School Accountability, Postsecondary Attainment and Earnings

David J. Deming, Harvard University and NBER*

Sarah Cohodes, Columbia University

Jennifer Jennings, New York University

Christopher Jencks, Harvard University

December 2015

Abstract

We study the impact of accountability pressure in Texas public high schools in the 1990s on postsecondary attainment and earnings, using administrative data from the Texas Schools Project (TSP). Schools respond to the risk of being rated Low-Performing by increasing student achievement on high-stakes exams. Years later, these students are more likely to have attended college and completed a four-year degree, and they have higher earnings at age 25. However, we find no overall impact of accountability pressure to achieve a higher rating, and large negative impacts on attainment and earnings for the lowest-scoring students.

* Corresponding author Deming: Harvard Graduate School of Education, 8 Appian Way, Gutman 411, Cambridge MA 02138 (email: david_deming@gse.harvard.edu. We thank Dick Murnane, Dan Koretz, David Figlio, Jonah Rockoff, Raj Chetty, John Friedman and seminar participants at Harvard, Stanford, Columbia, the University of Wisconsin, the NBER Summer Institute, CESifo, for helpful comments. This project was supported by the Spencer Foundation and the William T. Grant Foundation. Very special thanks to Maya Lopuch for invaluable research assistance. We gratefully acknowledge Rodney Andrews, Greg Branch and the staff of the UT-Dallas Education Research Center for making this research possible. The conclusions of this research do not necessarily reflect the opinions of the Texas Education Agency, the Texas Higher Education Coordinating Board, or the State of Texas.

Today's schools must offer a rigorous academic curriculum to prepare students for the rising skill demands of the modern economy (Levy and Murnane, 2012). Yet at least since the publication of *A Nation at Risk* in 1983, policymakers have acted on the principle that America's schools are failing. The ambitious and far-reaching No Child Left Behind Act of 2002 (NCLB) identified test-based accountability as the key to improved school performance. NCLB mandates that states conduct annual standardized assessments in math and reading, that schools' average performance on assessments be publicized, and that rewards and sanctions be doled out on the basis of student exam performance.

However, more than a decade after the passage of NCLB, we know very little about the impact of test-based accountability on students' long-run life chances. Previous work has found large gains on high-stakes tests, with some evidence of smaller gains on low-stakes exams that is inconsistent across grades and subjects (e.g. Koretz and Barron 1998, Klein et al. 2000, Carnoy and Loeb 2002, Hanushek and Raymond 2005, Jacob 2005, Wong, Cook and Steiner 2009, Dee and Jacob 2010, Reback, Rockoff and Schwartz 2014). There are many studies of strategic responses to accountability pressure, ranging from focusing instruction on marginal students, narrow test content and coaching, manipulating the pool of accountable students, boosting the nutritional content of school lunches, and teacher cheating (Haney 2000, McNeil and Valenzuela 2001, Jacob and Levitt 2003, Diamond and Spillane 2004, Figlio and Winicki 2005, Booher-Jennings 2005, Jacob 2005, Cullen and Reback 2006, Figlio and Getzler 2006, Vasquez Heilig and Darling-Hammond 2008, Reback 2008, Neal and Schanzenbach 2010).

When do improvements on high-stakes tests represent real learning gains? And when do they make students better off in the long-run? The main difficulty in interpreting accountability-induced student achievement gains is that once a measure becomes the basis of assessing performance, it loses its diagnostic value (Campbell 1976, Kerr 1995, Neal 2013). Previous research has focused on measuring performance on low-stakes exams, yet academic achievement is only one of many possible ways that teachers and schools may affect students (Chetty, Friedman and Rockoff 2014, Jackson 2012).

While there are many goals of public schooling, test-based accountability is premised on the belief that student achievement gains will lead to long-run improvements in important life outcomes such as educational attainment and earnings. High-stakes testing creates incentives for teachers and schools to adjust their effort toward improving test performance in the short-run. Whether these changes make students better off in the long-run depends critically on the correlation between the actions that schools take to raise test scores, and the resulting changes in earnings and educational attainment at the margin (Holmstrom and Milgrom 1991, Baker 1992, Hout et al., 2011).

In this paper we examine the long-run impact of test-based accountability in Texas public high schools. We use data from the Texas Schools Project, which links PK-12 records from all public schools in Texas to data on college attendance, degree completion and labor market earnings in their state. Texas implemented high-stakes accountability in 1993, and high school students in the mid to late 1990s are now old enough to examine outcomes in young adulthood. High schools were rated by the share of 10th grade students who received passing scores on exit exams in math, reading and writing. Schools were assigned an overall rating based on the pass rate of the lowest scoring test-subgroup combination (e.g math for whites), giving some schools strong incentives to focus on particular students, subjects and grade cohorts. School ratings were published in full page spreads in local newspapers, and schools that were rated as Low-Performing were forced to undergo an evaluation that could lead to serious consequences such as layoffs, reconstitution and/or school closure (TEA 1994, Haney 2000, Cullen and Reback 2006).

Our research design compares grade cohorts within a school that faced different degrees of accountability pressure due to policy-induced changes in the ratings thresholds. In 1995, at least 25 percent of all students in a high school were required to pass the 10th grade exit exam in each subject to receive a passing ("Acceptable") rating. This standard rose by 5 percentage points per year, up to 50 percent in 2000. Schools were also required to meet the passing standard for key subgroups. We use this policy variation to estimate the "risk" that a school will receive a particular rating, and we compare cohorts that are on the margin of receiving a particular rating to other cohorts that are plausibly "safe" from accountability pressure. Estimating schools' perceptions of accountability pressure is an inherently subjective exercise, and so we demonstrate that our results hold across a wide variety of alternative specifications. For example, we show that they are robust to comparison with placebo cohorts who would be "at risk" except that the lowest-scoring subgroup is below a minimum size threshold for accountability purposes.

We find that students score significantly higher on the 10th grade math exam when they are in a grade cohort that is at risk of receiving a Low-Performing rating. These students are more likely to graduate from high school "on time" and accumulate significantly more math credits, including in subjects beyond a 10th grade level. Later in life, they are more likely to attend and graduate from a four-year college, and they have higher earnings at age 25. The impacts are concentrated almost entirely among students with low 8th grade scores.

However, we find no impact on test scores of accountability pressure in schools that were close to receiving a high rating (called "Recognized"), and significant *declines* in math credit accumulation, attainment and earnings for low-scoring students. We present strong suggestive evidence that the negative impacts were due to strategic classification of low-scoring students as eligible for special education, and thus exempt from the "accountable" pool of test-takers.

We find that accountability pressure to avoid a Low-Performing rating leads to increases in labor market earnings at age 25 of around 1 percent. This is similar in size to the impact of having a teacher with 1 standard deviation higher “value-added”, and it lines up reasonably well with cross-sectional estimates of the impact of test score gains on young adult earnings (Chetty, Friedman and Rockoff 2014; Neal and Johnson 1996, Currie and Thomas 1999, Chetty et al. 2011). Broadly, our results indicate that school accountability led to long-run gains in schools that were at risk of falling below a minimum performance standard. Efforts to regulate school quality at a higher level (through the achievement of a Recognized rating), however, did not benefit students and may have caused long-run harm.

The accountability system adopted by Texas in 1993 was similar in many respects to the requirements of NCLB, which was enacted nine years later. NCLB required that states rate schools based on the share of students who pass standardized exams. It also required states to report subgroup test results, and to increase testing standards over time. Thus our findings may have broad applicability to the accountability regimes that were rolled out in other states over this period. However, because we compare schools that face different degrees of pressure within the same high-stakes testing regime, our study explicitly differences out any common trend in outcomes caused by school accountability. We estimate the net impact of schools’ responses along a variety of margins, including focusing on “bubble” students and subjects, teaching to the test, and manipulating the eligible test-taking pool. Our results are the net impact of schools’ responses along a variety of margins, and do not imply that school accountability in Texas was optimally designed (Neal 2013).

