similar identification strategy as mine, but his assumption is critically flawed. I briefly return to this issue in a later section.

2 Background

2.1 Texas Accountability System

Texas was one of the few states with a rigorous school accountability system before the well-known No Child Left Behind[4]. A basic form of standardized testing was in place in the early 1980s with exit exam requirements for high school diplomas. The Academic Excellence Indicator System (AEIS) that started in 1989 publicly reported a wide range of campus- and district-level student performance measures linked to monetary awards. This set of systems in Texas - standardized tests, school evaluations, consequent rewards, and penalties - formed one of the earliest school accountability systems in the nation and became a significant motivation for the nationwide reform of No Child Left Behind in 2001.

The early-stage Texas school accountability system in the 1990s was mainly based on several aggregate measures. Each year, schools and districts were given one of the four ratings: Exemplary, Recognized, Acceptable (Academically Acceptable), and Low-performing (Academically Unacceptable). Such ratings were determined by the three measures: pass rates of standardized tests, Texas Assessment of Academic Skills (TAAS), dropout rates, and attendance rates, as illustrated in Figure 1. To climb up to the next rating, each school or district had to satisfy all three criteria for all five student groups: All, Black, Hispanic, White, and Economically Disadvantaged. Failure in even one measure led to the next lower rating in principle[5]. Each student subgroup was considered only if it maintained a sufficiently large number of students.[6]

These accountability ratings were followed by significant consequences for schools, both explicitly and implicitly. Explicit incentives included monetary rewards[7], or exemptions from certain regulations and requirements for "Exemplary" campuses and districts. Schools with poor performances, typically classified as "Low-performing", had to conduct a hearing for residents and property owners. Further sanctions could follow if they did not show improvement afterward, which could even include school or district closure and consolidation[8].

[4]Carnoy and Loeb (2002) measures intensities of state accountability systems in 2000 with an index between 1 and 5. Texas was one of the four states with an intensity index of 5, along with New Jersey, New York, and North Carolina. Texas had both the earliest and most comprehensive system among those four. Meanwhile, most states- 32 states - had much weaker systems with an intensity of 2 or below. Two states - Iowa and Nebraska - did not even have any accountability system before NCLB.

[5]Failed requirements could be waived under the "Required Improvement" rule, which was applied to schools that showed significant improvements. However, its use was minimal due to strict eligibility conditions.

[6]For example, evaluation was waived for a student-subject group with less than 30 students.

[7]The Texas Successful Schools Award System and the Principal Performance Incentive Program provided the awards in the 1990s and 2000s. For example, TEA notified the distribution of TSSAS funds up to $500,000 in 2003, targeting schools that exhibited significant gains in student performances.

[8]Deming et al. (2016) shows that the effect of accountability pressure in Texas was concentrated at the

Implicit incentives involved impacts on the school's reputation, as all ratings were publicly available. Multiple studies indicate such publicly disclosed ratings could affect future school enrollment, closure, or even local property values (Figlio and Lucas, 2004; Nunes et al., 2015; Andrabi et al., 2017).

Some lauded the Texas system for demonstrating substantial improvements in student achievement. Texas showed impressive gains in TAAS pass rates across all subjects, ranging from 8 to 20 percentage points between 1994 and 1998, after the full-scale implementation of the school accountability system. The racial gap between black and white students narrowed from 38 to 30 percentage points, and dropout rates plunged from 2.8 to 1.6 percentage points during the same period(Haney, 2000)[9]. Supporters called this drastic improvement the "Texas Miracle", the term President Bush often used during his presidential campaign in 2000. Shortly after, the "Texas Miracle" motivated the legislation of No Child Left Behind, where the Texas-style school accountability system was mandated nationwide.

However, plenty of evidence suggested that such improvements were not "Miracle" but a myth resulting from schools' strategic behaviors. Unlike steep gains in high-stakes TAAS pass rates, students showed much less improvements with increasing racial gaps in NAEP, a low-stake nationwide assessment not considered by the accountability system (Haney, 2000; Klein et al., 2000). Schools were very likely to have manipulated their testing population by grade retention, making unreported dropouts, and excluding underperforming students by special education exemptions (Haney, 2000; Fielding, 2004; Heilig and Darling-Hammond, 2008). Interviews with veteran teachers also indicate such accomplishments were achieved through intensive "teaching to the test", rather than real gains in student learning (Ramzinski, 2019).

2.2 1999 Incorporation of Special Education Students

Special education services, mandated by the Individuals with Disabilities Education Act (IDEA), provided disabled students with additional resources necessary due to their conditions in schools. Qualified Texas students receive individually tailored education following the Individual Education Program (IEP) designed by their Admission, Review, and Dismissal (ARD) committees. An ARD committee consists of a student's parents and school personnel involved with the student. It controls the entire process around special education including initial referral, curriculum setup, giving accommodations or exemptions for testing, and managing requirements for grade promotion or graduation. Thus schools had

lowest margin, schools that could have been rated "Low Performing". Pressure to achieve higher ratings had no significant effect.

[9] Appendix Figure A.1 illustrates annual trends of TAAS math pass and dropout rates. Both measures show significant improvements after 1994.

considerable discretion over initial referrals, education, and evaluation of SE students.

Before 1999, the Texas accountability system did not consider TAAS scores of SE students. This was based on the idea that SE students could not be evaluated appropriately as general education students taking TAAS. TEA also intended to incentivize districts to actively include students with potential disabilities in the statewide assessment program[10]. On the other hand, this exacerbated the deliberate over-identification of special education students to protect school ratings by strategically placing low-performing students into special education (Nagle et al., 2006). TEA was aware of this risk as well, with rising shares of special education students and TAAS participation by them over the years.

Thus, TEA forced all TAAS scores of special education to be counted by the accountability system from 1999, under heavy pressure from disability advocacy groups. This created a significant new accountability pressure on Texas schools and districts. In 1998, around 14 percent of all students were in special education, most of them harshly underperforming compared to general education students. Test exemption rates skyrocketed in 1999, (Figure 2) and the share of SE students started to trend downward (Figure 3). While there had been no established causal relationship between increased test exemptions and the 1999 reform (Linton, 2000), teachers and district administrators found it highly likely that the increased accountability pressure caused the mass exemptions (Nagle et al., 2006).

Recognizing this problem, TEA banned test exemptions due to special education status in 2001[11]. However, this had no real effect on schools' behaviors. A new test, State-Developed Alternative Assessment (SDAA), developed for special education students, was introduced simultaneously to be an alternative to TAAS. Because the accountability system did not include the SDAA measures[12], the only real change to schools after 2001 was that they could put low-performing SE students into a low-stake exam, SDAA, instead of giving full test exemptions. Therefore, the measure did not affect schools' incentives around special education students and their consequent TAAS participation, as illustrated in Figure 2[13].

[10] Refer to Policy Research Report 9 of TEA (1997).

[11] In reality, a small portion of exemptions were still in place. 7.8 and 8.1 percent of students in special education got ARD test exemptions in 2001 and 2002, respectively.

[12] SDAA was later updated and included in the accountability system, starting in 2004. However, the rigor of SDAA-based accountability was questionable because of its low passing thresholds (Lewis, 2008) determined by ARDs.

[13] This is contrary to what Richardson (2015) assumes for his core identification strategy. His study assumes this prohibition effectively made SE students to be included in the accountability subset, which was not true.

3 Data

This study uses multiple individual-level administrative datasets provided by the Texas Education Research Center (ERC). I use K-12 educational records from TEA that cover all students in Texas public schools[14]. These records cover all aspects of educational information of each student, including enrollment, attendance, graduation, disciplinary actions, dropouts, and socio-demographics like age, gender, ethnicity, special education status, and free lunch eligibility. The data also provide detailed institutional information on schools and districts, including their types, geographics, and budgets. In this study, I use records of special education students enrolled between 1994, the earliest enrollment year in the data, and 2002, a year before the major overhaul in testing and the accountability system in Texas.

For short-run academic outcomes of students, I focus on high-stakes standardized exam (TAAS) participation and performances of special education students. I link the individual-level student enrollment records to the TAAS test records in $3 \sim 8$th grade and 10th grade, at which TAAS is administered. The TAAS subjects I use are reading and mathematics, as they are the two subjects that are tested across all grade levels. I convert the raw TAAS scores into normalized scores with a mean of zero and a standard deviation of one within each grade year to estimate the effect on test performances. I count only the first take of TAAS each year and subject for test retakes.

I do not include the earliest two years, 1994 and 1995, in my analysis because of a change in testing policy in 1996. Before 1996, schools could freely give TAAS exemptions to students without any alternative assessment to replace TAAS. TEA reverted this rule in 1996, mandating the provision of alternative assessments when students are exempt from TAAS. Since schools and ARD committees had to develop individualized assessments matched to each student's IEP, the change increased the costs of giving test exemptions. This naturally raised overall TAAS participation rates, proportional to the number of special education students and exemptions they had received. This could theoretically threaten my identification strategy based on the same variation[15]. Therefore, I look into the 7 years between 1996 and 2002 for my short-run analysis. Figure 4 illustrates this setup, where my analysis focuses on the last two periods.

I construct long-term student outcomes on three distinct datasets. First, I use TEA high school graduation records to observe whether each student successfully graduated from the

[14] This includes non-traditional public institutions such as charter school districts, alternative education campuses, and juvenile detention centers. The data do not cover private institutions in Texas.

[15] Event study results including the first two years show identical treatment effects in 1999. Testing rate results are consistent with my prediction and show steep increases in testing rates. However, it also shows this shock is clearly distinguished from the 1999 reform. Students were "indiscriminately" added to the testing pool, while the 1999 reform shows clear cream-skimming behaviors.

high school. Second, the TEA student record is linked to the Texas Higher Education Coordinating Board (THECB) data which include the students in higher education institutions within Texas. I build a measure of college degree acquisition by checking if a student has a two-year or four-year college degree after the high school graduation age. Third, I use administrative data the from Texas Workforce Commission (TWC) that contain all employees subject to the Unemployment Insurance (UI) benefits in Texas to build outcomes on wage and employment[16]. This study uses each student's wage and employment at the ages between 25 and 29.

Lastly, I use Academic Excellence Indicator System (AEIS) reports which are school reports published by the TEA. AEIS reports contain various school- or district-level information such as ratings, student demographics, and high-stakes test performances. While most of the data are also available through the TEA data at more micro levels, I use the official accountability ratings and indicators used to determine the actual accountability ratings to supplement the primary datasets I described above. Rating information lists accountability ratings including the "unrated" status of all schools and districts in Texas. The accountability indicators include aggregate TAAS pass rates, attendance rates, dropout rates, and the number of each student group considered by the accountability system.

Table 1 summarizes average individual characteristics, educational and labor market outcomes of general (columns 1 and 2) and SE students (columns 3 and 4). I show separate statistics for pre-period (columns 1 and 3) and post-periods (columns 2 and 4) around the 1999 reform. Students in special education were more likely to be male and have FRL benefits but showed little differences in racial distributions and limited English proficiency status. Two groups show stark differences in educational outcomes. SE students were less likely to take TAAS exams, with far lower TAAS scores if tested. They were also less likely to graduate high schools, attend and complete colleges. These negative traits continued in future labor markets, showing lower annual earnings and employment rates. One notable change between pre- and post-periods is TAAS test rates and scores of special education students. Students in special education experienced a sharp decrease in test rates and an increase in test scores with no changes in outcomes of general education students[17].

[16] Similar to TEA and THECB data, I am unable to track individuals who are employed outside of Texas. Such individuals are considered unemployed in this study.

[17] General education students show small decreases in normalized scores, but they are highly likely due to spillover effects from a rise in average scores of SE students. Note that SE students occupied significant portions of the overall student population.

4 Empirical Strategy

To estimate the effect of accountability pressure on students, this paper uses a difference-in-differences framework to compare changes in individual outcomes between schools with different accountability pressure imposed by the 1999 reform. In this section, I explain how I construct the measure of the imposed accountability pressure. Then, I describe sample construction processes and regression frameworks for both short- and long-run analyses.

4.1 Short-run Analysis

First, I estimate the causal relationship between the accountability pressure introduced by the 1999 reform and the short-run outcomes of SE students. Since there was no variation in implementation timing across schools in Texas, I use a cross-school variation in treatment intensity for this study. This strategy is based on the fact that the expected drop in aggregate pass rates after the reform was proportional to the initial number of SE students within each school. More SE students in a school meant more poor test scores getting into its accountability pool in 1999, which led to larger accountability pressure on schools. This approach is similar to that of Ballis and Heath (2021), where they use pre-policy SE shares to examine the effects of the Texas SE share cap in 2005.

While the SE student shares could serve as a useful proxy for accountability pressure by the reform, they are often an endogenous variable over which schools have significant control. For example, incorporating scores of the SE group could have reduced a school incentive to refer a student to special education. This is particularly likely as schools often placed low-performing students into special education before the reform[18]. Though moderate, the trend reversal of SE shares after 1999 in Figure 3 further supports this hypothesis. This is problematic not only because of the endogeneity issue but also because this could lead to different student compositions in special education after the reform, which could bias my estimates.

To circumvent this issue, I construct a balanced sample based on SE designation before the reform, using only SE student shares before 1999. More specifically, my short-run analysis sample is restricted to students who had been in special education before 1999 and stayed in Texas public schools for all 6 years across their 3–8th grades when they are required to take non-exit level TAAS. Furthermore, I use the school-level SE shares averaged between 1996 and 1998, before the 1999 reform was implemented. The distribution of the constructed

[18] Fielding (2004) conducts surveys on Texas educational diagnosticians, where over 78% responded that over half of initial referrals were primarily driven by poor TAAS performances. Teachers and school administrators explicitly ordered for qualification to SE services, often in "inappropriate ways".

measure is displayed in Figure 5. Students whose schools were not rated by the accountability system are excluded from the analysis sample. I test less restrictive specifications later in the robustness check section.

I compare changes in TAAS participation rates and scores of SE students between schools with different prior shares of SE students around 1999. This difference-in-differences framework examines whether the sharp accountability shock in 1999 incurred manipulative behaviors of excluding low-performing students from the testing pool as the literature suggests. I also look at test score changes to see if SE students benefited from the accountability pressure like many studies have reported for general students. I focus on reading math tests because they were the two subjects that were administered every grade under accountability.

I use this sample to run the following difference-in-differences regression analysis:

$$Y_{ist} = \alpha + \beta share_s^{pre} \times Post_t + f(X_{ist}) + \gamma_s + \tau_t + \epsilon_{ist} \tag{1}$$

where Y_{ist} is a short-run outcome of student i at year t, $share_s^{pre}$ is a school-level, time-invariant prior share of SE students of school s at year t and $Post_t$ is an indicator variable that turns on if $t \geq 1999$. $f(X_{ist})$ includes various student-level and school-level controls such as race, free lunch status, English proficiency status, county median income, and unemployment rates. I also include a school fixed effect γ_s and a year fixed effect τ_t. Standard errors are clustered at the school level. The coefficient of interest, β, captures the difference in changes in outcome Y_{ist} between a school with no SE students ($share = 0$) and a school with only SE students ($share = 1$) after the 1999 reform. During the rest of this paper, I divide this by 10 and interpret this as a treatment effect per 10 percent additional share of SE students within a school.

The validity of equation (1) relies on a critical assumption that without the 1999 reform, the trends of outcomes would have evolved similarly across different prior levels of special education student shares, conditional on other control variables and fixed effects. While this assumption is innately difficult to test directly, I use the following event study framework to support my previous specification:

$$Y_{ist} = \alpha + \sum_{k \neq 1998} \beta_k Share_s^{pre} \times Year_k(t) + f(X_{ist}) + \gamma_s + \tau_t + \epsilon_{ist} \tag{2}$$

where a series of coefficients β_k captures changes in outcome relative to the reference year 1998, across different levels of prior SE student shares. All other regression components are identical to equation 1.

Another interesting question of this study is whether increased test exemption due to the new accountability pressure, if it existed, was associated with "cream-skimming", selectively dropping low-performing students from their testing pool to inflate their aggregate pass rates. To answer this question, I estimate a heterogeneous effect model interacted with past test scores of each student:

$$\begin{aligned} Y_{ist} = \alpha + \beta_1 PrevScore_{it-1} \times Share_s^{pre} \times Post_t + \beta_2 PrevScore_{it-1} \times Post_t \\ + \beta_3 Share_s^{pre} \times Post_t + \theta PrevScore_{it-1} + f(X_{ist}) + \gamma_s + \tau_t + \epsilon_{ist} \end{aligned} \tag{3}$$

In equation 3, $PrevScore_{it-1}$ indicates average past normalized TAAS scores of student i up to year $t-1$. To address the endogeneity issue past 1999, I stop updating $PrevScore_{it-1}$ after 1999[19]. The coefficient of interest, β_1, thus estimates the additional treatment effect of a student with a 1 standard deviation (sd) higher past score compared to a student with a lower past test score. Similar to the base model, I supplement this with a corresponding event study model:

$$\begin{aligned} Y_{ist} = \alpha + \sum_{k \neq 1998} \beta_{1k} PrevScore_{ik-1} \times Share_s^{pre} \times Year_k(t) \\ + \sum_{k \neq 1998} \beta_{2k} PrevScore_{ik-1} \times Year_k(t) + \sum_{k \neq 1998} \beta_{3k} Share_s^{pre} \times Year_k(t) \\ + \theta PrevScore_{it-1} + f(X_{ist}) + \gamma_s + \tau_t + \epsilon_{ist} \end{aligned} \tag{4}$$

Likewise, β_{1t} shows the trend of additional treatment effects associated with past test scores better by 1 sd.

4.2 Long-run Analysis

My long-run analysis focuses on the impacts of accountability pressure on outcomes such as high school and college graduation, and earnings in adulthood. I cannot use the same sample as short-run analysis here because such long-term events happen only once at maximum to each individual. Instead, I exploit differential treatment across 9th-grade cohorts around 1999. Recall that Texas students take high-stakes exams only up to grade 10. A 9th grader in 1998 was expected to enter grade 10 in 1999, having 1 year of treatment by the 1999 reform. On the other hand, a one in 1997 would be entering grade 10 in 1998, not directly affected by the reform.

Therefore, I compare outcomes of 9th-grade cohorts who were in special education in their 8th grades in years 1994 $\sim$ 2002, using initial shares of SE students of high schools

[19] For example, a 8th-grade student in 2001 will have an average of 3–5th-grade TAAS scores as $PrevScore$. Students who do not have eligible test history are excluded from the sample.

they were enrolled in as main variation[20]. The analysis sample consists of 293,739 9th-grade students in special education from 9 years of cohorts. Like the previous section, the sample does not include high schools not rated by the accountability system.