I. Background

Beginning in the early 1990s, a handful of states such as Texas and North Carolina implemented “consequential” school accountability policies, where school performance on standardized tests was not only made public but also tied to rewards and sanctions (Carnoy and Loeb 2002, Hanushek and Raymond 2005, Dee and Jacob 2010, Figlio and Loeb 2011). The number of states with some form of school accountability rose from 5 in 1994 to 36 in 2000, and scores on high-stakes tests rose rapidly in states that were early adopters of school accountability (Hanushek and Raymond 2005, Figlio and Ladd 2007, Figlio and Loeb 2011). Under then Governor and future President George W. Bush, test-based accountability in Texas served as a template for the federal No Child Left Behind (NCLB) Act of 2002.

Figure 1 shows pass rates on the 8th and 10th grade reading and mathematics exams for successive cohorts of first-time 9th graders in Texas. Pass rates on the 8th grade math exam rose from about 58 percent in the 1994 cohort to 91 percent in the 2000 cohort, only six years later. Similarly, pass rates on the 10th grade exam, which was a high-stakes exit exam for students, rose from 57 percent to 78 percent, with smaller yet still sizable gains in reading. This rapid rise in pass rates has been referred to in the literature as the "Texas miracle" (Klein et al 2000, Haney 2000).

The interpretation of the "Texas miracle" is complicated by studies of strategic responses to high-stakes testing. Past research has found that scores on high-stakes tests improve, often dramatically, whereas performance on a low-stakes test with different format but similar content improves only slightly or not at all, a phenomenon known as "score inflation" (Koretz et al 1991, Koretz and Barron 1998, Klein et al 2000, Jacob 2005). Studies of the implementation of accountability in Texas and other settings have found that schools raised test scores by retaining low-performing students in 9th grade, classifying them as eligible for special education or otherwise exempt from taking the exam, and encouraging them to drop out (Haney 2000, McNeil and Valenzuela 2001, Jacob 2005, Cullen and Reback 2006, Figlio 2006, Figlio and Getzler 2006, Vasquez Heilig and Darling-Hammond 2008, McNeil et al 2008, Jennings and Beveridge 2009).

Performance standards that use short-run, quantifiable measures are often subject to distortion (Kerr 1975, Campbell 1976). As in the multi-task moral hazard models of Holmstrom and Milgrom (1991) and Baker (1992), performance incentives cause teachers and schools to adjust their effort toward the least costly ways of increasing test scores, possibly at the expense of actions that are important for students' long-run welfare. In the context of school accountability, the concern is that schools will focus on short-run improvements in test performance at the expense of higher-order learning, creativity, self-motivation, socialization and other important skills that are related to the long-run success of students. The key insight from Holmstrom and Milgrom (1991) and Baker (1992) is that the value of performance incentives depends on the correlation between a performance measure (high-stakes tests) and true productivity (attainment, earnings) *at the margin* (Hout et al, 2011). In other words, when schools face accountability pressure, do the actions they take to raise short-run test scores positively or negatively affect attainment, earnings and other long-run outcomes?

The literature on school accountability has focused on low-stakes tests, in an attempt to measure whether gains on high-stakes exams represent generalizable gains in student learning. Recent studies of accountability in multiple states have found achievement gains across subjects and grades on low-stakes exams (Ladd 1999, Carnoy and Loeb 2002, Greene and Winters 2003, Hanushek and Raymond 2005, Figlio and Rouse 2006, Chiang 2009, Dee and Jacob 2010, Cook and Steiner 2011, Allen and Burgess 2012).

Yet scores on low-stakes exams may miss important dimensions of responses to test pressure. Other studies of accountability have found that schools narrow their curriculum and instructional practices in order to raise scores on the high-stakes exam, at the expense of low-stakes subjects, students, and grade cohorts (Stecher et al 2000, Diamond and Spillane 2004, Booher-Jennings 2005, Hamilton et al 2005, Jacob 2005, Diamond 2007, Hamilton et al 2007, Reback 2008, Neal and Schanzenbach 2010, Lauen and Ladd 2010, Reback, Rockoff and Schwartz 2014, Dee and Jacob 2012). Increasing achievement is only one of many possible ways that schools and teachers may affect students (Chetty, Friedman and Rockoff 2014 Jackson 2012). Studies of early life and school-age interventions often find long-term impacts on outcomes despite "fade out" or non-existence of test score gains (Gould et al 2004, Belfield et al 2006, Booker et al 2009, Deming 2009, Chetty et al 2011, Deming 2011, Deming, Hastings, Kane and Staiger 2014).

A few studies have examined the impact of accountability in Texas on high school dropout, with inconclusive findings (e.g. Haney 2000, Carnoy, Loeb and Smith 2001, Mcneil et al 2008, Vasquez Heilig and Darling-Hammond 2008). To our knowledge, only two studies look at the long-term impact of school accountability on postsecondary outcomes. Wong (2008) compares the earnings of cohorts with differential exposure to school accountability across states and over time using the Census and ACS, and finds inconsistent impacts. Donovan, Figlio and Rush (2006) find that minimum competency accountability systems reduce college performance among high-achieving students, but that more demanding accountability systems improve college performance in mathematics courses. Neither of these studies asks whether schools that respond to accountability pressure by increasing students' test scores also make those students more likely to attend and complete college, to earn more as adults, or to benefit over the long-run in other important ways.

II. Data

The Texas Schools Project (TSP) at the University of Texas-Dallas maintains administrative records for every student who has attended a public school in the state of Texas. Students are tracked longitudinally from pre-kindergarten through 12th grade with a unique student identifier. From 1994 to 2003, state exams were referred to as the Texas Assessment of Academic Skills (TAAS). Students were tested in reading and math in grades 3 through 8 and again in grade 10, with writing exams also administered in grades 4, 8 and 10. Raw test scores were scaled using the Texas Learning Index (TLI), which was intended to facilitate comparisons across test administrations. For each year and grade, students are considered to have passed the exam if they reach a TLI score of 70 or greater. As we discuss in more detail in the next section, schools were rated based on the percentage of students who receive a passing score. After each exam, the test questions are released to the public, and the content of the TAAS remained mostly unchanged from 1994 to 1999 (e.g. Klein et al 2000).

Our analysis sample consists of five cohorts of first-time 9th grade students from Spring 1995 to Spring 1999. The TSP data begin in the 1993-1994 school year, and we need 8th grade test scores for our analysis. The first cohort with 8th grade scores began in the 1994-1995 school year. Our last cohort began high school in 1998-1999 and took the 10th grade exam in 1999-2000. We use these five cohorts because Texas' accountability system was relatively stable between 1994 and 1999, and because long-run outcome data are unavailable for later cohorts.

We assign students to a cohort based on the first time they enter 9th grade. We assign them to the first school that lists them in the six week enrollment records provided to the TEA. Prior work has documented the many ways that schools in Texas could manipulate the pool of "accountable" students to achieve a higher rating (Haney 2000, McNeil and Valenzuela 2001, Cullen and Reback 2006, Jennings and Beveridge 2009). Our solution is to assign students to the first high school they attend and to measure outcomes based on initial assignment. For example, if a student attends School A in 9th grade, transfers to School B in 10th grade and then graduates, she is counted as graduating from School A. This is similar in spirit to an "intent to treat" design.

High school students were required to pass each of the 10th grade exams to graduate from high school. The mathematics content on the TAAS exit exam was relatively basic – one analysis found that it was at approximately an 8th grade level compared to national standards (Stotsky 1999). Students who passed the TAAS exit exam in mathematics often struggled to pass end-of-course exams in Algebra I (e.g. Haney 2000). Although students were allowed to retake the 10th grade exit exams up to eight times, we use the first score only in our analysis. We also create an indicator variable equal to one if a student first took the test at the usual time for their 9th grade cohort. This helps us test for the possibility that schools might increase pass rates by strategically retaining, exempting or reclassifying students. Since the TSP data cover the entire state, we can measure graduation from any public school in the state of Texas, even if a student transfers several times, but we cannot track students who left the state.