I estimate the following regression for the long-run analysis:

$$Y_{ist} = \alpha + \beta Share_t^{pre} \times Expose_t + f(X_{ist}) + \gamma_s + \tau_t + \epsilon_{ist} \tag{5}$$

Here, Y_{ist} is a long-run outcome of a 9th-grade student who was i in a high school s in year t. $Expose_t$ equals 1 if the 9th-grade cohort t was expected to spend a positive number of years under the 1999 reform by 10th grade. I include a school fixed effect γ_s and cohort fixed effect τ_t to control for cohort-invariant school and school-invariant cohort characteristics respectively. $f(X_{ist})$ includes student- and school-level control variables similar to the short-run analysis. The coefficient of interest, β, estimates the effect of exposure to the 1999 reform on student long-run outcomes, compared between schools with no SE students and schools with only SE students. To examine the validity of this specification, I run the following event study regression as well, similar to the previous short-run case:

$$Y_{ist} = \alpha + \sum_{t \neq 1998} \beta_t Share_s^{pre} \times Expose_s + f(X_{ist}) + \gamma_s + \tau_t + \epsilon_{ist} \tag{6}$$

where β_t illustrates the changes in differences of long-term outcomes relative to the reference cohort, $t = 1998$.

5 Main Results

5.1 Short-run Analysis

5.1.1 Test Score

Figure 6.(a) and (b) illustrate raw data trends of TAAS scores of Texas SE students from 1996 to 2002. Each plot represents an average outcome separately by two groups with different treatment intensities. I define the high-share group as students enrolled in the top 25 percent of schools in terms of initial SE student shares and the low-share group as those in the bottom 25 percent. The TAAS scores are all normalized to have a mean of 0 with a standard deviation of 1. Both groups' TAAS scores trended similarly before 1999 but

[20]Note that I allow students in the sample to be referred to special education after 1999, unlike Ballis and Heath (2021), due to lack of pre-period cohorts. I test if this becomes a problem in the later robustness check section. Results are all robust to the sensitivity analysis.

exhibited very different responses in 1999 following the implementation of the accountability reform[21]. High-share schools showed much more significant improvements in the average test scores of SE students. This pattern was identical for both reading and math tests. Such clearly heterogeneous paths of test scores provide evidence that initial shares of SE students successfully capture variation in the treatment intensities of the 1999 reform, supporting the difference-in-differences framework of this paper[22].

Panels (a) and (b) of Figure 7 plot the regression estimates based on equation 2. I plot the estimated coefficients and 95 percent confidence intervals of $Share_s^{pre} \times Year_k(t)$, which are β_ks. Event study results show a pattern analogous to the previous raw data plots of Figure 6. Both panels (a) and (b) indicate SE students experienced drastic improvements in math and reading test scores after the 1999 reform which started to hold schools accountable for test scores of SE students. Such improvements are commonly observed regardless of the grades that SE students were in when the reform was introduced (see Appendix Figure A.2). While there are some signs of significant pre-trends before 1999 that potentially violate the parallel trends assumption, their sizes are relatively small and show the same direction as the estimated treatment effects.

Columns 1 and 3 of Table 2 report estimates of the same outcome after pooling all post-treatment years for the average effects of the 1999 reform, based on equation 1. They indicate there were approximately 0.1 standard deviations gain in reading and math test scores per 10 percent share of SE students within schools. These estimates imply schools in the top quartile in terms of the SE shares achieved around 0.057 standard deviations more gains in both test scores compared to schools in the bottom quartile. While these impacts on test scores were all large and significant, the fact that the gains were observed right after the 1999 reform with little further improvements afterward makes it difficult to conclude they were "real" gains. Thus, I examine the potential existence of compositional effects from changes in SE test exemptions in the next subsection.

5.1.2 Test Participation

Figure 8 displays raw data trends of average test participation rates of SE students from 1996 to 2002, separately reported between high-share and low-share schools. Interestingly, trends of test rates follow a pattern exactly opposite to that of test scores. SE students

[21]It is notable that there are considerable differences in average test scores before 1999. Appendix Table A1 describes summary statistics of both high- and low-share schools. Overall average statistics show that such differences between the two groups are not confined to SE students. This implies that students in low-share schools generally perform better than those in high-share schools.

[22]This assumes that there are no other confounding factors that are correlated with both SE share measures and outcome variables. Extensive investigation on the Texas accountability and education systems finds no such confounder in this period.

became much less likely to get tested after the 1999 reform, where students in high-share schools experienced greater decreases in test rates. Both reading and math tests showed similar declines in test participation of SE students.

Figure 9 presents event study results from equation 2 on test participation of SE students. Estimated coefficients follow similar patterns shown in the previous raw data plots. Increased accountability pressure led to significantly lower test rates for SE students, with effect sizes getting larger over the years. These sharp decreases in test rates in 1999 were again significant regardless of students' grade cohorts within the analysis sample, though older cohorts showed larger impacts (see Appendix Figure A.3). The impacts were identical between reading and math tests. Columns (1), (3), and (5) of Table 3 show corresponding point estimates from 1. They indicate that increased accountability pressure led to 7.34 and 7.01 percentage points lower participation rates in TAAS reading and math tests respectively per 10 percentage points additional SE shares. These were significant exclusions from testing, which were 12–13 percent drops from the pre-policy means.

Given that schools had considerable discretion over exempting SE students from testing, this drastic fall in the test rate was almost certainly due to increased test exemptions for SE students after the 1999 reform[23]. Still, it is challenging to verify whether schools intended these mass exemptions to protect their accountability ratings because such decision-making was usually carried out implicitly or in secret. One way to indirectly test the claim is to find out which students are getting excluded first. If schools were indeed trying to inflate ratings by giving test exemptions to SE students, it would be an optimal strategy for them to exclude low-performing students, while retaining high-performing ones in their testing pools.

Figure 10 illustrates estimates on test participation of SE students from equation 4 and provides evidence that schools were strictly engaging in such a cream-skimming behavior. Sharp jumps in estimated coefficients in 1999 imply that SE students with higher past test scores were more likely to get tested. In other words, SE students with lower past test scores became less likely to get tested after the reform[24]. Similar to the previous outcomes, these effects were identical across test subjects and different grade cohorts in my sample (see Appendix Figure A.4). Columns (2), (4), and (6) of Figure 3 report corresponding pooled estimates from equation 3. Per 10 percentage points in the initial SE shares in schools, having

[23]State-level TAAS participation statistics provided by AEIS reports strongly support this claim (https://rptsvr1.tea.texas.gov/perfreport/aeis/99/part/state.html). Though there is no separate information for SE students, the number of ARD exemptions, which were dedicated to SE students spiked in 1999. Other channels such as absence and LEP exemptions remained stable in the same period.

[24]Since the normalized test scores of SE students were mostly negative, most SE students were estimated to experience decreases in test rates unless they scored top 10 percent among tested SE students.

1 standard deviation lower past scores made the student 4.44 and 3.77 percentage points less likely to get tested after the reform. This suggests that schools selectively excluded students they thought were less likely to pass future tests. For example, a back-of-envelope calculation implies that a student in the bottom quartile became 36 percent less likely to get tested. In comparison, one in the top quartile experienced a negligible 2 percent drop in the test rate at an average school.

5.1.3 TAAS Score with Student-level Fixed Effects

Here, I return to the effects of accountability pressure on the test scores of SE students. Since previous results suggest schools actively manipulated their testing pools by excluding low-performing SE students from testing, one natural question is how much of the improvements in test scores in Figure 7 were from actual gains in student achievements, rather than from changes in compositions of tested students. For example, schools could still attempt to improve the education of some high-performing SE students if they were considered "promising" and thus kept getting tested, similar to what Neal and Schanzenbach (2010) showed. To examine this issue, I add student-level fixed effects to equation 1 and 2 so that I can estimate within-student treatment effects, addressing potential compositional effects.

Figure 11 presents the event study estimates on test scores from equation 2 with student-level fixed effects added. Surprisingly, it indicates that there was zero actual gain in both reading and math test scores, contrary to what Figure 7 suggested. Math scores even show a slight sign of deterioration in later years. Columns (2) and (4) of Table 2 report corresponding point coefficient estimates. Both estimates are statistically insignificant. These results imply schools completely relied on the exclusion of low-performing SE students to address the accountability pressure from the 1999 reform, without efforts to improve the accomplishments of SE students as the reform originally intended.

5.1.4 Heterogeneity Analyses

Urban vs. Rural Districts One interesting factor that could affect the degree of strategic responses to the accountability pressure is the competition schools face. Chakrabarti (2014) indicates that schools in more competitive environments show greater improvements in test scores. I test whether even manipulative behaviors shown in this study follow the same pattern, comparing urban and rural districts to proxy the extent of competition (Gibbons and Silva, 2008; van Maarseveen, 2021). I use district types defined by TEA by the number of population within districts[25]. Figure 12 illustrates the geographical

[25] See https://tea.texas.gov/reports-and-data/school-data/district-type-data-search/district-type-2020-21 for details of the district type definitions.

distribution of rural and urban districts by this definition, showing both narrow and wide definitions of rural districts.

Using this categorization, I examine the heterogeneous effects of the accountability pressure on test participation by district types. Figure 13 presents the event study results from equation 2. Though they share similar pre-policy trends, urban and rural districts exhibit significant heterogeneity after the reform. Exclusions of SE students were much more prevalent in urban districts than in rural districts. Furthermore, using a narrower definition of rural districts yielded starker heterogeneity between the two groups. This provides suggestive evidence that schools in more competitive neighborhoods responded much more actively to the accountability pressure.

School Performance Multiple studies have indicated that schools with poor previous performances that are more susceptible to sanctions from accountability systems are more sensitive to accountability pressure (Figlio and Rouse, 2006; Deming et al., 2016; Cilliers et al., 2021). I explore this heterogeneity by comparing the degree of exclusions in previously high-performing and low-performing schools. I define high- and low-performing schools by two measures: school-level average test scores and accountability ratings they received last year.

Figure 14 reports the estimated heterogeneity. Consistent with the literature, previously low-performing schools were more likely to exclude SE students after the 1999 reform. Schools in the top quartile in terms of average test scores made 2–3 times stronger responses compared to the schools in the bottom quartile, and this gap widened in the later years. On the other hand, heterogeneity by previous rating was not as large. These results imply that the threat of penalties from the accountability system was a significant driver of school responses.

School vs. District Incentives So far, I have presumed schools were responsible for the manipulative behaviors on SE students. This is because ARD committees consist of teachers of individual schools and parents, and thus, most decisions on the education of SE students were made at the school levels. However, district leadership could have exerted influence on its schools to inflate aggregate pass rates by excluding SE students. This is especially likely as the Texas accountability system published accountability ratings at both district and school levels, with similar rewards and sanctions at stake. Qualitative studies based on interviews with Texas teachers in the 2000s attest to the existence of such pressure from district leadership as well (Nagle et al., 2006; Ramzinski, 2019).

I look into this possibility by examining to which level of variation –school or district– the degree of exclusion was more sensitive. This analysis exploits the fact that school-level SE shares and aggregate district-level SE shares were often very different, though they were obviously correlated to each other[26]. Therefore, a school with a relatively low SE share could have been forced to exclude SE students from testing if its district had a high overall SE share. Figure 15 shows event study estimates from equation 2, using both school-level baseline SE shares and district-level SE shares as the main identifying variations. Estimated heterogeneity is clear: the degree of SE exclusion was more sensitive to district-level incentives than school-level incentives, which suggests that district leadership was more responsible for the exclusion of SE students than local schools and their teachers.

5.2 Long-run Analysis

Many studies have used test scores to examine the effect of school accountability systems on student achievements. However, my short-run analysis clearly demonstrates that schools attempted to manipulate their testing pools to inflate their aggregate test pass rates, which makes test outcomes unreliable measures of student accomplishments. Thus in this section, I focus on the effects of accountability pressure on the long-term outcomes of SE students to explore the true impacts of accountability pressure on SE students.

5.2.1 Exclusion in High Schools

First, I examine how increased accountability pressure from the 1999 reform affected the long-run outcomes of SE students. Figure 16 presents the estimates of the effects on two high school outcomes from equation 6: exit exam participation and whether they reached 10th grade when they were supposed to take the exit exam. Results are separately reported between high- and low-performing students based on their past test scores up to grade 8. Panel (a) shows that the exclusion of SE students from testing happened in high school as well. Exit-level TAAS participation rates of SE students with poor past test scores decreased by 5–10 percentage points per 10 percentage points of pre-policy SE shares. Panel (b) indicates an even more extreme form of exclusion. Low-performing 9th-grade students became 2.5 percentage points more likely to drop out before 10th grade per 10 percentage points of SE shares.

This increased dropouts, which could be interpreted as exclusion from schools themselves, were due to the uniqueness of high school exit-level exams. First, unlike other TAAS exams in 3–8th grades, the 10th-grade exit exam was the only high-stakes exam that fed into the accountability ratings of high schools. Second, SE students in Texas could legally drop out from grade 9. This meant high schools were under much heavier accountability pressure,

[26]The coefficient of correlation was $\rho = 0.67$, but more than 10 percent of schools had gaps between school-level and district-level SE shares larger than 5 percentage points. This was a significant difference as the average pre-policy SE share was around 13 percent.

with an additional means of exclusion other than giving test exemptions. It could still be hard to believe that schools intentionally encouraged dropouts just to stop low-performing SE students from getting tested when they still could exempt them from testing. However, a very similar pattern was already observed among general education students in Texas before 1999, by Haney (2000) which was a highly controversial report. He found many low-performing high school students "disappeared" from schools before 10th grade even without corresponding dropout records. He suspected that this was a result of strategic actions by schools to manipulate their exit-exam testing pools, similar to what I found in this study. I also find no sign of exclusion from official dropout data as well (see Appendix Figure A.5.

5.2.2 Educational Outcomes

Figure 17 shows the estimated effects on educational outcomes of SE students from equation 6. Panel (a), (b), and (c) present effects on high school graduation, college enrollment, and college completion respectively. Table 4 reports corresponding regression estimates from equation 5. The estimates indicate that increased accountability pressure from the 1999 reform led to a 0.73 percentage-point decrease in high school graduation or a 1.2 percent decrease compared to the pre-policy mean per 10 percentage points of initial high school SE shares. Effects on college enrollment and completion were statistically insignificant.

This small impact on college outcomes, compared to the effect on high school graduation could be because only a few SE students enter and complete college education: only 32 percent of 9th-grade SE students in 1998 eventually entered colleges, and less than 7 percent completed the program. In addition, as my previous results have suggested, the negative impact of accountability pressure was concentrated on low-performing SE students, who were even more unlikely to attend college in the future. Figure 18 and Table 5 present the same college outcomes as Figure 17 and Table 4, separated by types of institutions. As expected from the fact that most SE students enter 2-year colleges, the outcome of 2-year colleges drives a moderate decrease in college completion: a 5.9 percent decrease compared to the pre-policy mean. Both effects on high school graduation and college completion get larger as the number of years of exposure increases.

5.2.3 Labor Market Outcomes

Figure 19 and Table 6 present effects on labor market outcomes from age 25 to 29. Both indicate increased accountability pressure on SE students inflicted negative impacts on their outcomes. 10 percentage points of additional SE shares were associated with 309.4 dollars and 1.2 percentage points decreases in earning and employment in adulthood respectively.

These estimates were equivalent to 2.8 and 1.9 percent reduction compared to the pre-policy mean. These imply that impacts on the adulthood income were mostly in intensive margins, and log annual earnings conditional on employment do not show any significant impacts (see Appendix Figure A.6).

One common pattern of long-term impacts so far, especially on high school graduation and labor market outcomes is that a significant portion of the overall effect is observed right from the first year of exposure. This implies that students' exposure to the reform in 10th grade was a critical determinant of their long-term outcomes. Previously shown high school exclusions provide a good potential mechanism to support these results. An immediate decrease in exit-level exam participation and an increase in dropout rates of prospective 10th-grade cohorts could make them less likely to graduate high schools with consequential adverse effects on labor market outcomes. Cumulative effects of additional exposure to the reform are also observable through downward trends of event study estimates.

5.2.4 Heterogeneity Analyses

Next, I examine how the heterogeneous treatments in short-run exclusions in schools are reflected by the long-term outcomes I have covered so far. I test how heterogeneous treatments in short-run test rates are reflected by the long-term outcomes I have covered so far. More specifically, I focus on time-invariant variations associated with drastic differences in the degree of exclusions.

I showed the rate of exclusion SE students faced in schools critically depended on their past performances in high-stakes tests (Figure 10, 16). SE students who did well in the past tests were less likely to be exempted from tests and less likely to drop out compared to students who did not. I examine how effects on long-term outcomes differ between these two groups of SE students. I use past test scores up to grade 8 of 9th-grade SE students.

Figure 21 shows the results of the subsample analyses. It indicates that negative impacts on the long-term outcomes of SE students were mainly driven by low-performing students who had poor past test scores before entering high schools. While students with lower prior TAAS scores suffered from large adverse effects from increased accountability pressure on almost all outcomes, those with higher past scores showed null effects. This implies that deterioration in long-term outcomes witnessed before is likely to be a consequence of the exclusions in schools.

Figure 22 presents effects on long-term outcomes by district types. SE students in urban districts mainly drove adverse impacts, consistent with the prior results on short-run exclusions. Students in rural districts were mostly unaffected and even showed some positive estimates on college outcomes when I used wide definitions of rural districts. Estimates of rural districts under narrow definitions are highly imprecise due to the small size of the sample in the long-run analysis. Figure 23 illustrates heterogeneous impacts on long-term outcomes by school- and district-level incentives. It shows that effects on high school graduation and employment in adulthood were larger under district-level incentives, while the same does not hold for college outcomes.

Overall, long-run heterogeneity analyses imply that adverse impacts on SE students' long-term outcomes were highly likely due to exclusions they had experienced in schools. Results indicate student subgroups that were subject to more exclusions tend to suffer from greater deterioration in the long run as well. Though some estimates become largely inaccurate due to smaller sample sizes, they provide strong evidence that increased accountability pressure eventually ended up hurting SE students through strategic exclusions by schools.

5.3 Robustness Checks

In this section, I examine different empirical specifications to check the robustness of my short-run and long-run estimates. To test whether the restrictions in the short-run sample and variation were driving the results, I re-estimate the short-run analyses on TAAS participation rates using a full unrestricted sample and all annual school-level shares of special education students between 1994 and 2002. For the long-run results, I address the concerns about differences in student qualities between cohorts, by performing sensitivity analyses around the grades in which I pick the sample cohorts. The results are qualitatively robust to all these alternative specifications.

Columns 3 and 4 of Table 7 show estimates when I use a full sample of 3 $\sim$ 8th grade students who were ever in special education status. Compared to the base results of columns 1 and 2, the estimates are qualitatively identical, though the effect sizes tend to be smaller with the full sample. Columns 5 and 6 use all annual school-level shares of special education students, with similar results. Columns 7 and 8 use both unrestricted samples and all annual shares, showing no notable differences. These results indicate that the restrictions in sample and treatment variation did not affect my short-run findings significantly.

One potential concern in my long-run analyses is that differences between sample cohorts could be driving the estimated effects. Because schools' incentives to refer students to special education status potentially became weaker after the 1999 reform, 9th-grade students in 2000 who got special education status in 1999 could be systematically different from those who got the status in 1998. This compositional effect could have driven the immediate drop in high school graduation and employment rates shown before. If schools somehow foresaw the 1999 reform in 1997, then 8th-grade students newly referred to special education in 1997 could be more likely to be genuinely disabled, with worse expected long-term outcomes compared to non-disabled students who were strategically placed in special education. This difference would lead to a significant gap in outcomes of the 1998 and 1999 10th-grade cohorts. I test this hypothesis by changing the grades at which the sample 9th-graders were supposed to be in special education.

Figure 14 illustrates the long-run outcomes of section 5.2, but using different cohort specifications of SE status. Blue coefficients are results based on 9th-grade cohorts who were in special education at their 7th grade, while green coefficients are those who were in special education in 9th grade. Red coefficients are the same as the base estimates of section 5.2. The figure shows no difference across the three specifications, except for college enrollment, where all coefficients are nonetheless not statistically significant. This shows that the aforementioned compositional effect across cohorts does not drive my long-term effect estimates.

Next, I address the potential selection problem from using variations in past test scores. While there was very clear heterogeneity across students with different past score levels, a significant portion of SE students did not have any past test score histories. The heterogeneous effect model of equation 3 and 4 only included 53 percent of the full short-run analysis sample, and even in the long-run sample of 9th-grade SE students, only 60 percent of students had previous test records.

To expand the external validity of my heterogeneity results on student abilities, I use disability types of SE students as a proxy of their academic abilities. This approach is more extensive than using past test scores because all SE students are assigned types of disabilities to receive SE benefits. Table 8 describes the fraction and test outcomes of the four most common disability types[27]. It is easily notable that there are significant variations in academic performances across disability types. SE students with speech impairments perform far better than other SE students. Those with learning disabilities are most common but also show generally poorer outcomes.

First, I compare degrees of short-run exclusions between SE students with learning disabilities and speech impairments, as proxies of low- and high-performing students. Figure 24.(a) shows the event study estimates of the subsample analysis between the two groups. The results are consistent with the initial conjecture. Students with learning disabilities, therefore likely to be relatively low-performing experienced a much steeper decrease in test rates than those with speech impairment who tend to be high-performing.

However, one caveat here is that the number of SE students with speech impairments

[27] Though test scores are again limited measures, they still provide good information on the overall abilities of disability groups.

declines fast in higher grades. Unlike in 3–8th grades, few SE students (less than 2 percent) in 9th grade had speech impairments. I compare the effects of accountability pressure between SE students with learning disabilities and those without learning disabilities to circumvent this issue for long-run heterogeneity. Figure 24.(b) illustrates the results and again indicates that students with learning disabilities were more likely to be excluded from testing, which reflects their poorer academic abilities. Figure 25 reports corresponding estimates of effects on long-term outcomes. Results are similar to the heterogeneity analysis using past test scores, as shown in Figure 21 This implies low-performing groups faced more exclusions in schools with consequential negative impacts.

6 Conclusion

School accountability systems have been an irreversible part of contemporary education in the U.S., supported by numerous studies indicating improvements in student achievements at schools. On the other hand, some have also suggested the existence of undesirable school responses, often excluding students who actually need the most resources. Lack of data and appropriate empirical setting prevented researchers from identifying how the accountability pressure affects those underperforming students in both the short and long run. Understanding the causal impacts of accountability pressure is necessary to devise better incentive designs, ensuring they properly incentivize schools to put forth their best efforts for their students.

In this paper, I exploit an accountability reform in Texas that specifically targeted disabled students in special education and incorporated their performances into the rating measures. Using extensive student-level administrative data from Texas public schools, I examine how short- and long-run student outcomes differ after the onset of the reform, across schools with varying initial shares of special education students. I find that the reform caused schools to intentionally drop students in special education, especially those with poor past high-stakes scores. This widespread exclusion led to negative impacts on their future long-run outcomes, such as less high school graduation and employment in adulthood. From my understanding, this is the first study to estimate both short- and long-run effects of accountability pressure on disadvantaged students using a sharp identifying policy variation.

While the empirical setting of this paper is based on a very early stage of the Texas accountability system, similar problems persist even now, 24 years past 1999. Significant reforms have been made to appropriately accommodate all sorts of students to the rating system. Alternative assessments dedicated to students with special needs began to be included in the system. The overall accountability system itself became much more complicated, with certain measures to "close the gap" for disadvantaged students. However, Texas schools are still gaming the system, in much more sophisticated ways[28]. Therefore, policy-makers need to design both comprehensive and equitable accountability systems in order to provide the best incentives for their students' interests.

References

T. Andrabi, J. Das, and A. I. Khwaja. Report cards: The impact of providing school and child test scores on educational markets. *American Economic Review*, 107(6):1535–1563, 2017.

B. Ballis and K. Heath. The long-run impacts of special education. *American Economic Journal: Economic Policy*, 13(4):72–111, 2021.

M. Carnoy and S. Loeb. Does external accountability affect student outcomes? a cross-state analysis. *Educational evaluation and policy analysis*, 24(4):305–331, 2002.

R. Chakrabarti. Incentives and responses under no child left behind: Credible threats and the role of competition. *Journal of Public Economics*, 110:124–146, 2014.

H. Chiang. How accountability pressure on failing schools affects student achievement. *Journal of Public Economics*, 93(9-10):1045–1057, 2009.

J. Cilliers, I. M. Mbiti, and A. Zeitlin. Can public rankings improve school performance?: Evidence from a nationwide reform in tanzania. *Journal of Human Resources*, 56(3):655–685, 2021.

S. G. Craig, S. A. Imberman, and A. Perdue. Does it pay to get an a? school resource allocations in response to accountability ratings. *Journal of urban Economics*, 73(1):30–42, 2013.

J. B. Cullen and R. Reback. Tinkering toward accolades: School gaming under a performance accountability system. In *Improving school accountability*, volume 14, pages 1–34. Emerald Group Publishing Limited, 2006.

T. S. Dee and B. Jacob. The impact of no child left behind on student achievement. *Journal of Policy Analysis and management*, 30(3):418–446, 2011.

[28]See related article in https://www.washingtonpost.com/education/2022/10/30/new-miracle-texas-school-district/

D. J. Deming, S. Cohodes, J. Jennings, and C. Jencks. School accountability, postsecondary attainment, and earnings. *Review of Economics and Statistics*, 98(5):848–862, 2016.

O. Eren and O. Ozturk. School accountability, long! run criminal activity, and self! suffi ciency. 2022.

C. Fielding. Low performance on high-stakes test drives special education referrals: A texas survey. In *The Educational Forum*, volume 68, pages 126–132. Taylor & Francis, 2004.

D. N. Figlio. Testing, crime and punishment. *Journal of Public Economics*, 90(4-5):837–851, 2006.

D. N. Figlio and L. S. Getzler. Accountability, ability and disability: Gaming the system? In *Improving school accountability*, volume 14, pages 35–49. Emerald Group Publishing Limited, 2006.

D. N. Figlio and M. E. Lucas. What's in a grade? school report cards and the housing market. *American economic review*, 94(3):591–604, 2004.

D. N. Figlio and C. E. Rouse. Do accountability and voucher threats improve low-performing schools? *Journal of Public Economics*, 90(1-2):239–255, 2006.

S. Gibbons and O. Silva. Urban density and pupil attainment. *Journal of Urban Economics*, 63(2):631–650, 2008.

W. Haney. The myth of the texas miracle in education. *Education policy analysis archives*, 8:41–41, 2000.

E. A. Hanushek and M. E. Raymond. Does school accountability lead to improved student performance? *Journal of Policy Analysis and Management: The Journal of the Association for Public Policy Analysis and Management*, 24(2):297–327, 2005.

J. V. Heilig and L. Darling-Hammond. Accountability texas-style: The progress and learning of urban minority students in a high-stakes testing context. *Educational Evaluation and Policy Analysis*, 30(2):75–110, 2008.

B. A. Jacob. Accountability, incentives and behavior: The impact of high-stakes testing in the chicago public schools. *Journal of public Economics*, 89(5-6):761–796, 2005.

B. A. Jacob and S. D. Levitt. Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *The Quarterly Journal of Economics*, 118(3):843–877, 2003.

J. L. Jennings and A. A. Beveridge. How does test exemption affect schools' and students' academic performance? *Educational evaluation and policy analysis*, 31(2):153–175, 2009.

S. P. Klein, L. Hamilton, D. F. McCaffrey, B. Stecher, et al. What do test scores in texas tell us? *Education policy analysis archives*, 8:49–49, 2000.

D. L. Lewis. *Ending the bigotry of low expectations? No Child Left Behind and the Texas state alternative assessment for students with disabilities*. Texas State University-San Marcos, 2008.

T. H. Linton. High stakes testing in texas: An analysis of the impact of including special education students in the texas academic excellence indicator system. 2000.

K. Nagle, C. Yunker, and K. W. Malmgren. Students with disabilities and accountability reform: Challenges identified at the state and local levels. *Journal of Disability Policy Studies*, 17(1):28–39, 2006.

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

L. C. Nunes, A. B. Reis, and C. Seabra. The publication of school rankings: A step toward increased accountability? *Economics of Education Review*, 49:15–23, 2015.

L. E. Ramzinski. *Texas miracle on whose account?: An oral history of retired veteran Texas teachers on accountability*. PhD thesis, 2019.

R. Reback. Teaching to the rating: School accountability and the distribution of student achievement. *Journal of public economics*, 92(5-6):1394–1415, 2008.

R. Reback, J. Rockoff, and H. L. Schwartz. Under pressure: Job security, resource allocation, and productivity in schools under no child left behind. *American Economic Journal: Economic Policy*, 6(3):207–241, 2014.

J. Richardson. Accountability incentives and academic achievement: Distributional impacts of accountability when standards are set low. *Economics of Education Review*, 44:1–16, 2015.

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

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

TEA. Expanding the scope of the texas public school accountability system, 1997.

R. van Maarseveen. The urban–rural education gap: do cities indeed make us smarter? *Journal of Economic Geography*, 21(5):683–714, 2021.

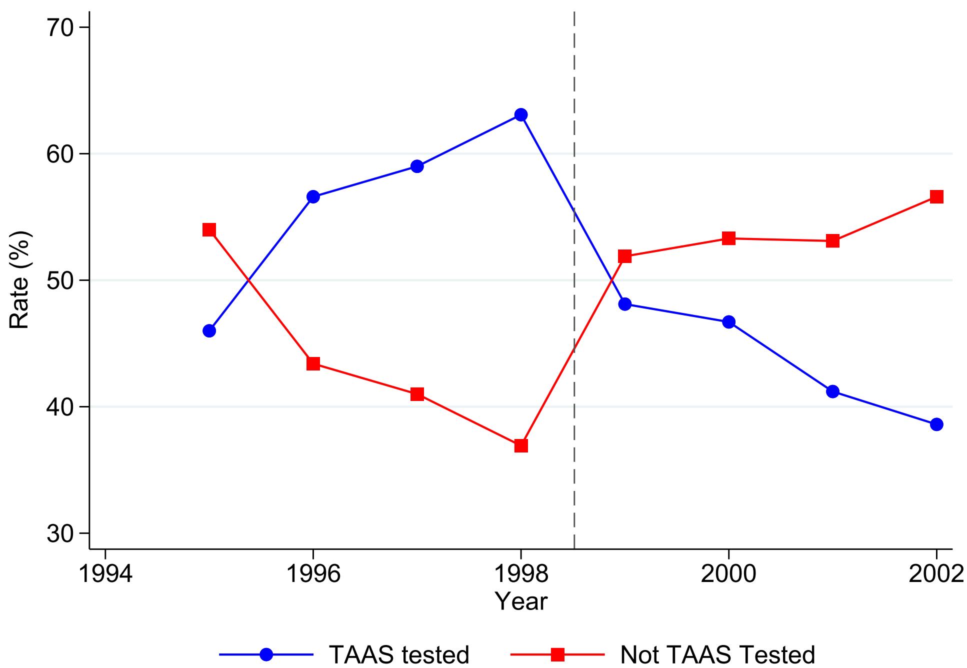
7 Figures and Tables

Figure 1: 1996 Texas School Accountability Manual

	Exemplary †	Recognized †	Academically Acceptable / Acceptable	Academically Unacceptable / Low-performing
Base Indicator Standards				
Spring '96 TAAS • Reading • Writing • Mathematics	at least 90.0% passing each subject area *(all students & each student group*)*	at least 70.0% passing each subject area *(all students & each student group*)*	at least 30.0% passing each subject area *(all students and each student group*)*	below 30.0% passing any subject area *(all students and each student group*)*
1994-95 Dropout Rate	1.0% or less *(all students and each student group*)*	3.5% or less *(all students and each student group*)*	6.0% or less *(all students and each student group*)* ‡	above 6.0% *(all students or any student group*)* ‡
1994-95 Attendance Rate	at least 94% (grades 1-12)☆	at least 94% (grades 1-12)☆	at least 94% (grades 1-12) ɸ	at least 94% (grades 1-12) ɸ

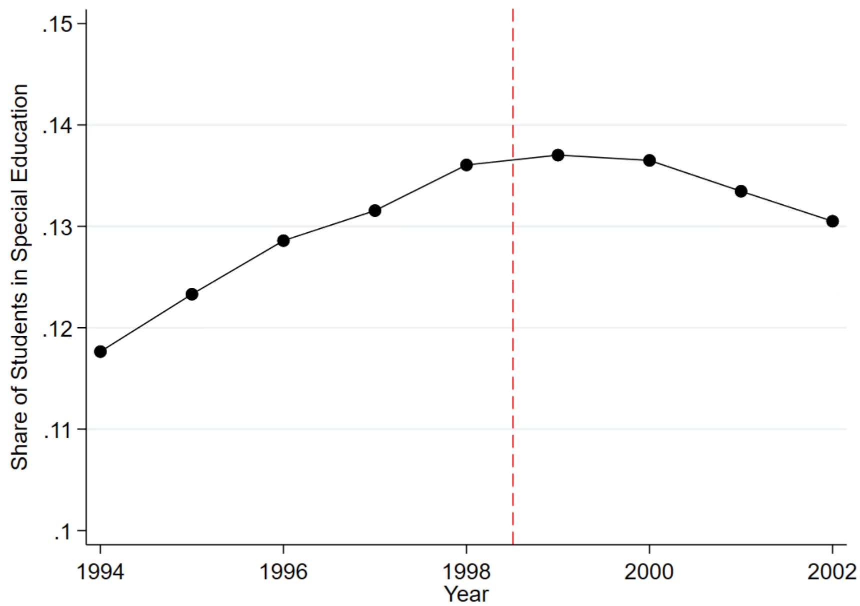
Notes: This figure shows a part of the 1996 Texas School Accountability Manual distributed to district and school personnel by TEA. The full manuals for 2004–current year are available here: https://rptsvr1.tea.texas.gov/perfreport/account/.

Figure 2: TAAS Participation Trends, 1995 $\sim$ 2002



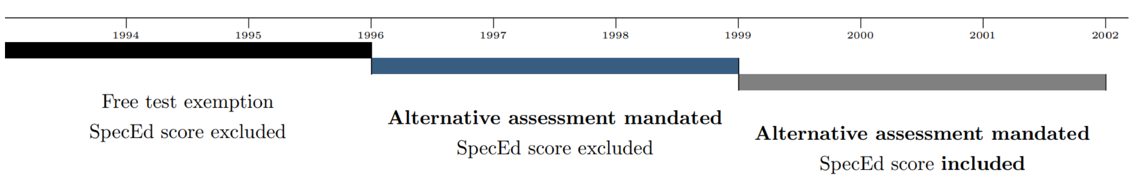
Notes: This figure illustrates trends of TAAS participation, exemption, and SDAA participation of special education students between 1995 and 2002, sourced from Academic Excellence Indicator System (AEIS) school reports. AEIS did not provide TAAS participation information separately for special education students in 1994. The blue line depicts shares of special education students who took TAAS each year, while the red line depicts shares of those who got exempted from the tests. The green line shows shares of special education students who took only SDAA after its introduction in 2001. The sum of the three does not necessarily add up to 1, due to the existence of absence.

Figure 3: Trends of Special Education Student Shares, 1994 ~ 2002



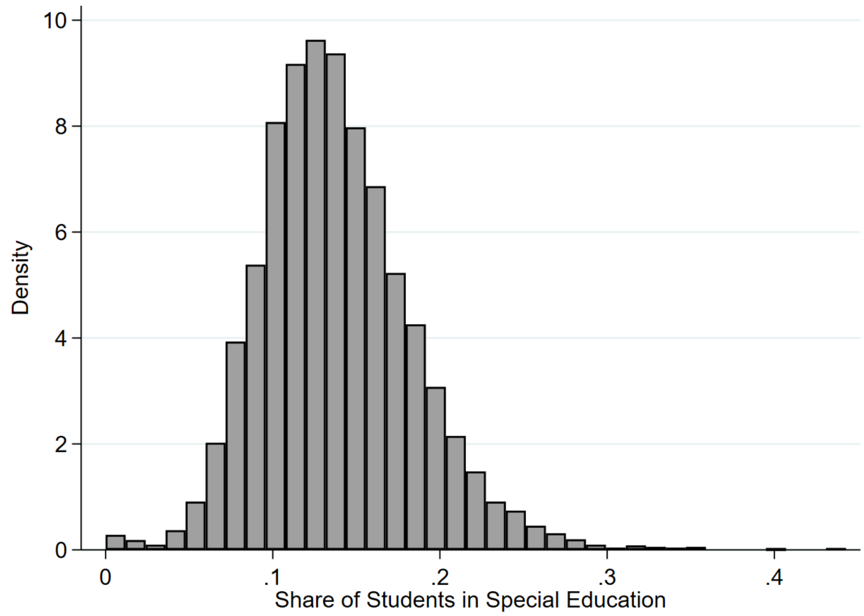
Notes: This figure presents the trend of state-level special education population shares in Texas public schools between 1994 and 2002.