The TSP links PK-12 records to postsecondary attendance and graduation data from the Texas Higher Education Coordinating Board (THECB). The THECB data contain records of enrollment, course-taking and matriculation for all students who attended public institutions in the state of Texas. While the TSP data do not contain information about out-of-state college enrollment, less than 9 percent of graduating seniors in Texas who attend college do so out of state, and they are mostly high-scoring students.[2] Our main postsecondary outcomes are whether the student ever attended a four-year college or received a bachelor's degree from any public or private institution in Texas.[3]

The TSP has also linked PK-12 records to quarterly earnings data for 1990-2010 from the Texas Workforce Commission (TWC). The TWC data covers wage earnings for nearly all formal employment. Importantly, students who drop out of high school prior to graduation are covered in the TWC data, as long as they are employed in the state of Texas. Our main outcomes of interest here are annual earnings in the age 23-25 years (i.e. the full calendar years that begin 9 to 11 years after the student's first year in 9th grade). Since the earnings data are available through 2010, we can measure earnings in the age 25 year for the 1995 through 1999 9th grade cohorts. We also construct indicator variables for having any positive earnings in the age 19-25 years and over the seven years combined. Zero positive earnings could indicate a true lack of labor force participation, having UI-ineligible earnings, or employment in another state.

[2] Authors' calculation based on a match of 2 graduating classes (2008 and 2009) in the TSP data to the National Student Clearinghouse (NSC), a nationwide database of college attendance.
[3] Our youngest cohort of students (9th graders in Spring 2001) had 7 years after their expected high school graduation date to attend college and complete a BA. While a small number of students in the earlier cohorts received a BA after year 7, almost none attended a four-year college for the first time after 7 years.

Table 1 presents descriptive statistics for our overall analysis sample, and by race and 8th grade test scores. The sample is about 14 percent African-American and 34 percent Latino. 38 percent of students are eligible for free or reduced price lunch (meaning their family income is less than 185 percent of the Federal poverty line). About 76 percent of all students, 59 percent of blacks and 67 percent of Latinos pass the 10th grade math exam on the first try (roughly 20 months after entering 9th grade). There is a strong relationship between 8th grade and 10th grade pass rates. Only 40 percent of students who failed an 8th grade exam passed the 10th grade math exam on the first attempt, and only 62 percent ever passed the 10th grade math exam. In contrast, over 90 percent of students who passed both of their 8th grade exams also passed the 10th grade math exam, almost always on the first attempt.

III. Policy Context

Figure 2 summarizes the key Texas accountability indicators and standards from 1995 to 2002. Schools were grouped into one of four possible performance categories – Low-Performing, Acceptable, Recognized and Exemplary. Schools and districts were assigned performance grades based on the overall share of students that passed TAAS exams in reading, writing and mathematics, as well as attendance and high school dropout. Indicators were also calculated separately for 4 key subgroups - White, African-American, Hispanic, and Economically Disadvantaged (based on the Federal free lunch eligibility standard for poverty) – but only if the group constituted at least 10 percent of the school's population.

Beginning in 1995, schools received the overall rating ascribed to their lowest performing indicator-subgroup combination. This meant that high schools could be held accountable for as many as 20 total performance indicators (5 measures by 4 subgroups). The TAAS passing standard for a school to receive an Acceptable rating rose by 5 percentage points every year, from 25 percent in 1995 to 50 percent in 2000. The standard for a Recognized rating also rose, from 70 percent in 1995 and 1996 to 75 percent in

1997, and 80 percent from 1998 onward. In contrast, the dropout and attendance rate standards remained constant over the period we study.

The details of the rating system meant that math scores were almost always the main obstacle to improving a school's rating. The lowest subgroup-indicator was a math score in over 90 percent of cases. Since schools received a rating based on the lowest scoring subgroup, racially and economically diverse schools often faced significant accountability pressure even if they had high overall pass rates.[4]

Schools had strong incentives to respond to accountability pressure. School ratings were made public, published in full page spreads in local newspapers, and displayed prominently inside and outside of school buildings (Haney 2000, Cullen and Reback 2006). Schools were required to give to each parent a standardized report card which included the school's overall rating and TAAS performance overall and by subgroup (Izumi and Evers 2002). School accountability ratings have been shown to affect property values and private donations to schools (Figlio and Lucas 2004, Figlio and Kenny 2009, Imberman and Lovenheim 2013). Additionally, school districts received an accountability rating based on their lowest-rated school – thus Low-Performing schools faced informal pressure to improve from the district-wide bureaucracy. A TEA-sponsored survey of school and district administrators found that principals perceived their job security as tied directly to the school's rating, with several principals indicating that they would not have their contracts renewed if their school failed to receive a high rating (Toenjes and Garst 2000).

Schools rated as Low-Performing were also forced to undergo an evaluation process that carried potentially serious consequences, such as allowing students to transfer out, firing school leadership, and reconstituting or closing the school (TEA 1994, Cullen and Reback 2006). Although the most punitive

[4] Appendix Table A1 presents descriptive statistics for high schools by the accountability ratings they received over our sample period. Appendix Figure A1 displays the importance of subgroup pressure by plotting each school's overall pass rate on the 10th grade math exam against the math rate for the lowest-scoring subgroup in that school, for the 1995 and 1999 cohorts.

sanctions were rarely used, surveys of principals and teachers indicate that threat of dismissal or transfer for failing to achieve a particular rating was more common (Toenjes and Garst 2000, Evers and Walberg 2002, Lemons, Luschei and Siskin 2003, Mintrop and Trujillo 2005). Schools receiving high ratings were eligible for cash bonuses of up to $5,000 per school, and higher rated schools did indeed receive additional funding as a performance incentive (Izumi and Evers 2002, Craig, Imberman and Perdue 2013).

The TEA did not provide additional funding for low-performing schools (Izumi and Evers 2002). However, regional education service centers (run by the TEA) were encouraged to contact low-performing schools and could offer various forms of assistance such as data analysis, visits from management teams and additional instructional staff in some cases (Izumi and Evers 2002). However, these services were formally available to all schools upon request (Izumi and Evers 2002). In some cases schools that had previously received a low-performing rating were targeted with modest external improvement efforts, such as management teams sent from the district office and focused remediation outside of school hours (Scheurich, Skrla and Johnson 2000, Evers and Walberg 2002, Lemons, Luschei and Siskin 2003).

The Texas accountability system was in many ways the template for the Federal No Child Left Behind Act of 2002. NCLB incorporated most of the main features of the Texas system, including reporting and rating schools based on exam pass rates, reporting requirements and increased focus on performance among poor and minority students, and rising standards over time.

IV. Measuring Accountability Pressure

Figure 1 shows that test scores rose rapidly in Texas after the introduction of school accountability. Did the "Texas Miracle" represent a real gain in student learning? A careful analysis of TAAS content across years found that the test content got progressively easier from 1995 to 1998 (Stotsky 1999).

Since the focus of our study is on long-run outcomes, we first examine descriptive evidence of trends in four-year college attendance and earnings at age 25 for the five cohorts of first-time 9th grade students in Texas included in our study. Appendix Figures A2 and A3 show that college attainment and earnings rose modestly for successive cohorts following the introduction of school accountability.[5]

However, the secular increase in postsecondary attainment and earnings in Texas could be due to other factors besides school accountability. An ideal experiment would randomly assign schools to test-based accountability, and then observe the resulting changes in test scores and long-run outcomes such as attainment and earnings. However, because of the rapid rollout of high-stakes testing in Texas and (later) nationwide, such an experiment is not possible, at least in the U.S. context. Unfortunately, data limitations preclude us from looking at prior cohorts of students who were not part of the high-stakes testing regime.

We aim to isolate the causal impact of accountability pressure by using quasi-experimental variation in the *relative degree of pressure* felt by some grade cohorts within a school over time. Using the full analysis sample, we estimate by logistic regression the probability that each student passes each 10th grade exit exam as follows:

$$Pr[I(Pass\ 10th\ grade\ exam)]_{ijsc}^{t} = \beta X_{ijsc} + \gamma_c + \varepsilon_{isc} \quad (1)$$

The X vector includes demographic characteristics fully interacted with a third order polynomial in 8th grade reading and math scores for student i in school j, subgroup s and cohort c. Equation (1) also includes cohort fixed effects γ_c, which account for yearly changes in test difficulty or any other common cohort shock. We estimate equation (1) separately by test t.