Figure 4: Timeline of Special Education Assessment and Accountability in Texas



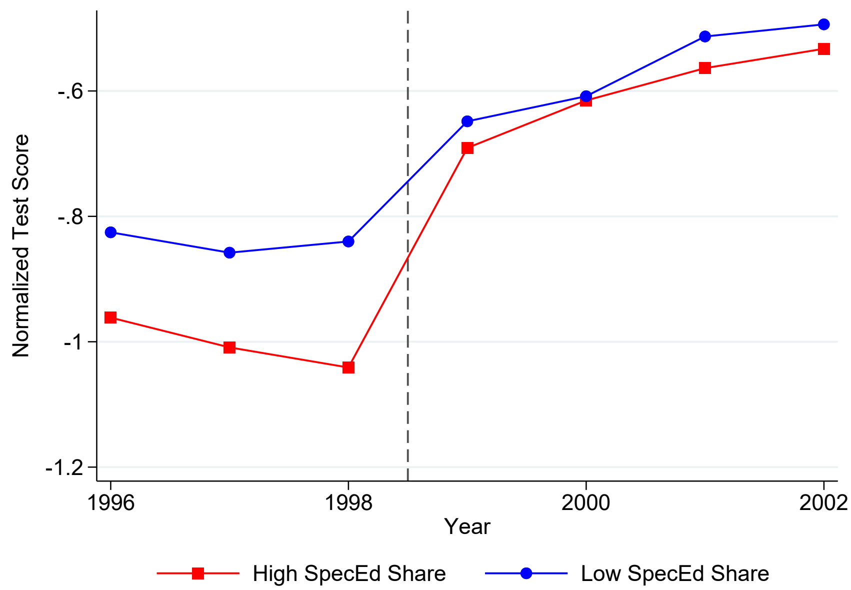
Notes: This figure depicts the two changes in Texas's SE assessment and accountability system between 1994 and 2002. This study focuses on the last two periods between 1996 and 2002, where alternative assessments for exempted students were mandated for all 7 years. The accountability system was halted in 2003 for overhaul and resumed in 2004 with new assessment (TAKS) and accountability provisions.

Figure 5: School-level Special Education Student Shares, 1998

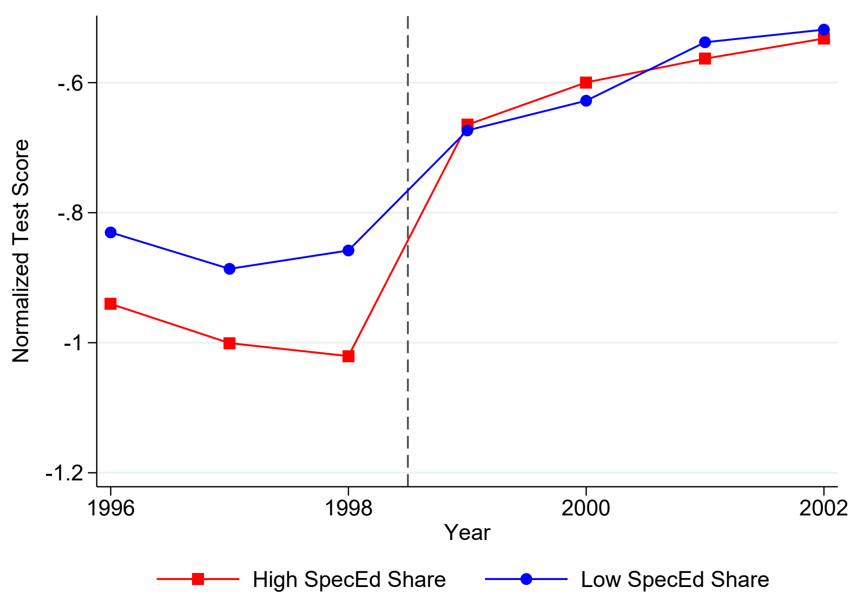


Notes: This figure shows the distribution of school-level special education student shares in 1998, a year before the 1999 reform. The mean share was 0.138 with a standard deviation of 0.046. Shares were calculated using the student population in the accountability subset (grade 3 $\sim$ 8, 10). Schools not under the accountability system or with too small numbers of students (less than 30) were excluded.

Figure 6: Raw Data Plot: High treatment groups showed stronger improvements in average SE test scores.



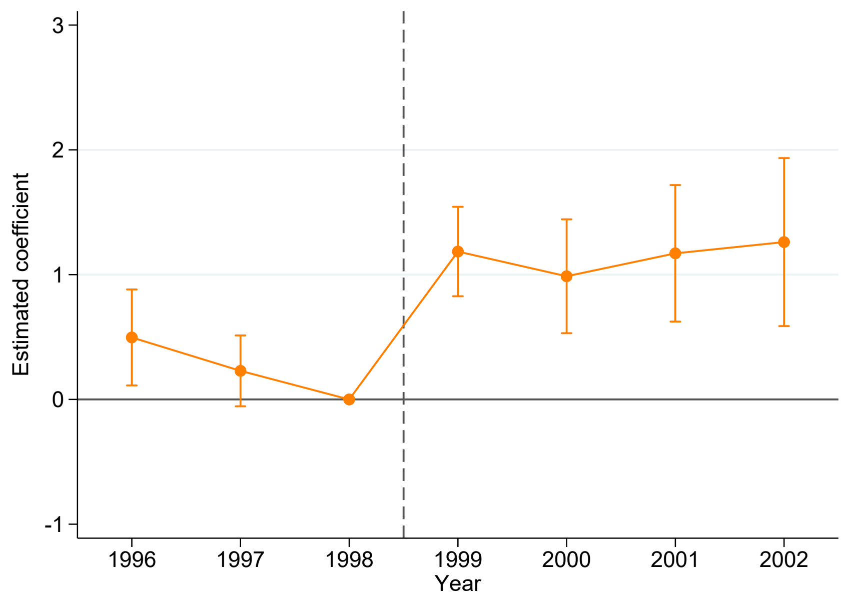
(a) Reading Score



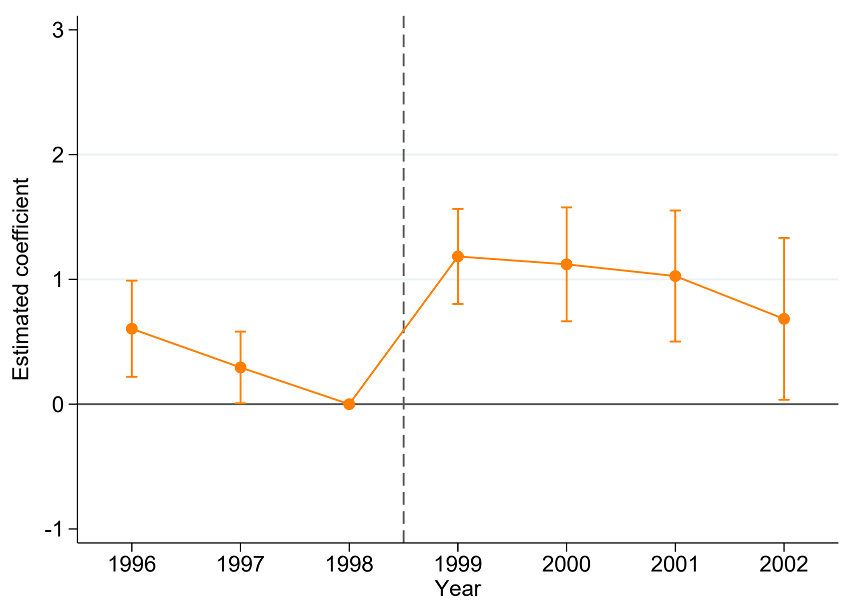
(b) Math Score

Notes: The figure illustrates raw data trends of TAAS reading and math scores of schools with high and low shares of SE students. I define "high share" schools as the top 25 percent schools in terms of SE shares and "low share" schools as the bottom 25 percent. The cutoffs of shares for the two groups were 16.5 percent and 10.7 percent respectively. Test scores are normalized to have a mean of 0 with a standard deviation of 1 within each grade level, subject, and year.

Figure 7: Event study: Accountability pressure increased average test scores of SE students.



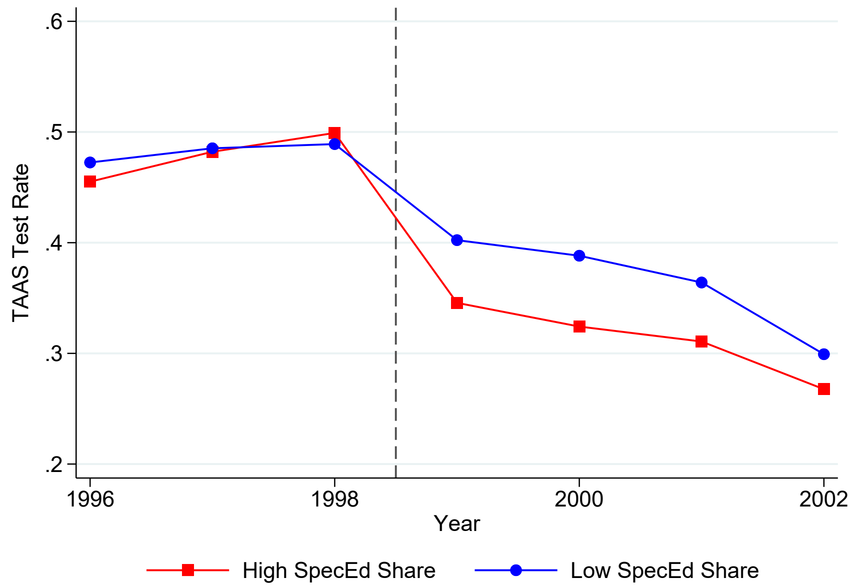
(a) Reading Score



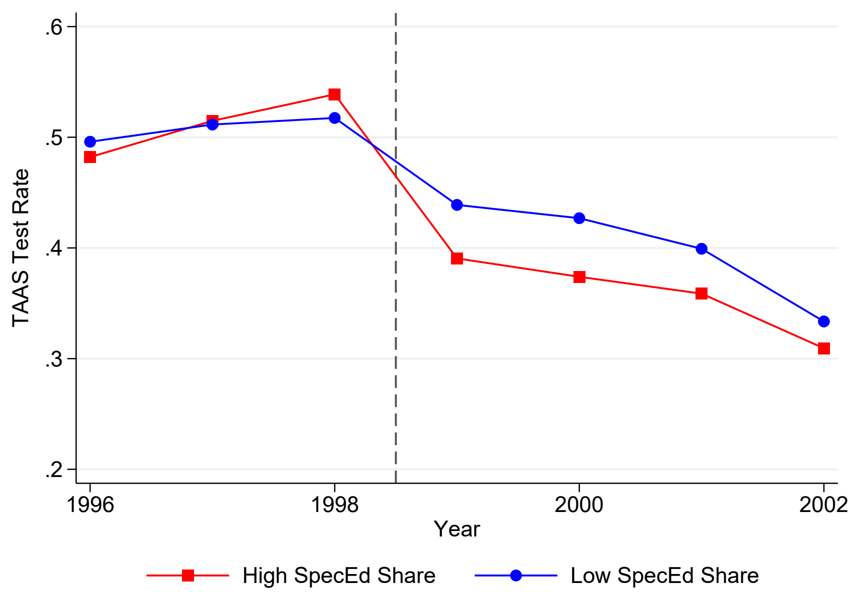
(b) Math Score

Notes: The figure plots event study estimates based on equation 2. Panels (a) and (b) show the results using the balanced sample I described. I plot the estimated coefficients of interest with corresponding 95 percent confidence intervals. Standard errors are clustered at the school levels. Test scores are normalized to have a mean of 0 with a standard deviation of 1 within each grade level, subject, and year.

Figure 8: Raw Data Plot: High treatment groups showed steeper decreases in SE test participation.



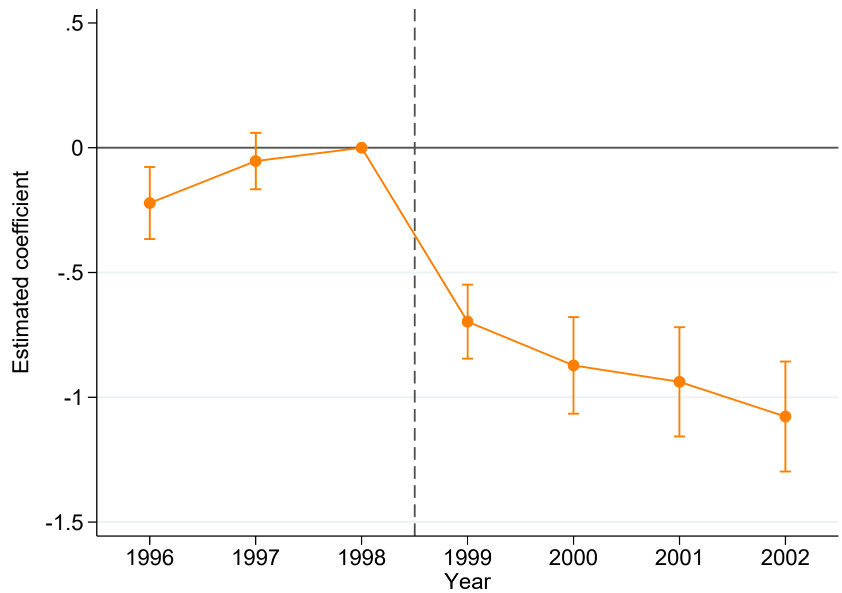
(a) Reading Test Participation



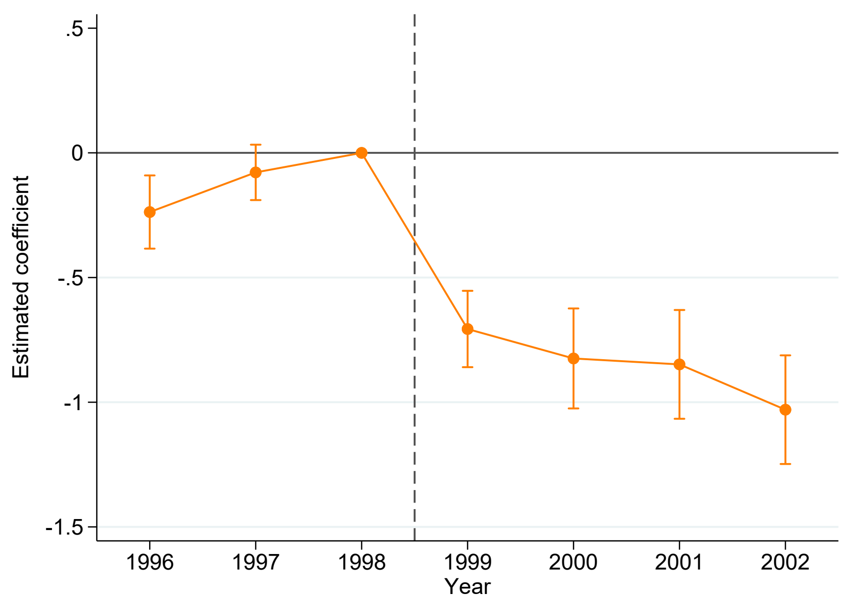
(b) Math Test Participation

Notes: The figure illustrates raw data trends of TAAS testing rates of schools with high and low shares of SE students. I define "high share" schools as the top 25 percent schools in terms of SE shares and "low share" schools as the bottom 25 percent. The cutoffs of shares for the two groups were 16.5 percent and 10.7 percent, respectively.

Figure 9: Event study: Accountability pressure decreased test participation of SE students.



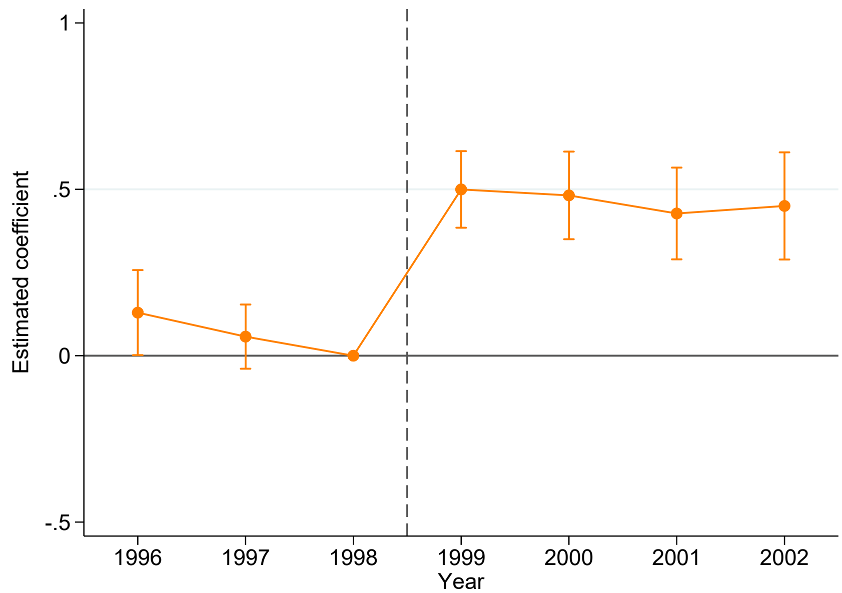
(a) Reading Test Participation



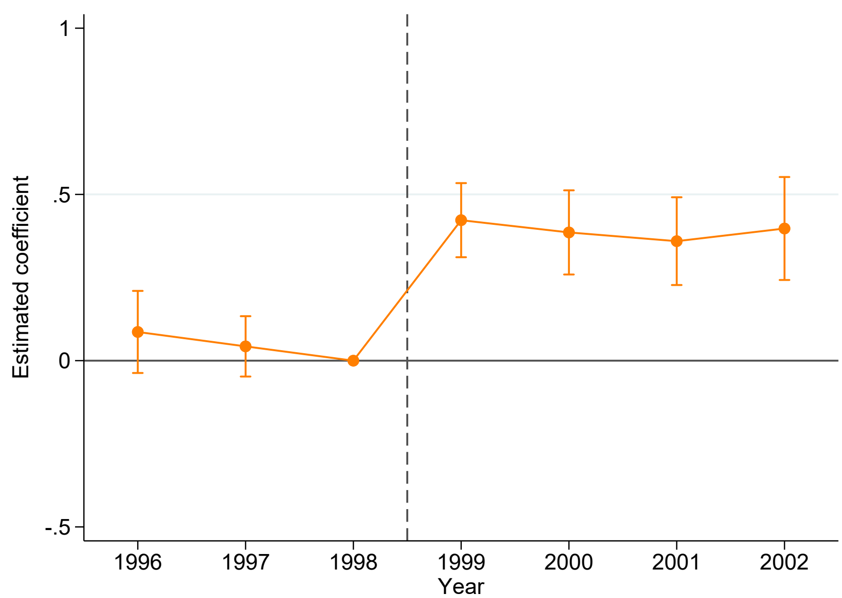
(b) Math Test Participation

Notes: The figure plots event study estimates based on equation 2. Panels (a) and (b) show the results using the balanced sample I described. I plot the estimated coefficients of interest with corresponding 95 percent confidence intervals. Standard errors are clustered at the school levels.

Figure 10: Event study: Students with lower past scores were more likely to get excluded.



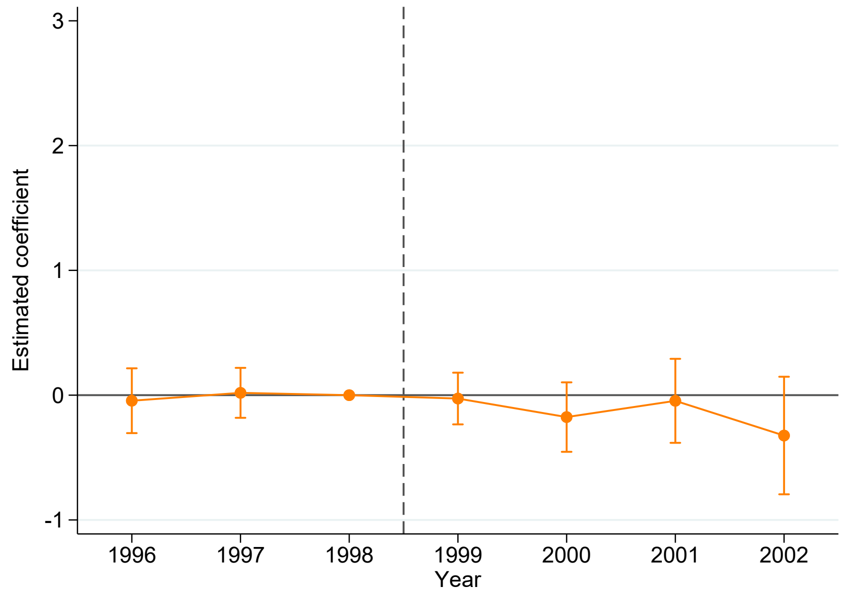
(a) Reading Test Participation



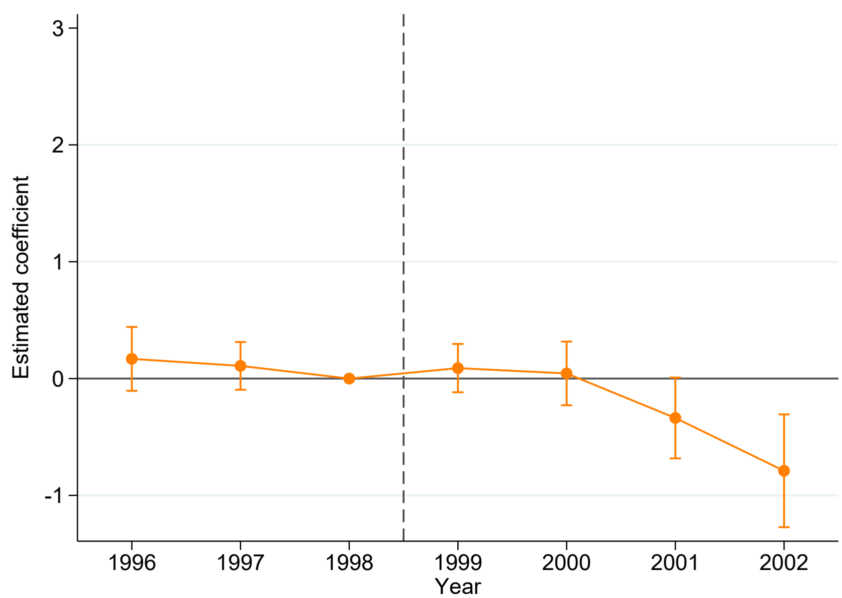
(b) Math Test Participation

Notes: The figure plots event study estimates based on equation 4. Panels (a) and (b) show the results using the balanced sample I described. I plot the estimated coefficients of interest with corresponding 95 percent confidence intervals. Standard errors are clustered at the school levels.

Figure 11: Event study: Within comparison shows no actual gain in student test scores.

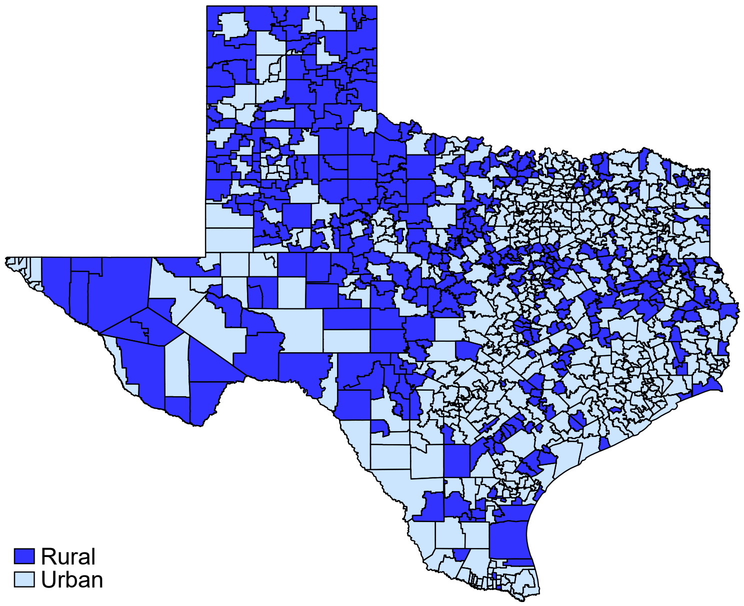


(a) Reading Score

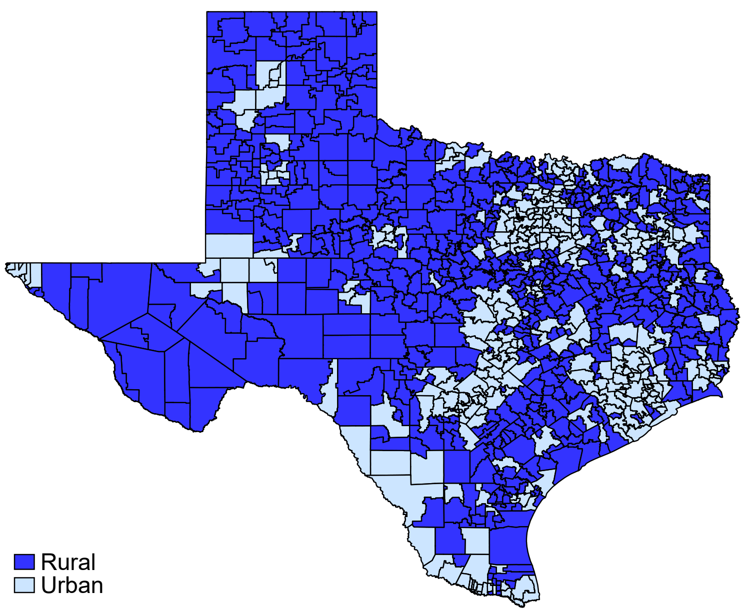


(b) Math Score

Notes: The figure plots event study estimates using a model that adds student-level fixed effects to equation 2. Panels (a) and (b) show the results using the balanced sample I described. I plot the estimated coefficients of interest with corresponding 95 percent confidence intervals. Standard errors are clustered at the school levels. Test scores are normalized to have a mean of 0 with a standard deviation of 1 within each grade level, subject, and year.



(a) Narrow Definition of Rural Districts

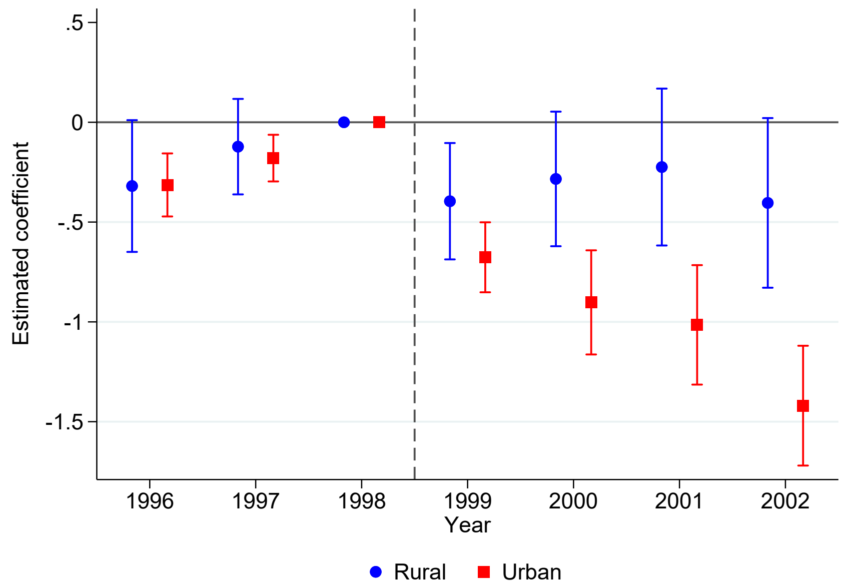


(b) Wide Definition of Rural Districts

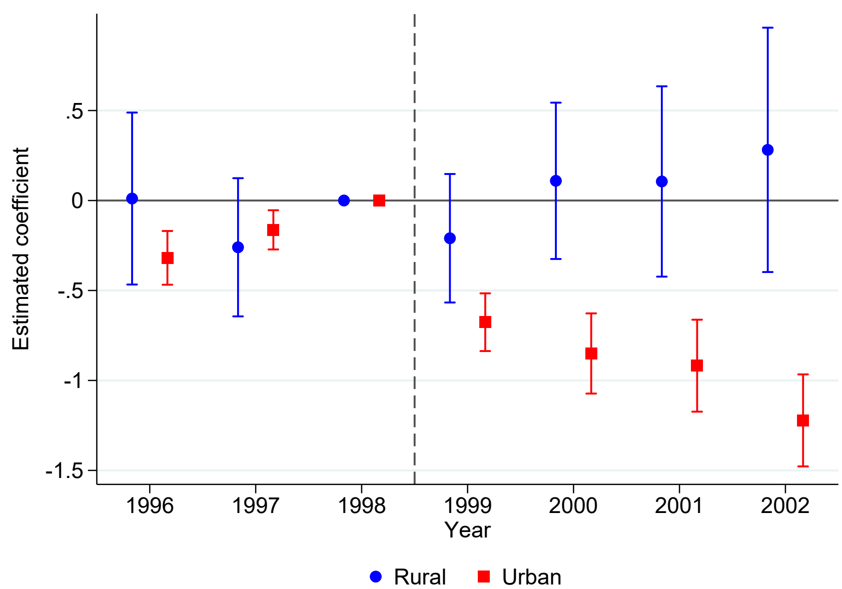
Figure 12: Definition of rural and urban districts

Notes: The figures illustrate the geographical distribution of rural and urban districts in Texas. In panel (a), only districts tagged as "Rural" are treated as rural districts. In panel (b), districts tagged as "Rural", "Non-Metropolitan Stable", and "Non-Metropolitan Fast Growing" are treated as rural districts. I follow the 2007 definition of districts by TEA, which is the earliest data available.

Figure 13: Event Study: Urban districts show larger exclusion of SE students.



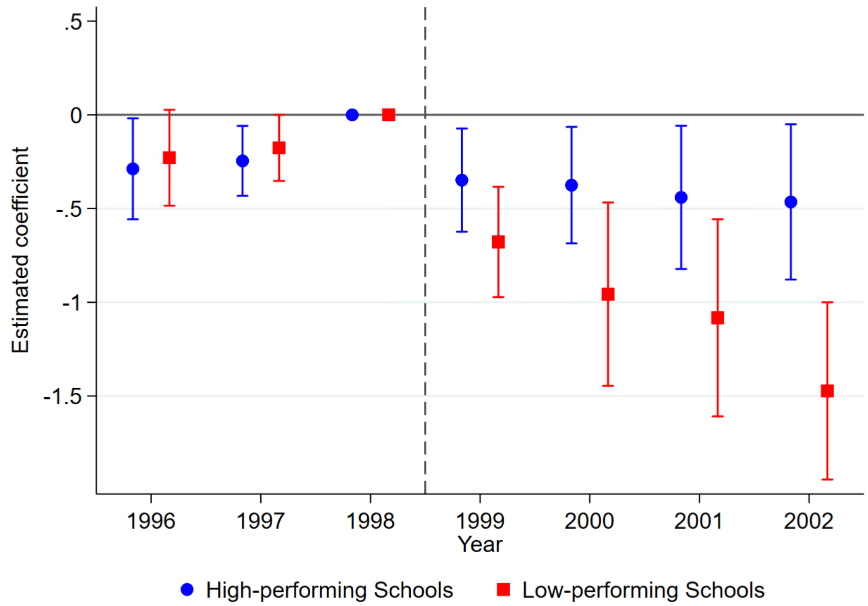
(a) Wide Definition of Rural Districts



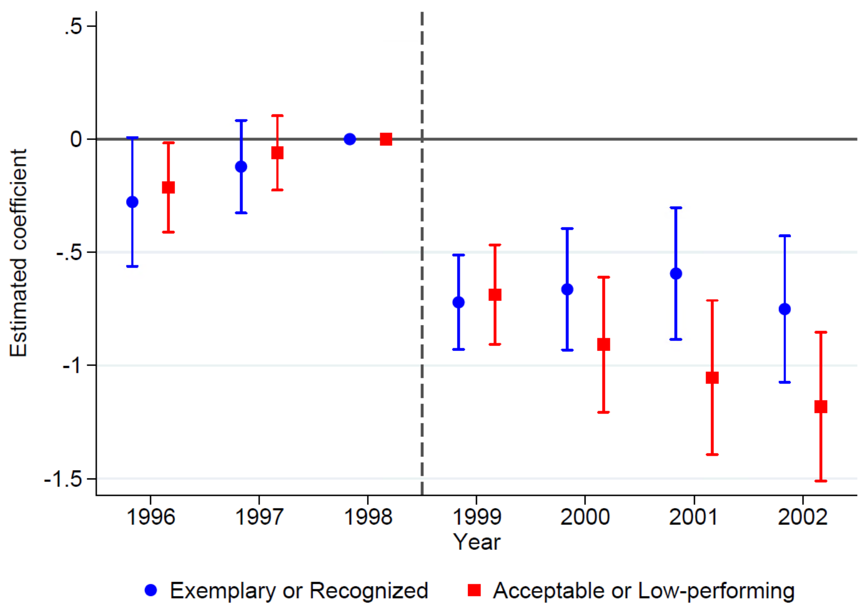
(b) Narrow Definition of Rural Districts

Notes: This figure shows heterogeneous effect estimates on TAAS participation rates by district geographic characteristics. Blue and red plots represent estimates from equation 2 based on students in urban and rural districts respectively. The definition of urban and rural districts follows that of Figure 15. In panel (a), only districts tagged as "Rural" are treated as rural districts. In panel (b), districts tagged as "Rural", "Non-Metropolitan Stable", and "Non-Metropolitan Fast Growing" are treated as rural districts. I plot the estimated coefficients of interest with corresponding 95 percent confidence intervals. Standard errors are clustered at the school levels.

Figure 14: Event Study: Schools with poorer performance show larger exclusion of SE students.



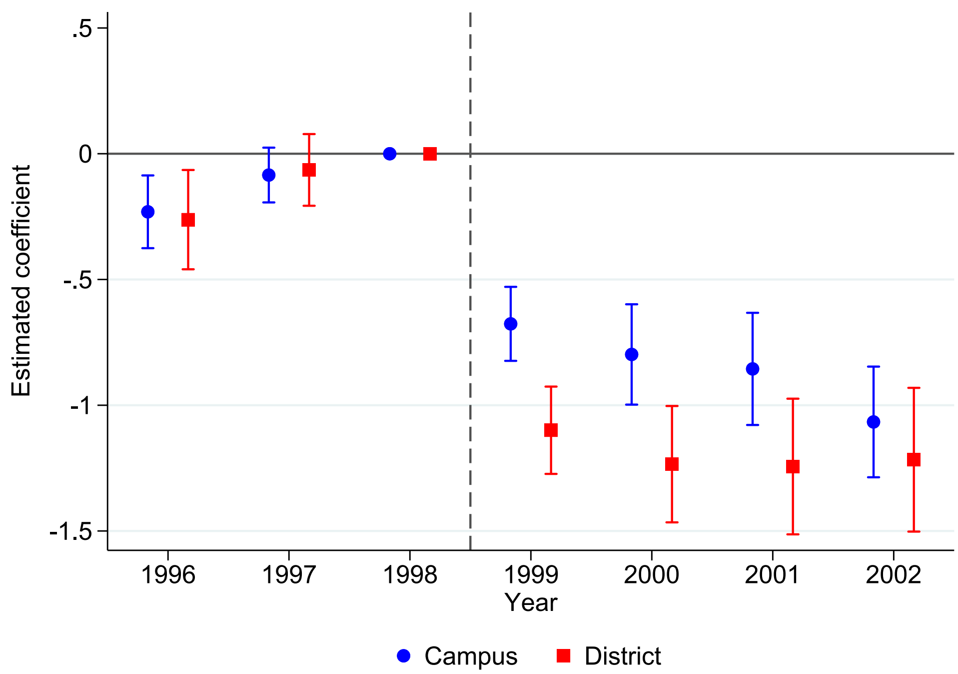
(a) By Previous Average Test Score



(b) By Previous Rating

Notes: This figure shows heterogeneous effect estimates on TAAS participation rates by measures of school performance. Blue and red plots represent estimates from equation 2 based on students in high- and low-performing schools respectively. In panel (a), high-performing schools are those with the top quartile of average test scores, and low-performing schools are those with the bottom quartile. I plot the estimated coefficients of interest with corresponding 95 percent confidence intervals. Standard errors are clustered at the school levels.