[5] An exception to this pattern is the decline in earnings during 2009-2010, which probably reflects the impact of the Great Recession.

We aggregate the individual predictions up to the school-subgroup-test level to estimate the "risk" that schools will receive a particular rating. [6] The prediction proceeds in three steps. First, we use the predicted values from the student level regressions in equation (1) to form mean pass rates and standard errors at the school-subgroup-test level, i.e. $\overline{PassRate_{js}^{t}}$.

Second, we integrate over the mean pass rates and standard errors to get predicted accountability ratings for each subgroup, school and test. For example, if the predicted pass rate for white students in school A on the math exam is 35 percent with a standard error of 2.5 percent, our model would predict the probability of receiving an "Acceptable" rating as 50 percent in 1997 (since the threshold was at exactly 35 percent) but only about 5 percent in 1998 (since the threshold increased to 40 percent, which is two standard deviations above the mean).

Third, since Texas' accountability rating system specifies an overall school rating that is based on the lowest subgroup-test pass rate, the probability that a school receives a rating of Acceptable or higher (and likewise for other ratings) is equal to the probability that *every eligible subgroup* rates Acceptable or higher on each test. [7] Thus we simply multiply the probabilities for each subgroup and test together to get the probability that school j in cohort c receives a particular rating.

There are two sources of variation in perceived accountability pressure within schools over time - 1) changes over time in the ratings thresholds shown in Figure 2, and 2) changes in the demographics and

[6] Appendix Figure A4 compares our predicted ratings to the actual ratings received by each school in each year. Among schools in the highest risk quintile for a Low-Performing rating, about 40 percent actually receive the Low-Performing rating, and this share declines smoothly as the predicted probability decreases.

[7] Formally, $Pr(Rating \geq Acceptable)_{jc} = \prod_{s=1}^{S} \prod_{t=1}^{T} \Pr(Rating \geq Acceptable)_{jstc}$. Consider the following example for a particular high school. Based on the predicted pass rates on the 10th grade mathematics exam in math, reading and writing for each of the 4 rated subgroups, we calculate that white students have a 96.3 percent chance of receiving an A rating and a 3.7 percent chance of receiving an R rating. Black students have an 18.8 percent chance of receiving an LP rating and an 81.2 percent chance of receiving an A rating. Latinos have a 4.7 percent chance of receiving an LP rating and a 95.3 percent chance for an A rating. Economically Disadvantaged students have an 11.3 percent chance of receiving an LP rating and an 88.7 percent chance for an A rating. Since only whites have any chance of getting an R, and the rating is based on the lowest rated subgroup and test, the probability of getting an R is zero. The probability of an A rating is equal to the probability that all subgroups rate A or higher, which is (0.963+0.037)*(0.812)*(0.953)*(0.887) = 0.766. The probability of an LP rating is equal to 1 minus the summed probabilities of receiving each higher rating, which in this case is 1-0.766 = 0.234. This calculation is conducted separately for all 3 tests to arrive at an overall probability, although in almost all cases math is the only relevant test since math scores are so much lower than reading and writing.

prior test scores of a school's incoming grade cohort, which may have altered the school's incentives to focus on particular subgroups.

However, cohort characteristics may have changed endogenously over time in response to accountability pressure and school performance. For example, a low accountability rating in earlier years may affect subsequent cohorts' high school enrollment decisions. For this reason, we initially compute a single average prediction across all five cohorts. We then allow the ratings thresholds to vary around this single prediction, which isolates policy variation in accountability pressure.

In principle, we could also hold student characteristics constant by computing the prediction using the demographic information from the first cohort only. Our results are very similar but also less precise when we adopt this approach, because the prediction sample is only 20 percent as large.

One limitation of computing a single prediction across cohorts is that it discards potentially useful variation, such as whether a particular subgroup is large enough to count toward the rating. Moreover, there is much less yearly variation along the Acceptable/Recognized rating threshold. Thus we also present results that employ separate risk predictions by cohort (formally, we compute $\overline{PassRate_{jsc}^{t}}$ rather than $\overline{PassRate_{js}^{t}}$ in step 1 above). The bottom line is that our results are not sensitive to a variety of reasonable approaches to measuring accountability pressure.

Our approach is similar in spirit to Reback, Rockoff and Schwartz (2014), who compare students across schools that faced differential accountability pressure because of variation in state standards. We follow their approach in constructing subgroup and subject specific pass rate predictions based on measures of prior achievement. Several papers have studied the impact of *actually receiving* a low school rating, in a regression discontinuity (RD) framework (e.g. Figlio and Lucas 2004, Chiang 2009, Rockoff and Turner 2010, Rouse, Hannaway, Goldhaber and Figlio 2013). Our approach focuses on the much larger group of schools that feel pressure to avoid a Low-Performing rating.

V. Results

V.1 Event Study Using Policy Variation

For an initial graphical examination of accountability pressure, we align each school's predicted pass rates with the ratings threshold in an event study framework. Many schools, particularly in the early years, have a predicted pass rate that is far above the Low-Performing threshold – formally, their "risk" of being rated Low-Performing (according to the estimation procedure above) approaches zero. Depending on each school's average 8th grade test scores and demographic characteristics, the model predicts that they will have some positive probability of being rated Low-Performing beginning in a particular year. Because the policy threshold for a Low-Performing rating only rises over time (see Figure 2) and the prediction does not vary by cohort, once a school is "at risk" it remains so in subsequent cohorts. We organize schools according to the first year they have a positive probability of being "at risk" and estimate:[8]

$$Y_{isc} = \sum_{c=-4}^{4} \delta_{sc} I[Cohort\ C, School\ S] + \beta X_{isc} + \gamma_c + \eta_s + \varepsilon_{isc} \qquad (2)$$

The X vector includes the same covariates as equation (1) above. However, in this specification we have added school fixed effects (η_s) to account for persistent differences across schools in unobserved factors such as parental education, wealth, or school effectiveness. Intuitively, we ask whether the school-specific trend in outcomes varies systematically around the first year that a school was "at risk" of being rated Low-Performing. Because we only have 5 cohorts, the panel is unbalanced for any individual school. However, by controlling for cohort fixed effects (γ_c), we can obtain estimates for up to four years before and after the first year a school was "at risk" of being rated Low-Performing. Since

[8] In Appendix Table A2 we allow the impacts to vary by tercile of predicted risk (1 to 33 percent, 34 to 66 percent, 67 to 100 percent) and find no meaningful difference.

our main independent variables are nonlinear functions of generated regressors, we adjust the standard errors by block bootstrapping at the school level here and for the remainder of the paper.[9]

Figures 3 through 5 present results from equation (2) for the three key outcomes in the paper – 10th grade math pass rates, four-year college attendance, and earnings in the 11th calendar year after the students 9th grade cohort, which we refer to from here onward as the “age 25” year. Estimates for each cohort include 95 percent confidence intervals, with the last year a school is “safe” as the baseline.

Figure 3 shows that students in the same school and with similar prior characteristics are about 2 percentage points more likely to pass the math exam on time (defined as the year after the first time a student enters 9th grade) if their grade cohort is the first to be “at risk”. This difference is statistically significant at the 95 percent level. However, we also find evidence of pre-trends in math pass rates – the difference between 2 years and 1 year prior to being “at risk” is also statistically significant.

This result appears puzzling at first glance, since schools were being rated based on student pass rates on the 10th grade exam. However, the estimated impact in Figure 3 is net of strategic responses such as grade retention and special education classification that alter the test-taking pool. Prior studies of accountability in Texas have shown that schools boosted their ratings by delaying grade progression or strategically exempting students from the test (e.g. Haney 2000, McNeil and Valenzuela 2001, Cullen and Reback 2006). In the next section, we will examine strategic responses directly.

The broader point is that such strategic responses would result in lower performance on the measure in Figure 3 – passing the 10th grade exam “on time”. Since strategic responses are endogenous and affect who takes the test, it is not possible for us to construct a single measure of “true” achievement for all affected students.

Our main interest is in long-run outcomes, which are less easily manipulated. Figure 4 presents results from equation (2) for four-year college attendance. Students in the first grade cohort “at risk” are

[9] Estimates that use the parametric Murphy-Topel (1985) adjustment or no adjustment are very similar to the main results.

about 0.9 percentage points more likely to attend a four-year college within 8 years of the first time they enter 9th grade, and the difference is statistically significant at the 5 percent level. Moreover, we see no significant evidence of pre-trends. We also see that the impact on four-year college attendance continues to rise for subsequent cohorts (who are also "at risk").

The pattern is very similar for earnings – Figure 5 shows that students in the first cohort "at risk" earn about $300 more at age 25 (this estimate is significant at the 10 percent level), and the impact rises gradually over time with no evidence of pre-trends. Thus it appears that the pressure to avoid a Low-Performing rating led to gains in postsecondary attainment and earnings for students in Texas. Note that point estimates are always less precise for years farther away from the last year a school is "safe". This is because of the unbalanced nature of the panel – with only five cohorts, estimates at either end are identified using fewer years of data.[10]

V.2 Regression Results Using Policy Variation

Table 2 presents regression results from a specification that pools all "at risk" grade cohorts together, producing estimates that rely only on policy changes for the relevant variation. We estimate:

$$Y_{isc} = \delta I[pr(LP)_{sc} > 0] + \theta I[pr(R)_{sc} > 0] + \beta X_{isc} + \gamma_c + \eta_s + \varepsilon_{isc} \quad (3)$$

In this setup, grade cohorts that are "safe" (i.e. the probability of being rated Acceptable rounds up to 100 percent) are the omitted category. Equation (3) also allows us to jointly estimate results for schools at risk of both types of ratings (Low-Performing and Recognized). We do not have enough power to estimate results for the small number of schools on the margin between a Recognized and Exemplary rating.

[10] We attempted to construct a similar event study analysis for schools on the margin between an Acceptable and Recognized rating. However, the passing standard for Recognized exhibits much less variation over time, rendering our estimates too imprecise to draw any firm conclusions.

The results for schools at risk of being rated Low-Performing are generally similar to what we find in the event study graphs. There are two key differences between the event study models and the regression models. First, the regression results allow schools to switch back to being “safe” after being “at risk” in an earlier year. If the impact of accountability pressure in a particular year persists for future cohorts, as Figures 3 through 5 suggest, the regression setup will understate the impact on subsequent cohorts. The second key difference is that the regression results allow us to jointly estimate the impact of accountability pressure along both margins. Over the five cohorts in our analysis sample, some schools shift from being “at risk” of Low-Performing to “at risk” of Recognized, and the regression results allow for this variation.

Table 2 shows that students in grade cohorts that were at risk of being rated Low-Performing were about 0.8 percentage points more likely to pass the 10th grade math exam on time (Column 1), and scored about 0.3 scale score points (about 0.05 SDs) higher overall (Column 2). We also find statistically significant increases in the probability of four-year college attendance (0.6 percentage points, Column 3) and receipt of a bachelor’s degree by age 25 (0.37 percentage points, Column 4). The impact on earnings is positive but not statistically significant. In contrast, we find no significant impacts of accountability pressure to achieve a Recognized rating.

Since the accountability metric is based on pass rates, schools had strong incentives to focus on lower-achieving students. One reliable predictor of low high school achievement is whether a student failed an 8th grade exam (e.g. Izumi and Evers 2002). In Panel B we present results that allow the impact to of accountability pressure to vary by whether a student failed either 8th grade exam.

We find that all of the gains from accountability pressure to avoid a Low-Performing rating are concentrated among students who previously failed an exam. These students are about 4.7 percentage points more likely to pass the math exam (Column 1), and they score about 1.3 scale score points (0.2

SDs) higher on the exam overall. More importantly, they are significantly more likely to attend a four-year college (1.9 percentage points, Column 3) and earn a bachelor's degree (1.27 percentage points, Column 4). These impacts, while small in absolute terms, represent about 19 and 30 percent of the mean for students who previously failed an 8th grade exam. We also find that they earn about $298 more at age 25, and that impact is statistically significant at the 5 percent level.

In contrast, we find statistically significant *negative* long-run impacts for low-scoring students in grade cohorts that face pressure to achieve a Recognized rating. Students who previously failed an exam are about 1.8 percentage points less likely to graduate from a four-year college and 0.7 percentage points less likely to earn a bachelor's degree, and they earn $748 less at age 25. We find no impacts of either type of accountability pressure on higher-achieving students.

V.3 Regression Results Using All Cohort Variation

While using only policy variation is the cleanest and most transparent approach, it also throws out some potentially useful variation. Schools naturally vary in the demographics and prior test scores of their incoming students, and this natural variation is likely to also affect the school's perceived risk. This is particularly true when certain subgroups within a school fluctuate around the minimum size requirement of 10 percent of the cohort. In some cases, whether a group "counts" makes a large difference in the probability that a school will receive a Low-Performing or Recognized rating.

To make use of cohort variation in prior characteristics, we estimate equation (1) again but with separate predictions for each school and cohort. This allows for much more flexibility in schools' perceptions of accountability pressure over time – for example, a school may be at risk initially because of a particular subgroup, then switch to "safe" because the group becomes too small in subsequent cohorts.[11]

[11] We follow the minimum size requirements outlined by accountability policy and exclude subgroups that are less than 10 percent of the 9th grade cohort in this calculation. We also incorporate into the model a provision known as Required

Table 3 presents results from equation (3), estimated using this new set of risk predictions. Overall, the results are very similar to the model in Table 2 that uses only policy variation. There are two main differences. First, while the overall impact of accountability pressure to avoid a Low-Performing rating is very similar, the impacts in Table 3 are more evenly distributed across lower and higher-achieving students. Second, in schools that faced pressure to achieve a Recognized rating, the negative impact of accountability pressure on the postsecondary attainment of low-achieving students is considerably higher.

Some schools would be at risk of being rated Low-Performing or Recognized because of a particular subgroup, but are actually "safe" because that subgroup is too small to count toward the rating. Thus the minimum subgroup size requirement provides us with a useful placebo test. In Appendix Table A5 we show that estimated impacts for placebo subgroups are near zero and statistically significantly smaller than subgroups that are truly "at risk".[12]

V.4 Robustness Checks

One potential concern is that the relationship between 10th grade scores and 8th grade characteristics is contaminated by endogenous responses to perceived risk. Concretely, if the prediction model in equation (1) used an identical set of covariates as equation (2), our estimates would be identified purely from functional form. However, the timing of perceived risk is a function of policy variation that is not in the prediction model. As a check on the endogeneity of the prediction model, in Appendix Table A6 we simply allow impacts to vary by the 8th grade pass rate of the lowest-scoring subgroup in a school, rather than estimating risk directly.[13]

Improvement, which allows schools to avoid receiving a Low-Performing rating if the year-to-year increase in the pass rate was large enough to put them on pace to reach a target of 50 percent within 5 years. Appendix Table A4 presents a transition matrix that shows the relationship between schools' predicted ratings in year T and year T+1.

[12] We select 8 percent as the placebo because schools face some uncertainty around the threshold, which is based on 10th grade cohorts rather than first-time 9th graders.

[13] The results in Table A6 are obtained by calculating the share of students in an incoming high school cohort who passed the 8th grade exam for all test-subgroup combinations (e.g. Latinos in reading, blacks in math, etc.) We then take the difference

Another concern is that the timing of a school's predicted rating is correlated with other contemporaneous shocks that might also affect long-run outcomes. We test for the possibility of contemporaneous shocks in Appendix Table A6 by regressing a school's predicted risk of being rated Low-Performing on time-varying high school inputs such as principal turnover, teacher pay and teacher experience.[14]

Our data only cover postsecondary attendance and employment in the state of Texas. Hence our estimates would be biased if accountability pressure increases out-of-state migration, particularly if out-of-state migrants are more likely to attend and graduate from college and have higher earnings. In Appendix Tables A9 and A10, we find that our results are robust to imputing missing earnings values and to separately estimating results for schools that send large shares of students out-of-state.