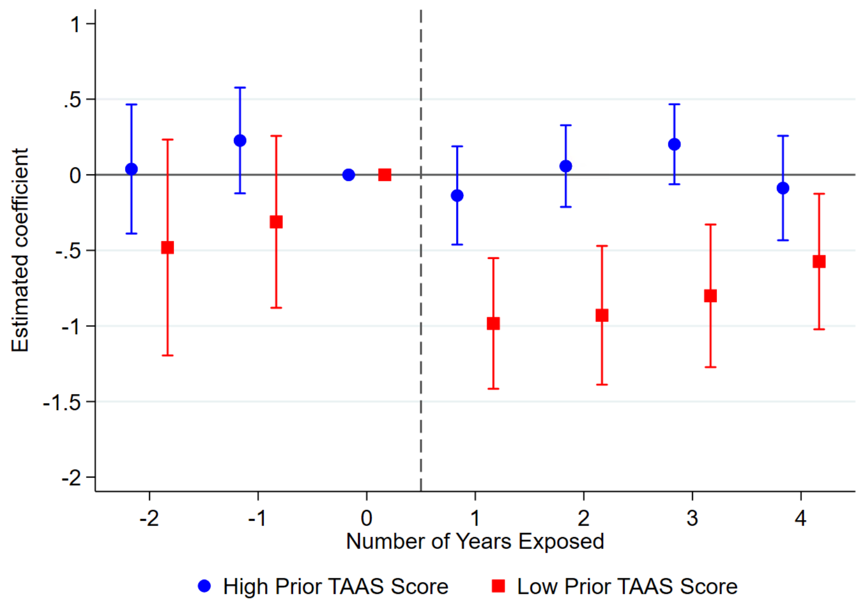
Figure 15: Event Study: The degree of exclusion was more responsive to district-level incentives.



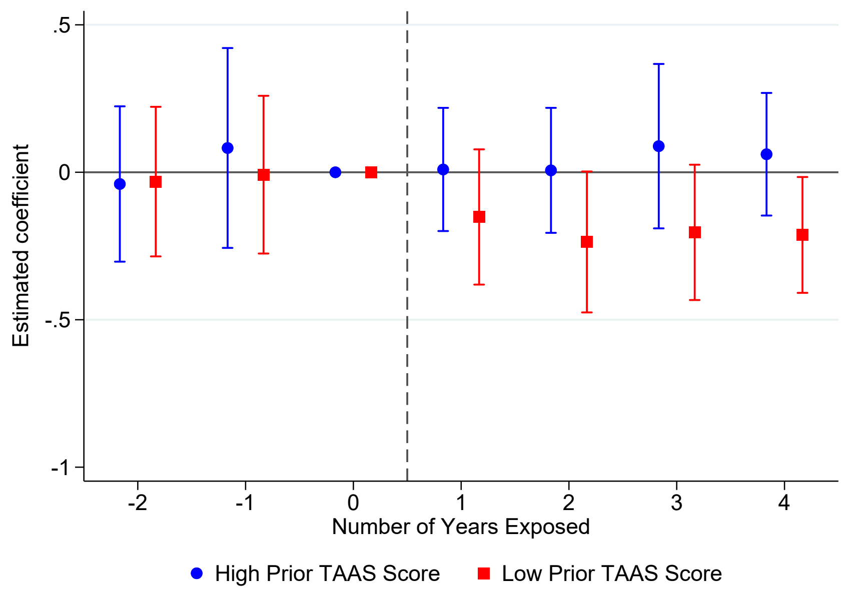
(a) Campus- vs. District-level Variation

Notes: This figure shows heterogeneous effect estimates on TAAS participation rates by levels of SE shares. Blue and red plots represent estimates from equation 2, based on school- and district-level SE share variation. I plot the estimated coefficients of interest with corresponding 95 percent confidence intervals. Standard errors are clustered at the school levels.

Figure 16: Event Study: Low-performing SE students faced exclusions in high school, including more dropouts.



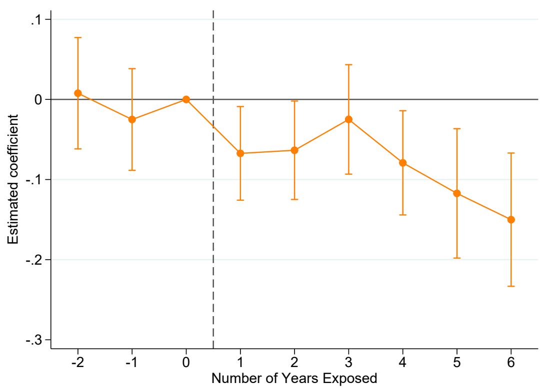
(a) Exit Exam Participation



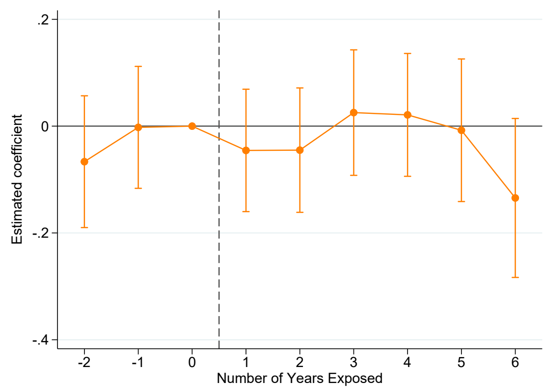
(b) Reaching 10th Grade

Notes: This figure shows long-run mechanism event study results based on equation 6 by students' past TAAS scores. I define "high performance" groups as students in the top tertile in terms of their past TAAS scores and vice versa. I assume that a student took the exit-level TAAS if he took at least one subject of the exam. I plot the estimated coefficients of interest with corresponding 95 percent confidence intervals. Standard errors are clustered at school levels.

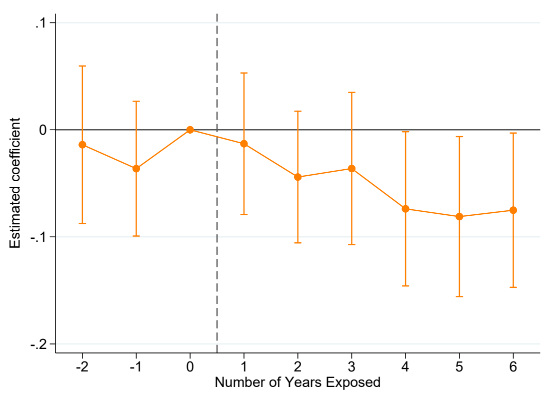
Figure 17: Event study: The accountability pressure negatively affected long-run outcomes.



(a) High School Graduation



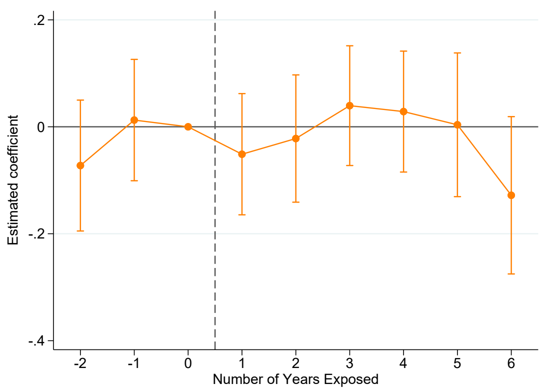
(b) College Enrollment



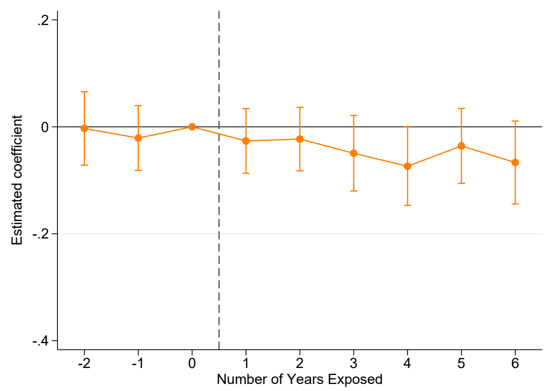
(c) College Degree

Notes: The figure plots event study estimates based on equation 6. I plot the estimated coefficients of interest with corresponding 95% confidence intervals. Standard errors are clustered at the school levels.

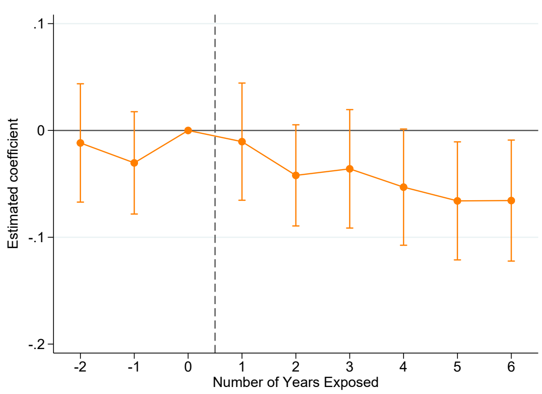
Figure 18: Event study: College Outcomes were mainly driven by 2-year colleges.



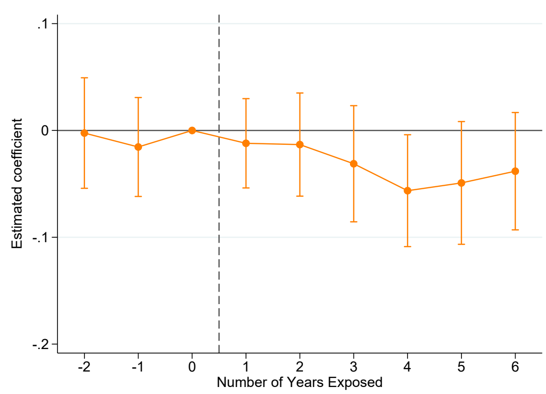
(a) College Enrollment, 2 Year



(b) College Enrollment, 4 Year



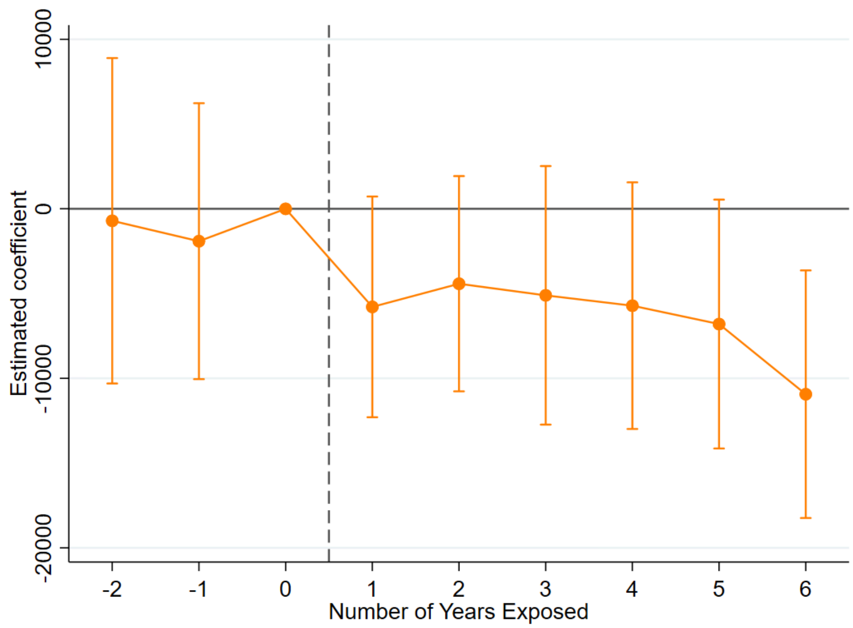
(c) College Degree, 2 Year



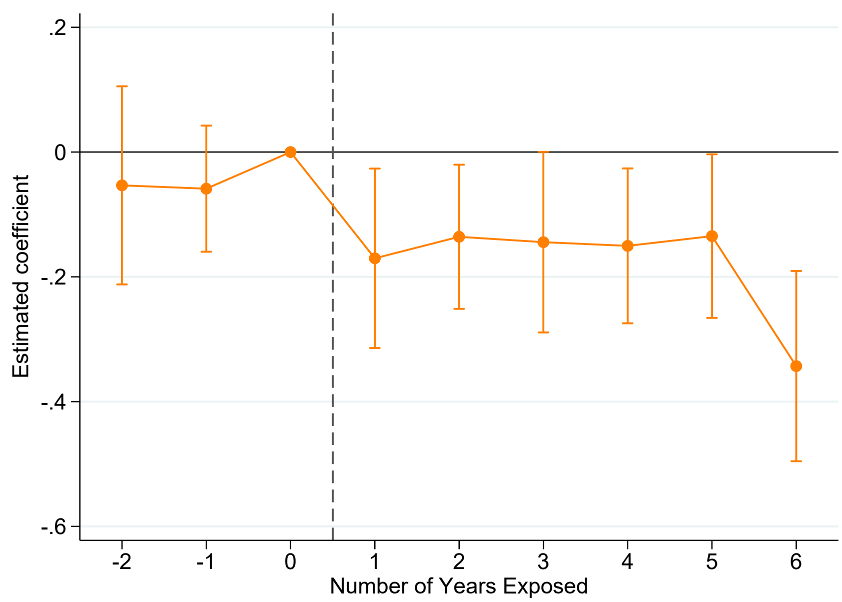
(d) College Degree, 4 Year

Notes: The figure plots event study estimates based on equation 6. I plot the estimated coefficients of interest with corresponding 95% confidence intervals. Standard errors are clustered at the school levels.

Figure 19: Event study: Wage levels decreased in extensive margins.



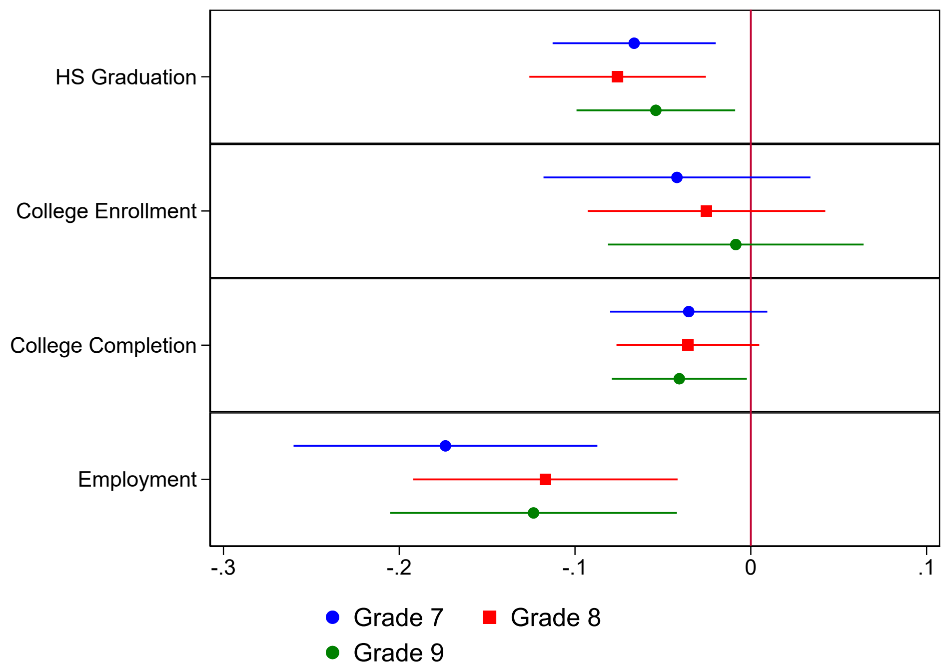
(a) Earning



(b) Employment

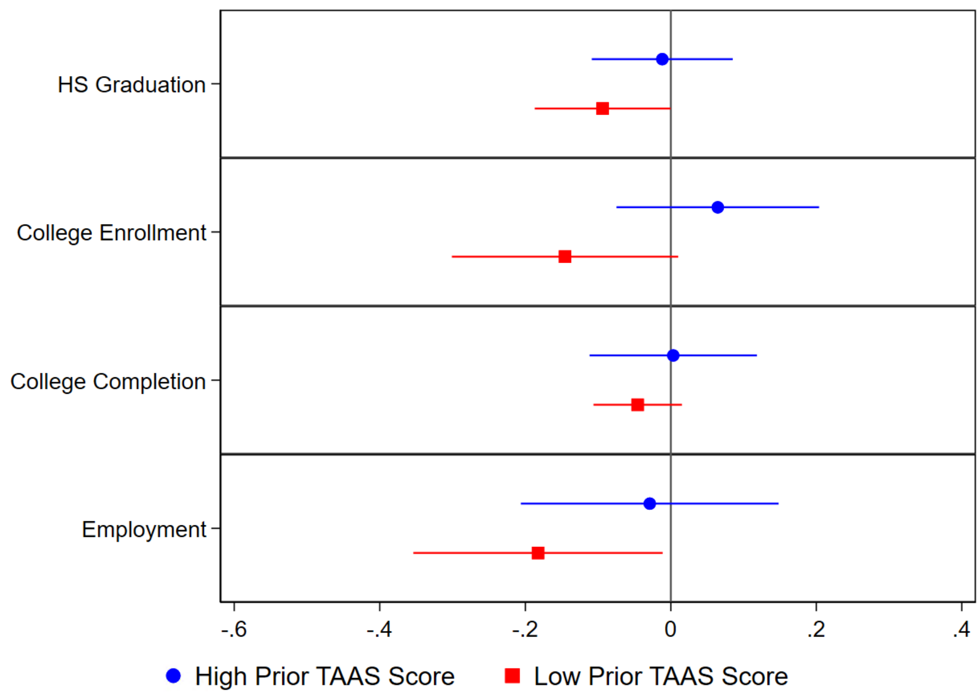
Notes: The figure plots event study estimates based on equation 6. I plot the estimated coefficients of interest with corresponding 95% confidence intervals. Standard errors are clustered at the school levels. All labor market outcomes are measured at the age between 25 and 29. Earnings include zero values.

Figure 20: Sensitivity Analysis: Timing of Special Education Status



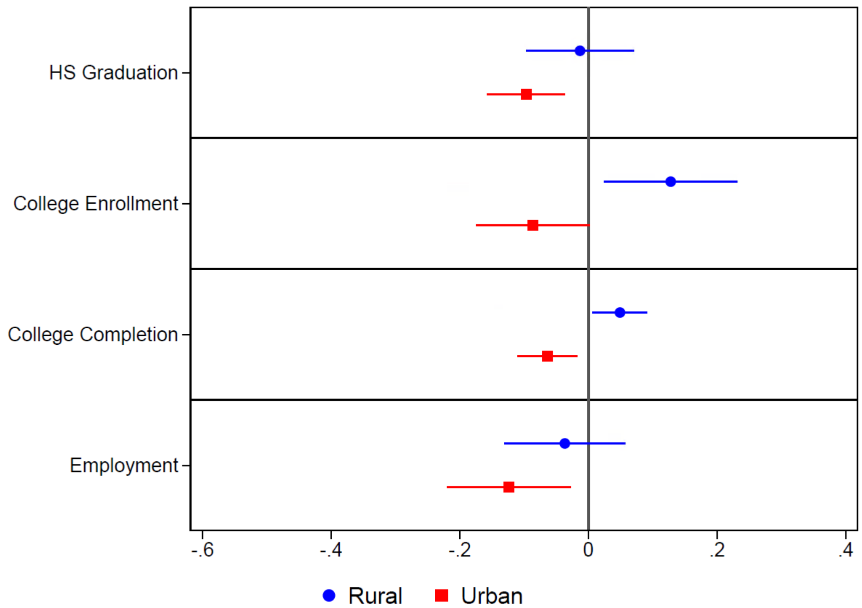
Notes: The figure depicts estimates of equation 5 using three different sample specifications. The blue plots present regression estimates when I use 9th-grade students who were in special education in 7th grade. Similarly, the red and green plots present estimates based on students who were in special education in 8th and 9th grades respectively. I plot the estimated coefficients of interest with corresponding 95% confidence intervals. Standard errors are clustered at the school levels.

Figure 21: Other outcomes deteriorated for low-performing students.

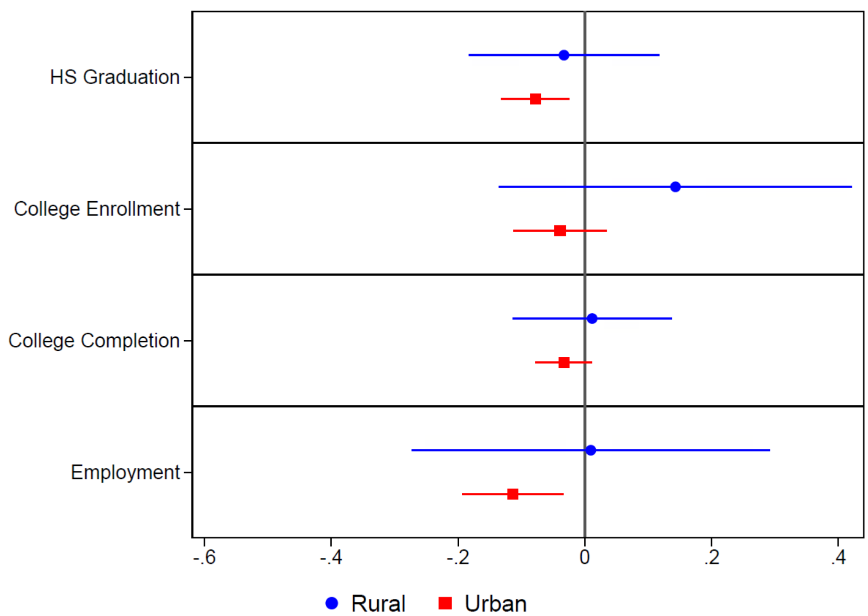


Notes: The figure reports estimated coefficients on long-run outcome based on equation 5 by students' past TAAS scores. I define "high performance" groups as students in the top tertile in terms of their past TAAS scores and vice versa. I plot the estimated coefficients of interest with corresponding 95% confidence intervals. Standard errors are clustered at school levels.

Figure 22: Urban districts drove adverse effects.



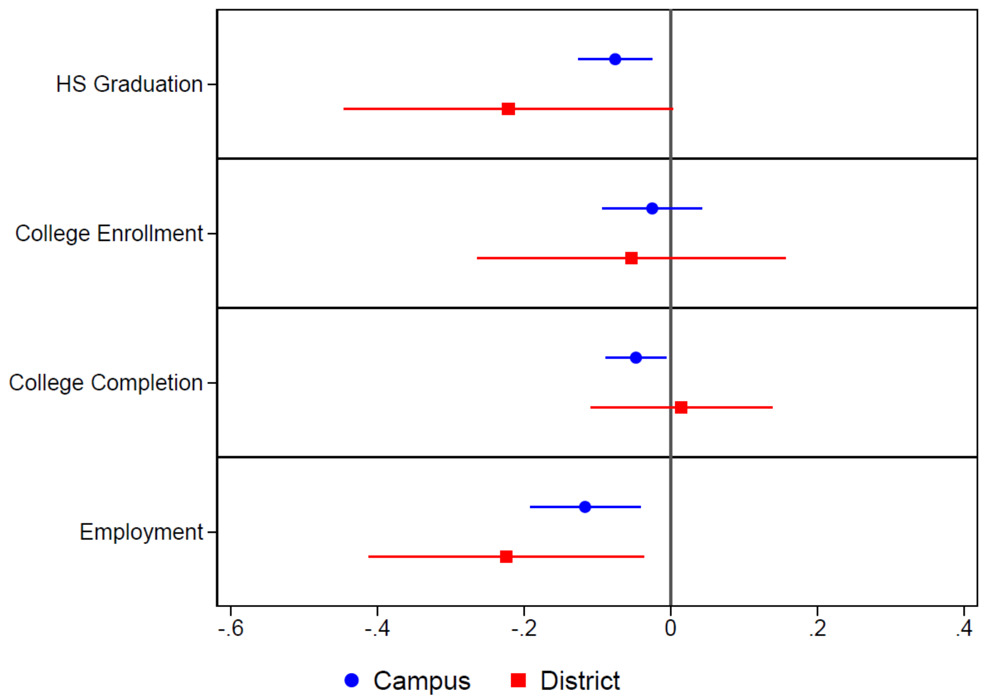
(a) Wide Definition of Rural Districts



(b) Narrow Definition of Rural Districts

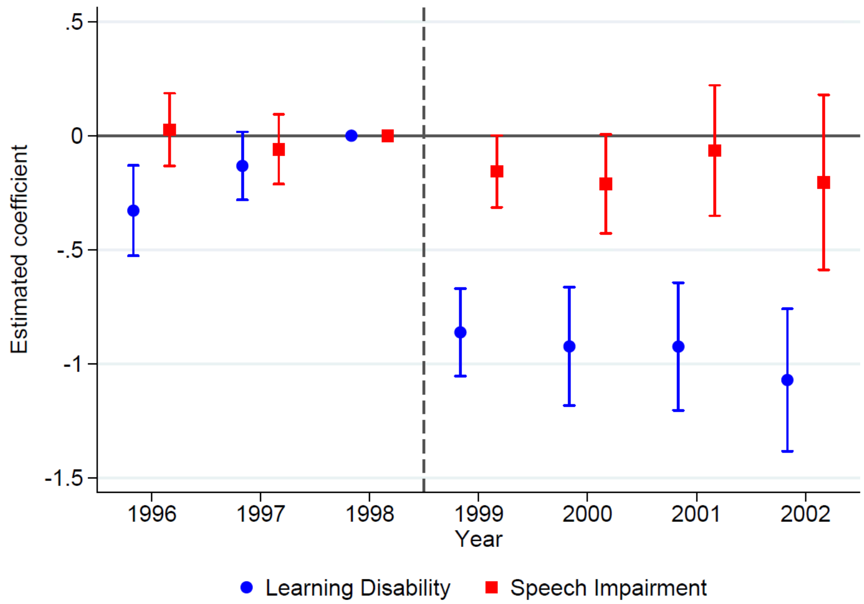
Notes: The figure reports estimated coefficients on long-run outcome based on equation 5 by district geographic characteristics. The definition of urban and rural districts follows that of Figure 15. In panel (a), only districts tagged as "Rural" are treated as rural districts. In panel (b), districts tagged as "Rural", "Non-Metropolitan Stable", and "Non-Metropolitan Fast Growing" are treated as rural districts. I plot the estimated coefficients of interest with corresponding 95% confidence intervals. Standard errors are clustered at school levels.

Figure 23: High school graduation and employment were more sensitive to district incentives.

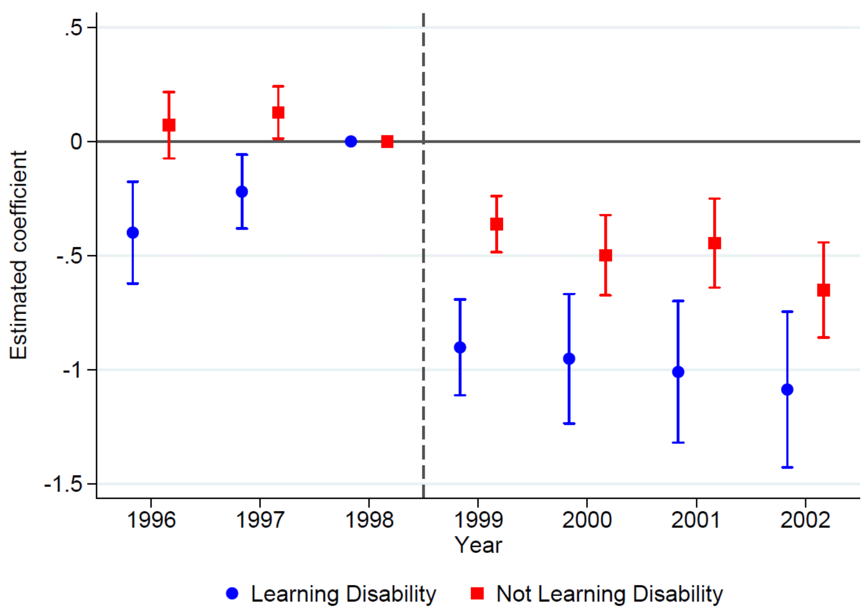


Notes: This figure shows heterogeneous effect estimates on long-term outcomes by levels of SE shares. Blue and red plots represent estimates from equation 5, based on school- and district-level SE share variation. I plot the estimated coefficients of interest with corresponding 95 percent confidence intervals. Standard errors are clustered at the school levels.

Figure 24: Event Study: Students with learning disabilities experienced more exclusions.



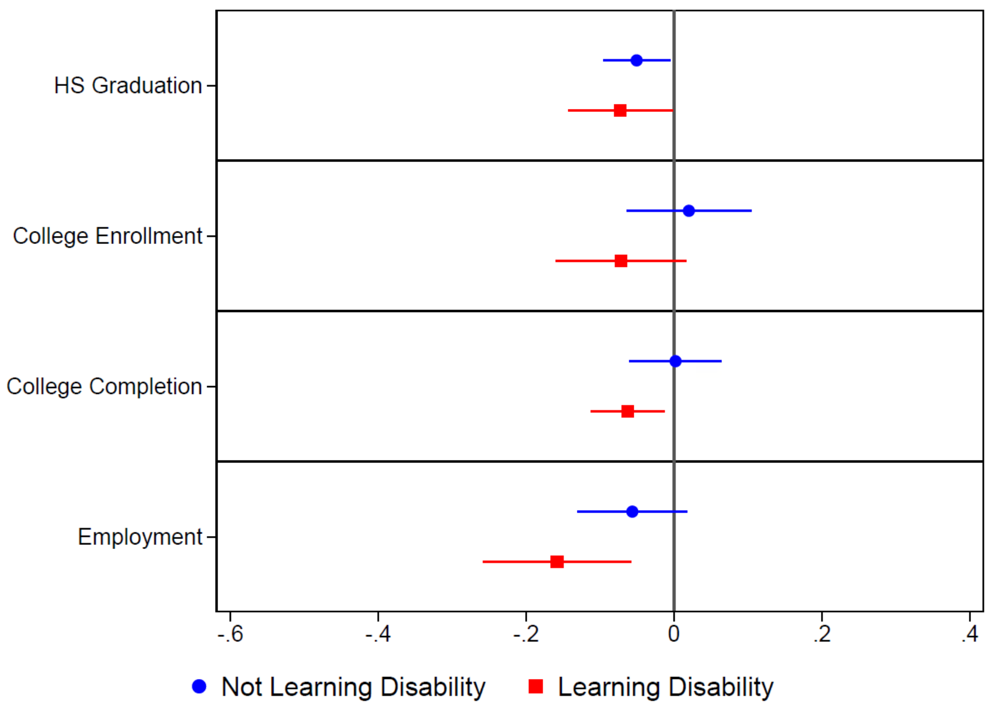
(a) Learning Disability vs. Speech Impairment



(b) Learning Disability vs. Others

Notes: This figure shows heterogeneous effect estimates on TAAS participation rates by disability types of SE students. Blue and red plots represent estimates from equation 2. I plot the estimated coefficients of interest with corresponding 95 percent confidence intervals. Standard errors are clustered at the school levels.

Figure 25: Adverse impacts on long-term outcomes were more severe on the learning disability group.



Notes: This figure shows heterogeneous effect estimates on long-term outcomes by disability types of SE students. Blue and red plots represent estimates from equation 5. I plot the estimated coefficients of interest with corresponding 95 percent confidence intervals. Standard errors are clustered at the school levels.

Table 1: Summary Statistics - 8th Grade Cohorts Between 1996 and 2002

	General Education		Special Education	
	1996 ∼ 1998 (1)	1999 ∼ 2002 (2)	1996 ∼ 1998 (3)	1999 ∼ 2002 (4)
Individual Characteristics				
Male	0.49	0.49	0.68	0.67
White	0.49	0.47	0.46	0.43
Black	0.13	0.13	0.18	0.19
Hispanic	0.36	0.37	0.35	0.37
Free/reduced-price lunch	0.41	0.42	0.56	0.58
Limited English proficiency	0.08	0.07	0.09	0.1
Educational Outcomes				
TAAS tested, reading	0.82	0.85	0.54	0.44
TAAS tested, math	0.82	0.85	0.53	0.42
Normalized score, reading	0.12	0.09	–1.07	–0.82
Normalized score, math	0.13	0.09	–1.12	–0.85
High school graduation	0.71	0.75	0.6	0.65
College enrollment	0.51	0.54	0.24	0.26
College enrollment, 4 year	0.25	0.26	0.05	0.05
College completion	0.22	0.24	0.06	0.06
College completion, 4 year	0.17	0.19	0.03	0.03
Labor Market Outcomes				
Annual income ($)	17,303	18,573	10,980	11,307
Employment	0.7	0.71	0.63	0.66
Rural district	0.14	0.13	0.17	0.16
Number of individuals	725,356	995,490	105,318	156,501

Notes: This table presents average individual characteristics, educational outcomes, and labor market outcomes of students in general and special education. I categorize students into general education unless they are specified as special education students in the data. Labor market outcomes are calculated between age 25 and 29. The annual income measure includes unemployed individuals with zero earnings and is deflated using 2000 CPI.

Table 2: Short-Run Effects on TAAS Scores

	Reading Score		Math Score	
	(1)	(2)	(3)	(4)
Share x Post	0.990***	–0.0631	0.951***	0.0234
	(0.176)	(0.102)	(0.181)	(0.105)
Individual FE	No	Yes	No	Yes
Observations	432,504	417,724	458,191	444,411
R-squared	0.266	0.807	0.266	0.829

Notes: This table presents estimated coefficients of equation 1. Significance levels at the 1%, 5%, and 10% are denoted by ***, **, and * respectively. Standard errors are clustered at the school level.

Table 3: Short-Run Effects on TAAS Participation Rates

	Tested		Tested, Reading		Tested, Math	
	(1)	(2)	(3)	(4)	(5)	(6)
Share × Post	–0.678***	–0.196***	–0.734***	–0.304***	–0.701***	–0.272***
	(0.0819)	(0.0640)	(0.0818)	(0.0665)	(0.0835)	(0.0643)
PrevScore × Share × Post		0.416***		0.444***		0.377***
		(0.0521)		(0.0524)		(0.0511)
Observations	819,087	435,709	819,087	435,709	819,087	435,709
R-squared	0.253	0.299	0.271	0.317	0.256	0.296

Notes: This table presents estimated coefficients of equation 1 and 2. Significance levels at the 1%, 5%, and 10% are denoted by ***, **, and * respectively. Standard errors are clustered at the school level.

Table 4: Long-run Effects on Educational Outcomes

	High School Graduation	College Enrollment	College Completion
	(1)	(2)	(3)
Share x Expose	-0.0732***	–0.138	–0.0367
	(0.0254)	(0.0824)	(0.0201)
Observations	355,239	355,239	355,239
R-squared	0.155	0.112	0.051

Notes: This table presents estimated coefficients of equation 5. Significance levels at the 1%, 5%, and 10% are denoted by ***, **, and * respectively. Standard errors are clustered at the school level.

Table 5: Long-run Effects on College Outcomes

	Enrollment, 2 year	Enrollment, 4 year	Completion, 2 year	Completion, 4 year
	(1)	(2)	(3)	(4)
Share x Expose	–0.131	–0.0742	–0.0354**	–0.0211
	(0.0798)	(0.0486)	(0.0145)	(0.0156)
Observations	355,239	355,239	355,239	355,239
R-squared	3103	0.078	0.022	0.054

Notes: This table presents estimated coefficients of equation 5. Significance levels at the 1%, 5%, and 10% are denoted by ***, **, and * respectively. Standard errors are clustered at the school level.

Table 6: Long-run Effects on Labor Market Outcomes

	Earning	Employment
	(1)	(2)
Share x Expose	–3094.801*	–0.121***
	(1786.117)	(0.0370)
Observations	395,561	395,561
R-squared	0.056	0.04

Notes: This table presents estimated coefficients of equation 5. Significance levels at 1%, 5%, and 10% are denoted by ***, **, and * respectively. Standard errors are clustered at the school level. Earnings include zero values.