We also measure possible attrition directly by constructing an indicator variable that is equal to one if a student has zero earnings and never attends any college between the ages of 19 and 25. This provides an upper bound on students who left the state and did not return (incarcerated or deceased students would have a value of zero, for example). In Table 1, we see that the mean of this variable is 13 percent for the full sample.[15] When we estimate the impact of accountability pressure on this indicator for possible attrition, the estimate is -0.001 with a standard error of 0.002 for Low-Performing and 0.004 (0.003) for Recognized. Thus there is no evidence of differential attrition, and our standard errors allow us to rule out all but very small impacts.

---

between the minimum 8th grade test-subgroup pass rate for each cohort and the threshold for an Acceptable rating when that cohort takes the TAAS two years later, in 10th grade, and divide schools into bins based on their relative distance from the yearly threshold. In this approach, there is no mean reversion or correlated estimation error, because we do not estimate anything.

[14] Appendix Table A7 conducts a similar exercise using a linear trend interacted with overall and subgroup-specific pass rates going back to 1991, three years prior to the beginning of school accountability in Texas. While high school inputs and test score trends are strong predictors of accountability ratings across schools, they have little predictive power across cohorts within the same school, once we account for 8th grade test scores and year fixed effects.

[15] Data from the 2000 Census indicate that only 8 percent of youths age 14-18 who were enrolled in school (not college) in Texas were living in another state or country 5 years ago. Among blacks and Latinos those figures are 6.2 and 7.8 percent respectively. Moreover, out-of-state college attendance is relatively rare. Only 10.3 percent of all undergraduates ages 19-21 who lived in Texas 5 years earlier were enrolled in colleges outside of Texas.

Our empirical strategy sometimes compares students who are only one or two grades apart in the same school. If accountability pressure causes schools to shift resources toward some students at the expense of others (e.g. Reback 2008), comparisons across cohorts may be problematic. In Appendix Tables A11 and A12 we therefore restrict our analysis to 1) non-consecutive cohorts (i.e. 1995, 1997 and 1999) and 2) non-overlapping cohorts (i.e. 1995 and 1999). In the latter case, students who progressed through high school "on time" and in four years would never be in the building together. Our results are robust to these sample restrictions.

VI. What explains the pattern of results?

The theoretical literature on incentive design and multi-task moral hazard predicts that high-stakes testing will cause teachers and schools to adjust their effort toward the least costly (in terms of dollars or effort) way of increasing test scores, possibly at the expense of other salutary actions (Holmstrom and Milgrom 1991, Baker 1992). Thus one natural way to try to understand the difference in impacts along the two ratings thresholds is to ask – what was the least costly method of achieving a higher rating?

In our data, schools "at risk" of being rated Low-Performing were on average 23 percent African-American, 32 percent Latino and 44 percent poor, with a mean cohort size of 212 and a mean pass rate on the 8th grade math exam of 56 percent. Since the overall cohort and each tested subgroup was on average quite large, these schools could only escape a Low-Performing rating through widespread improvement in test performance.

In contrast, schools "at risk" of being rated Recognized were only about 5 percent African-American, 10 percent Latino and 16 percent poor, with a mean cohort size of only 114 and a mean pass rate on the 8th grade math exam of 84 percent. Thus many of these schools could achieve a higher rating by affecting only a small number of students.

Why does this matter? Many of the strategic responses documented in prior work are most effective in small numbers. One example is strategic classification of students in order to influence who "counts" toward the rating. During this period in Texas, special education students were allowed to take the 10^{th} grade TAAS but their scores did not count toward the school's accountability rating. They also were not required to pass the 10^{th} grade exam to graduate (Fuller 2000). Cullen and Reback (2006) find that schools in Texas during this period strategically classified students as eligible for special education services to keep them from lowering the school's accountability rating. It is much easier to strategically exempt or reclassify 5 percent of a grade cohort than 50 percent of a grade cohort.

In Table 5 we provide some evidence on possible mechanisms by estimating results for additional outcomes in high school. The outcome in Column 1 is an indicator for whether a student is receiving special education services in the 10^{th} grade year, *but did not receive special education services in 8^{th} grade.* Panel B of Column 1 shows strong evidence of strategic special education classification in schools that had a chance to achieve a Recognized rating. Low-scoring students in these schools are 2.4 percentage points more likely to be newly designated as eligible for special education, an increase of over 100 percent relative to the baseline mean of 2 percent. We also find a smaller (0.5 percentage points) but still highly significant *decrease* in special education classification for high-scoring students in these schools.

These results provide strong evidence that schools trying to achieve a Recognized rating did so by strategically exempting students from the high-stakes test. In Appendix Table A13, we show that controlling for 10^{th} grade special education status eliminates the negative impacts of pressure to achieve a Recognized rating on low-scoring students, which further suggests a strong mediating role for strategic special education classification. Additionally, in results not shown, we find larger impacts on strategic special education classification and (negatively) on long-run outcomes when fewer students in the cohort had previously failed an 8th grade exam, allowing for greater strategic targeting of particular students.

Column 2 shows results for high school graduation within 8 years of the student's first time entering 9th grade. We find an overall increase in high school graduation of about 1 percentage point in schools that face pressure to avoid a Low-Performing rating. Interestingly, we find an *increase* (significant at the 10 percent level) in high school graduation for low-scoring students in schools that faced pressure to achieve a Recognized rating, despite finding negative long-run impacts on postsecondary attainment and earnings. When we examine results separately by type of diploma (not shown), we find that the increase is driven by special education diplomas (for students who are not required to pass the exit exam). It is possible that marginal students were placed in less-demanding courses and acquired fewer skills.

Finally, Column 3 shows impacts on total math credits accumulated in four state-standardized high school math courses – Algebra I, Geometry, Algebra II and Pre-Calculus. We find an increase of about 0.06 math course credits in schools that face pressure to avoid a Low-Performing rating. We also find a decline of about 0.11 math course credits for students with low baseline scores in schools that were close to achieving a Recognized rating. Both estimates are statistically significant at the less than 1 percent level. In results not reported, we find that the impacts on both math credits and long-run outcomes increase with cohort size and with the number of students who previously failed an 8th grade exam, suggesting that students benefited from accountability pressure when school-wide efforts were necessary.

Increased knowledge of mathematics is a plausible mechanism for long-run impacts on postsecondary attainment and earnings. Using cross-state variation in the timing of high school graduation requirements, Levine and Zimmerman (1995) and Rose and Betts (2004) also find that additional mathematics coursework in high school is associated with increases in labor market earnings.

Cortes, Goodman and Nomi (2015) find increases in high school graduation and college attendance for students who are assigned to a “double dose” Algebra I class in 9th grade.

In Appendix Table A14, we show that controlling for math coursework reduces the estimates of accountability pressure on bachelor’s degree receipt and earnings at age 25 to nearly zero, and lowers the impact on four-year college attendance by about 50 percent. This suggests that increases in math coursework are a key mediator for explaining the long-run impacts of accountability pressure. In Appendix Table A15, which contains results for a number of additional high school outcomes, we show that these increases in math credits extend beyond the requirements of the 10th grade math exit exam, to upper level coursework such as Algebra II and Pre-Calculus.

Did accountability pressure lead to increases in instructional resources devoted to “at risk” students? Appendix Figure A5 presents estimates of the impact of accountability pressure on the allocation of regular classroom and remedial classroom teacher FTEs, using the setup in equation (3). We find some evidence that schools respond to the risk of being rated Low-Performing by increasing staffing, particular in remedial classrooms. Given the across-cohort design, it is most likely that these differences are driven by short-run allocation of floating teachers or tutors rather than permanent staffing changes.