Table 7: Robustness Check : Alternative Sample and Treatment Intensity Specification

	Base		Unbalanced		Annual Shares		Unbalanced, Annual Shares	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Share x Post	–0.678***	–0.196***	–0.495***	–0.200***	–0.670***	0.0720	–0.625***	–0.272***
	(0.0819)	(0.0640)	(0.0639)	(0.0600)	(0.0531)	(0.0531)	(0.0556)	(0.0643)
PrevScore x Share x Post		0.416***		0.225***		0.458***		0.377***
		(0.0521)		(0.0474)		(0.0442)		(0.0511)
Observations	819,087	435,709	2,288,044	824,397	823,444	441,511	2,351,265	836,447
R-squared	0.253	0.299	0.231	0.231	0.252	0.299	0.209	0.234

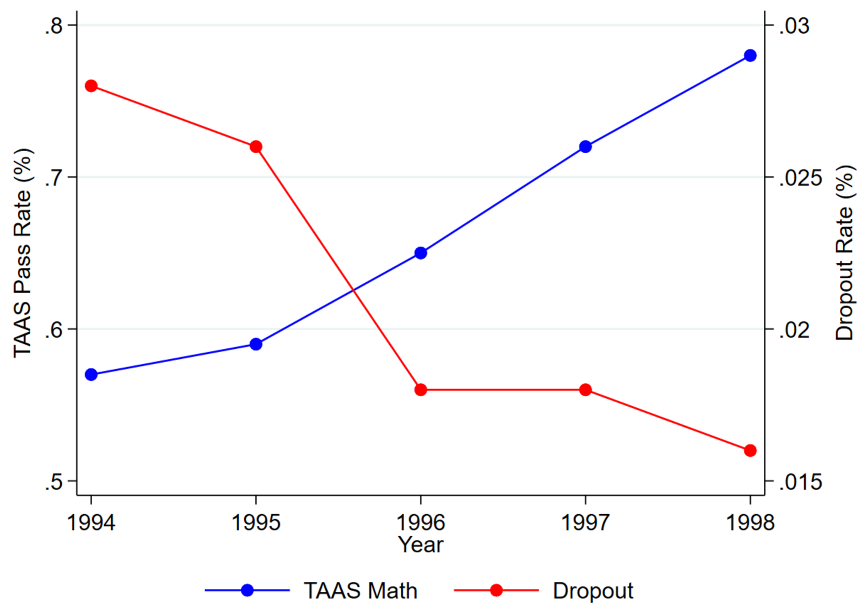
Notes: This table presents estimated coefficients of equation 1 and 2. Significance levels at the 1%, 5%, and 10% are denoted by ***, **, and * respectively. Standard errors are clustered at the school level.

Table 8: Summary Statistics - Test Outcomes by Disability Types

	Learning Disability (1)	Speech Impairment (2)	Emotional Disturbance (3)	Mental Retardism (4)
Short-run Analysis (G3–8)				
Fraction	0.612	0.187	0.056	0.053
TAAS Participation				
Reading	0.43	0.79	0.45	0.35
Math	0.49	0.8	0.47	0.38
TAAS Score				
Reading	-1.18	-0.25	-0.67	-1.97
Math	-1.08	-0.19	-0.82	-2.06
Long-run Analysis (G9)				
Fraction	0.699	0.016	0.114	0.064
TAAS Participation				
Reading	0.55	0.77	0.51	0.05
Math	0.55	0.78	0.49	0.05
TAAS Score				
Reading	-1.45	-0.72	-0.99	-2.37
Math	-1.38	-0.69	-1.12	-2.46

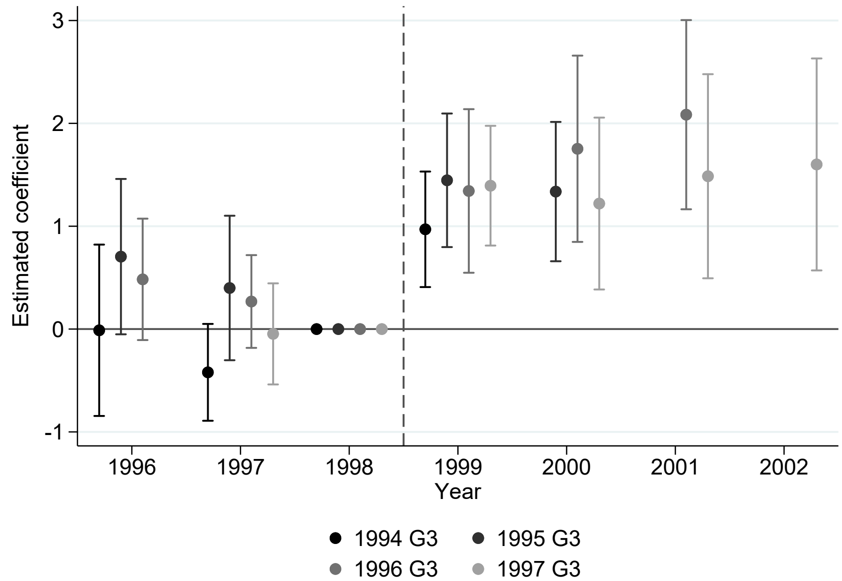
A Appendix Figures

Figure A.1: Trends Educational Outcomes After Introduction of Accountability System

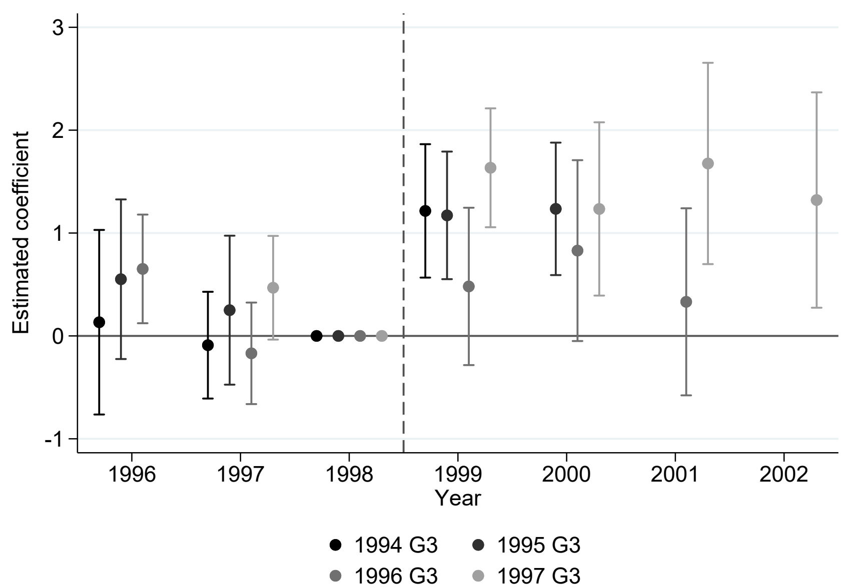


Notes: The figure plots state-level average educational outcomes of Texas after the implementation of the full-scale school accountability system in 1994. Refer to Haney (2000) for more details.

Figure A.2: Event study: Test Score by Grade Cohorts.



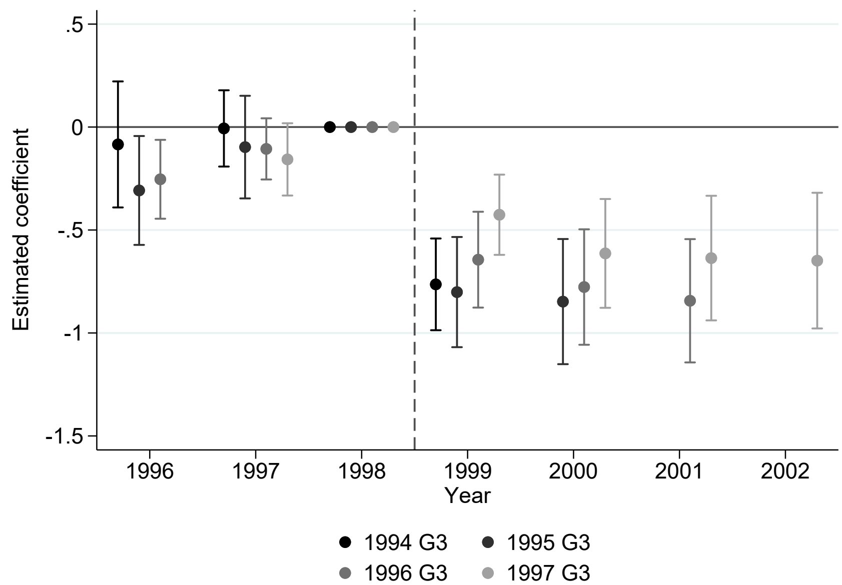
(a) Reading Score



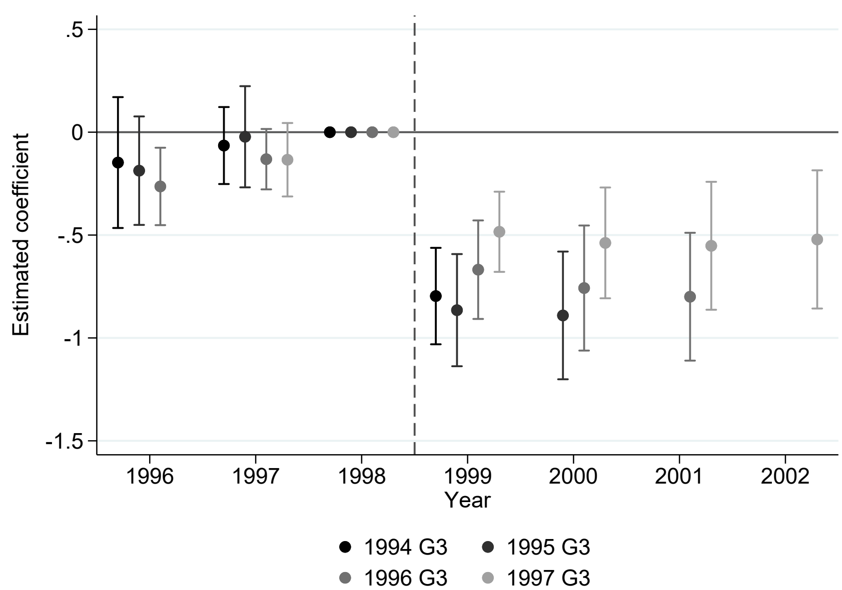
(b) Math Score

Notes: The figure plots event study estimates based on equation 2. Panels (a) and (b) report the results separately for each cohort of the balanced sample. I plot the estimated coefficients of interest with corresponding 95 percent confidence intervals. Standard errors are clustered at the school levels. Test scores are normalized to have a mean of 0 with a standard deviation of 1 within each grade level, subject, and year.

Figure A.3: Event study: Test Participation by Grade Cohorts



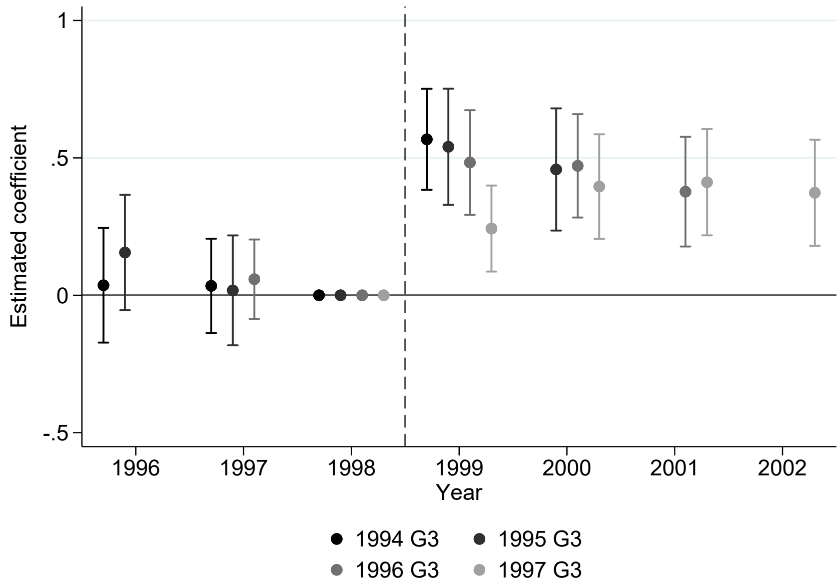
(a) Reading Test Participation



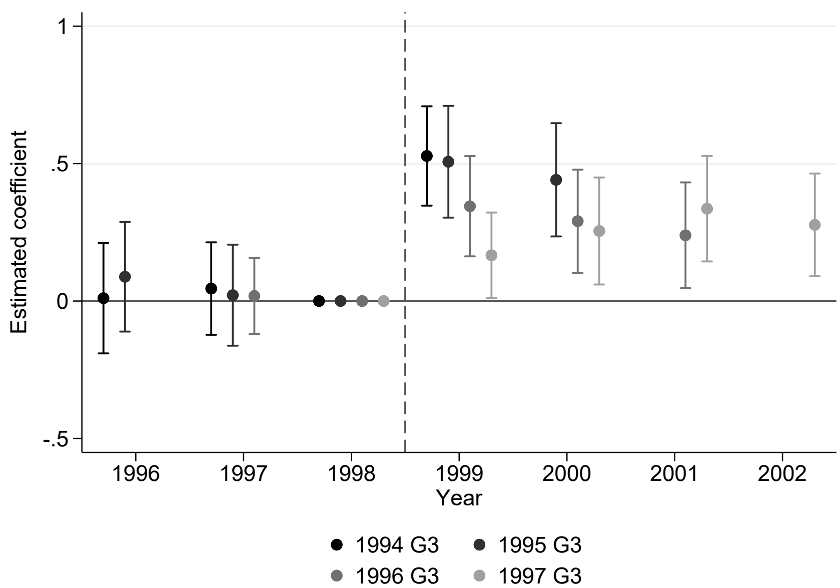
(b) Math Test Participation

Notes: The figure plots event study estimates based on equation 2. Panels (a) and (b) report the results separately for each cohort of the balanced sample. I plot the estimated coefficients of interest with corresponding 95 percent confidence intervals. Standard errors are clustered at the school levels.

Figure A.4: Event study: Heterogeneous Effect on Test Participation by Grade Cohorts



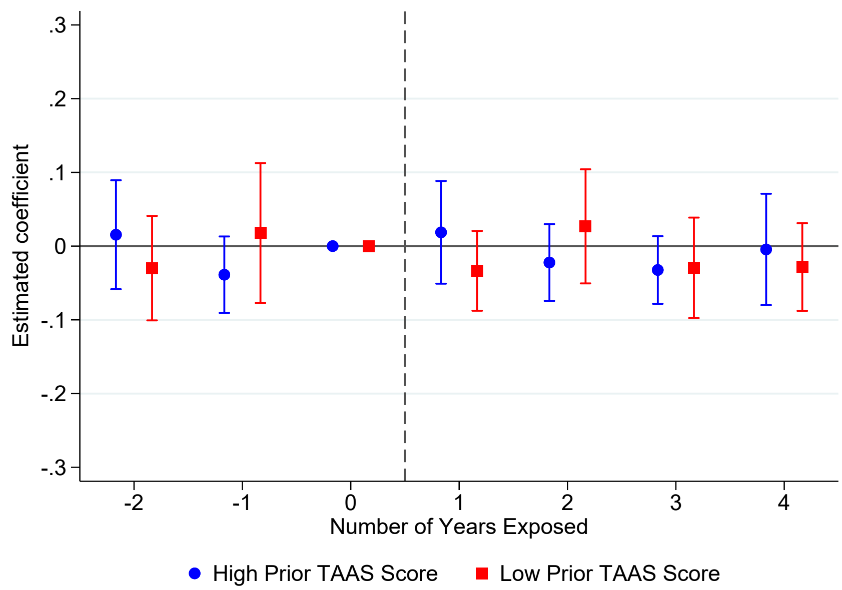
(a) Reading Test Participation



(b) Math Test Participation

Notes: The figure plots event study estimates based on equation 4. Panels (a) and (b) show the results using the balanced sample I described. I plot the estimated coefficients of interest with corresponding 95 percent confidence intervals. Standard errors are clustered at the school levels.

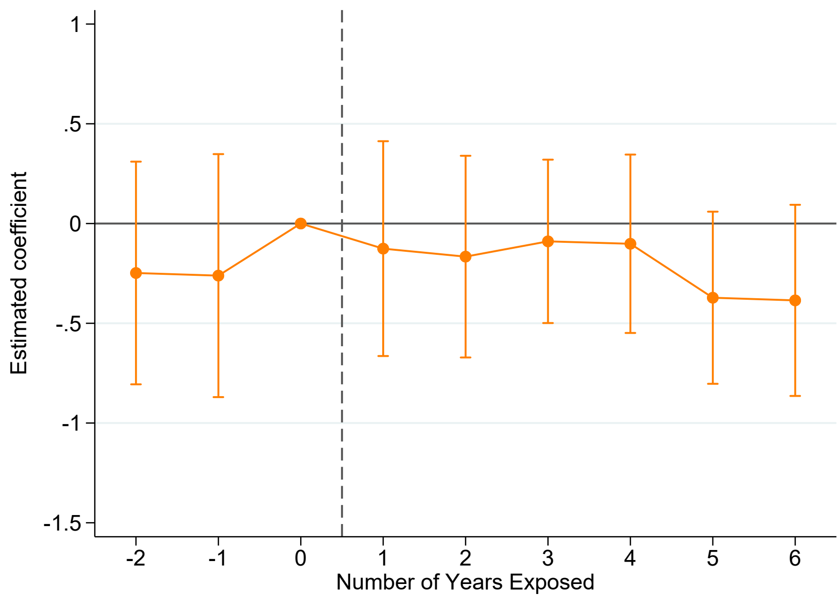
Figure A.5: Event Study: Effects on Dropouts from Official Dropout Records



(a) Dropout

Notes: This figure shows event study results analogous to Figure 16.(b) based on equation 6 and the official TEA dropout data. I define "high performance" groups as students in the top tertile in terms of their past TAAS scores and vice versa. I assume that a student took the exit-level TAAS if he took at least one subject of the exam. I plot the estimated coefficients of interest with corresponding 95 percent confidence intervals. Standard errors are clustered at school levels.

Figure A.6: Event Study: Effects on Earnings Conditional on Employment



(a) Earnings, Employed

Notes: This figure shows event study results on log wage based on equation 6. All observations with zero earnings are excluded from the analysis. I plot the estimated coefficients of interest with corresponding 95 percent confidence intervals. Standard errors are clustered at school levels.

B Appendix Tables

Table A1: Summary Statistics by SE Share Levels (Grade 8)

	High SE Share	**Low SE Share**
	(1)	(2)
Individual Characteristics		
Male	0.48	0.49
White	0.52	0.46
Black	0.13	0.12
Hispanic	0.33	0.37
Free/reduced-price lunch	0.46	0.38
Limited English proficiency	0.04	0.09
Educational Outcomes		
TAAS tested, reading	0.87	0.82
TAAS tested, math	0.87	0.82
Normalized score, reading	0.08	0.18
Normalized score, math	0.09	0.19
High school graduation	0.70	0.72
College enrollment	0.49	0.54
College enrollment, 4 year	0.22	0.29
College completion	0.19	0.25
College completion, 4 year	0.15	0.21
Labor Market Outcomes		
Annual income ($)	16,731	17,800
Employment	0.73	0.68
Number of individuals	96,168	232,334

Notes: This table presents average individual characteristics, educational outcomes, and labor market outcomes of all students in schools with high and low levels of SE shares. Labor market outcomes are calculated between ages 25 and 29. The annual income measure includes unemployed individuals with zero earnings and is deflated using 2000 CPI. Samples are limited to 8th graders between 1996 and 1998.