VII. Discussion and Conclusion

Why do some students benefit from accountability pressure while others are harmed? Based on the pattern of results discussed above, we argue that heterogeneous responses to accountability pressure stemmed from schools choosing the path of least resistance. The typical school at risk of receiving a Low-Performing rating was large, majority nonwhite and with many students who had previously failed an 8th grade exam. Thus the scope for strategic classification of particular students as eligible for special education services was quite limited. Students in schools at risk of being rated Low-Performing were more likely to pass the 10^{th} grade math exam on time, acquired more math credits in high school (beyond a 10^{th} grade level), and were more likely to graduate from high school on time. In the long-run, they had higher rates of postsecondary attainment and earnings. These gains were concentrated among students at the greatest risk of failure.

The typical school facing pressure to achieve a Recognized rating, on the other hand, was small and had lower shares of poor and minority students. Because ratings were assigned based on the lowest scoring subgroup, and because special education students were exempt from the ratings calculation, schools faced strong incentives to strategically classify particular students. In these schools, we find that low-scoring students were more than twice as likely to be newly deemed eligible for special education. This designation exempted students from the normal high school graduation requirements, which then led to lower total accumulation of math credits. In the long-run, low-scoring students in schools that faced pressure to achieve a Recognized rating had significantly lower postsecondary attainment and earnings.

We find that accountability pressure to avoid a Low-Performing rating leads to increases in labor market earnings at age 25 of around 1 percent. By comparison, Chetty, Friedman and Rockoff (2014) find that having a teacher in grades 3 through 8 with 1 SD higher "value-added" also increases earnings at age 25 by about 1 percent. Chetty et al (2011) also find that students who are randomly assigned to a kindergarten classroom that is 1 SD higher quality earn nearly 3 percent more at age 27. Our results also line up fairly well with the existing literature on the connection between test score gains and labor market earnings. Neal and Johnson (1996) estimate that high school-age youth who score 0.1 SD higher on the Armed Forces Qualifying Test (AFQT) have 2 percent higher earnings at ages 26-29. Similarly, Currie and Thomas (1999) and Chetty et al (2011) find cross-sectional relationships between test scores at age 5-7 and adult earnings that are similar in size to our results for high school students.

Since accountability policy in Texas was in many ways the template for No Child Left Behind, our findings may have broad applicability to the similarly structured accountability regimes that were rolled out later in other states. However, many states (including Texas itself) have changed their rating systems over time, both by incorporating test score growth models and by limiting the scope for strategic behavior such as special education exemptions. At least in our setting, school accountability was more effective at ensuring a minimum standard of performance than improving performance at a higher level.

References

Allen, R. & Burgess, S., 2012. *How should we treat under-performing schools? A regression discontinuity analysis of school inspections in England*, University of London.

Baker, G.P., 1992. Incentive Contracts and Performance Measurement. *Journal of Political Economy*, 100(3), pp.598–614.

Booher-Jennings, J., 2005. Below the bubble: "Educational triage" and the Texas accountability system. *American Educational Research Journal*, 42(2), pp.231–268.

Booker, K. et al., 2011. The Effects of Charter High Schools on Educational Attainment. *Journal of Labor Economics*, 29(2), pp.377–415.

Campbell, D.T., 1976. *Assessing the impact of planned social change*, Hanover, NH: Dartmouth College, Public Affairs Center.

Carnoy, M. & Loeb, S., 2002. Does external accountability affect student outcomes? A cross-state analysis. *Educational Evaluation and Policy Analysis*, 24(4), pp.305–331.

Carnoy, M., Loeb, S. & Smith, T.L., 2001. Do higher state test scores in Texas make for better high school outcomes. In *American Educational Research Association Annual Meeting (April)*.

Chetty, R., Friedman, J.N., Hilger, N., et al., 2011. How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project Star. *The Quarterly Journal of Economics*, 126(4), pp.1593–1660.

Chetty, R., Friedman, J.N. & Rockoff, J.E., 2014. Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood. *American Economic Review*, 104(9), pp.2633-2679.

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

Cortes, K., Goodman, J. & Nomi, T., 2015. Intensive Math Instruction and Educational Attainment: Long-Run Impacts of Double-Dose Algebra. *Journal of Human Resources*, 50(1), pp.108-158.

Craig, S.G., Imberman, S.A., & Perdue, A. (2013). Does it pay to get an A? School resource allocations in response to accountability ratings. *Journal of Urban Economics*, 73(1), pp.30-42.

Cullen, J.B. & Reback, R., 2006. Tinkering toward accolades: School gaming under a performance accountability system. In T. Gronberg & D. Jansen, eds. *Advances in Applied Microeconomics*. Elsevier.

Dee, T.S. & Jacob, B., 2011. The Impact of No Child Left Behind on Student Achievement. *Journal of Policy Analysis and Management*, 30(3), pp.418–446.

Deming, D., 2009. Early Childhood Intervention and Life-Cycle Skill Development: Evidence from Head Start. *American Economic Journal: Applied Economics*, 1(3), pp.111–134.

Deming, D., Hastings, J.S., Kane, T.J., &S Staiger, D.O, 2014. School choice, school quality and academic achievement. *American Economic Review,* 104(30, pp.991-1013.

Deming, D.J., 2011. Better Schools, Less Crime? *The Quarterly Journal of Economics*, 126(4), pp.2063–2115.

Diamond, J. & Spillane, J., 2004. High-stakes accountability in urban elementary schools: Challenging or reproducing inequality? *The Teachers College Record*, 106(6), pp.1145–1176.

Diamond, J.B., 2007. Where the rubber meets the road: Rethinking the connection between high-stakes testing policy and classroom instruction. *Sociology of Education*, 80(4), pp.285–313.

Donovan, C., Figlio, D.N. & Rush, M., 2006. *Cramming: The effects of school accountability on college-bound students*, Cambridge, MA: National Bureau of Economic Research.

Evers, W.M. & Walberg, H.J., 2002. *School accountability*, Hoover Press.

Figlio, D. & Loeb, S., 2011. School Accountability. In *Handbook of the Economics of Education*. pp. 383–421.

Figlio, D.N. & Getzler, L.S., 2006. Accountability, ability and disability: Gaming the system? In T. Gronberg & D. Jansen, eds. *Advances in Applied Microeconomics*. Elsevier, pp. 35–49.

Figlio, D.N. & Kenny, L.W., 2009. Public sector performance measurement and stakeholder support. *Journal of Public Economics*, 93(9), pp.1069–1077.

Figlio, D.N. & Ladd, H.F., 2008. School accountability and student achievement. In *Handbook of Research in Education Finance and Policy*. pp. 166–182.

Figlio, D.N. & Lucas, M.E., 2004. Whats in a Grade? School Report Cards and the Housing Market. *American Economic Review*, 94(3), pp.591–604.

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

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

Gould, E.D., Lavy, V. & Paserman, M.D., 2004. Immigrating to opportunity: Estimating the effect of school quality using a natural experiment on Ethiopians in Israel. *The Quarterly Journal of Economics*, 119(2), pp.489–526.

Greene, J., Winters, M. & Forster, G., 2004. Testing High-Stakes Tests: Can We Believe the Results of Accountability Tests? *Teachers College Record*, 106(6), pp.1124–1144.

Hamilton, L.S. et al., 2007. *Standards-Based Accountability Under No Child Left Behind: Experiences of Teachers and Administrators in Three States*, Santa Monica, CA: RAND Corporation.

Haney, W., 2000. The Myth of the Texas Miracle in Education. *Education Policy Analysis Archives*, 8(41).

Hanushek, E.A. & Raymond, M.E., 2005. Does school accountability lead to improved student performance? *Journal of Policy Analysis and Management*, 24(2), pp.297–327.

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

Holmstrom, B. & Paul Milgrom, 1991. Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design. *Journal of Law, Economics and Organization*, 7, pp.24–52.

Hout, M. & Elliott, S.W., 2011. *Incentives and test-based accountability in education*, National Academies Press.

Izumi, L. T., & Evers, W. M. (2002). State accountability systems. *School accountability: An assessment by the Koret Task Force on K-12 education*.

Jacob, B.A., 2005. Accountability, incentives and behavior: the impact of high-stakes testing in the Chicago Public Schools. *Journal of Public Economics*, 89(5-6), pp.761–796.

Jacob, B.A. & Levitt, S.D., 2003. Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating. *The Quarterly Journal of Economics*, 118(3), pp.843–877.

Jennings, J.L. & Beveridge, A.A., 2009. How Does Test Exemption Affect Schools' and Students' Academic Performance? *Educational Evaluation and Policy Analysis*, 31(2), pp.153–175.

Kerr, S., 1975. On the Folly of Rewarding A, While Hoping for B. *Academy of Management Journal*, 18(4), pp.769–783.

Klein, S.P. et al., 2000. *What do test scores in Texas tell us?*, Santa Monica, CA: Rand.

Koretz, D.M. & Barron, S.I., 1998. *The Validity of Gains in Scores on the Kentucky Instructional Results Information System (KIRIS).*, Santa Monica, CA: RAND.

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

Lemons, R., Luschel, T., & Siskin, L. (2003). Leadership and the demands for standards-based accountability. *The new accountability: High schools and high-stakes testing*, 99-128.

Levy, F. & Murnane, R.J., 2012. *The new division of labor: How computers are creating the next job market*, Princeton University Press.

McNeil, L. et al., 2008. Avoidable losses: High-stakes accountability and the dropout crisis. *Education Policy Analysis Archives*, 16(3), p.1.

McNeil, L. & Valenzuela, A., 2001. The Harmful Impact of the TAAS System of Testing in Texas: Beneath the Accountability Rhetoric. In *Raising Standards or Raising Barriers? Inequality and HIgh Stakes Testing in Public Education*. New York, NY: Century Foundation, pp. 127–150.

Mintrop, H., & Trujillo, T. (2005). Corrective Action in Low Performing Schools: Lessons for NCLB Implementation from First-generation Accountability Systems. *Education Policy Analysis Aarchives*, *13(28.*

Neal, D. & Schanzenbach, D.W., 2010. Left Behind by Design: Proficiency Counts and Test-Based Accountability. *Review of Economics and Statistics*, 92(2), pp.263–283.

Neal, D.A., 2013. *The consequences of using one assessment system to pursue two objectives*, Cambridge, MA, NBER Working Paper 19214.

Neal, D.A. & Johnson, W.R., 1996. The Role of Premarket Factors in Black-White Wage Differences. *Journal of Political Economy*, 104(5), pp.869–895.

Reback, R., 2008. Teaching to the rating: School accountability and the distribution of student achievement. *Journal of Public Economics*, 92(5), pp.1394–1415.

Reback, R., Rockoff, J. & Schwartz, H.L., 2014. Under pressure: Job security, resource allocation, and productivity in schools under NCLB. *American Economic Journal: Economic Policy*, 6(3), pp.207-241.

Rockoff, J. & Turner, L.J., 2010. Short-Run Impacts of Accountability on School Quality. *American Economic Journal: Economic Policy*, 2(4), pp.119–147.

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

Skrla, L., Scheurich, J.J. & Johnson, J.F., 2000. Equity-driven achievement-focused school districts: A report on systemic school success in four Texas school districts serving diverse student populations. *Austin, TX: Charles A. Dana Center*.

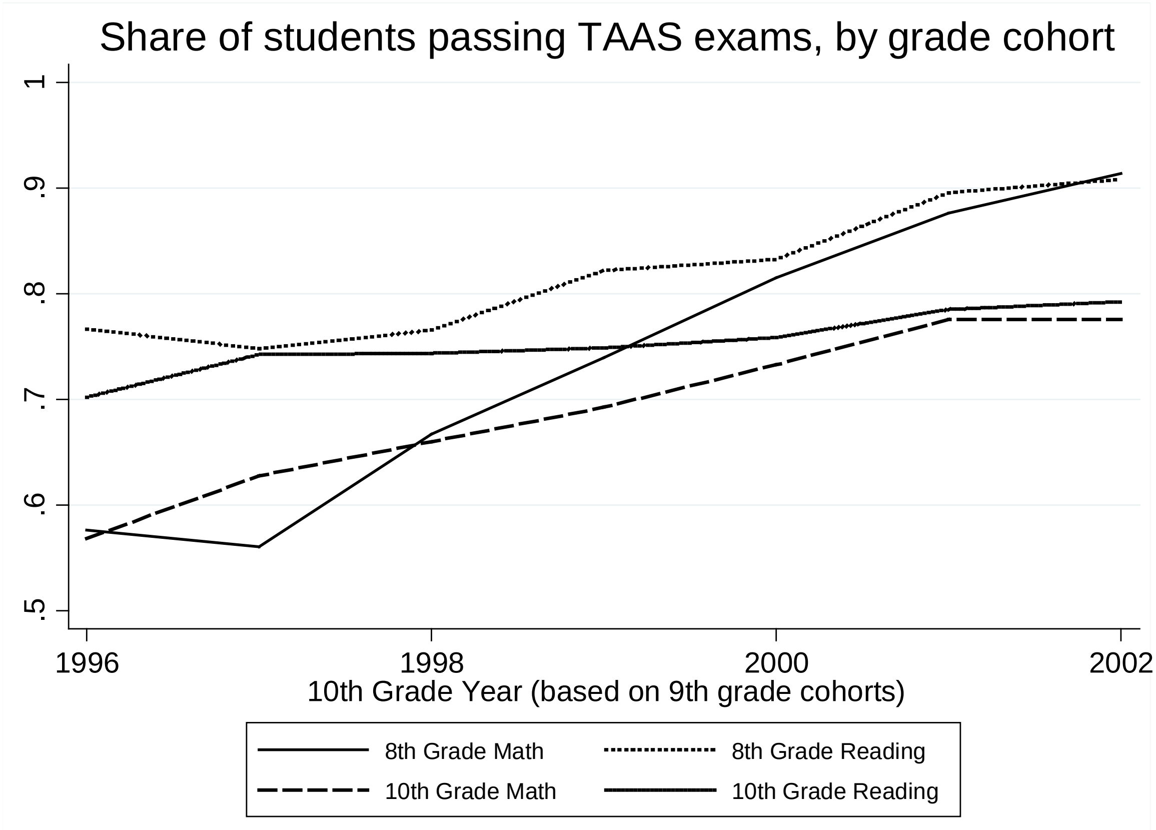
Spillane, J.P., Parise, L.M. & Sherer, J.Z., 2011. Organizational Routines as Coupling Mechanisms Policy, School Administration, and the Technical Core. *American Educational Research Journal*, 48(3), pp.586–619.

Stecher, B.M. et al., 2000. *The Effects of the Washington State Education Reform on Schools and Classrooms*, Santa Monica, CA: RAND Corporation.

Toenjes, L.A. & Garst, J.E. (2000) *Identifying High Performing Texas Schools and School Districts and their Methods of Success.* Texas Education Agency.

Wong, K., 2008. *Looking Beyond Test Score Gains: State Accountability's Effect on Educational Attainment and Labor Market Outcomes*, University of California, Irvine.

Figure 1



Notes: The figure above shows time trends in the share of students in Texas who pass the 8th and 10th grade exams in math and reading. Students are assigned to cohorts based on the first time they enter 9th grade.

Figure 2

ACCOUNTABILITY INDICATORS AND STANDARDS 1995 TO 2002

	1995	1996	1997	1998	1999	2000	2001	2002
TAAS PASSING STANDARD FOR READING, WRITING, AND MATHEMATICS (GR. 3-8, 10) ***[for "all students" and each student group]***								
Exemplary	>=90.0%	>=90.0%	>=90.0%	>=90.0%	>=90.0%	>=90.0%	>=90.0%	>=90.0%
Recognized	>=70.0%	>=70.0%	>=75.0%	>=80.0%	>=80.0%	>=80.0%	>=80.0%	>=80.0%
Academically Acceptable * */ Acceptable*	>= 25.0%	>= 30.0%	>= 35.0%	>= 40.0%	>= 45.0%	>= 50.0%	>= 50.0%	>= 55.0**%
Academically Unacceptable * */ Low-performing*	< 25.0%	<30.0%	<35.0%	<40.0%	<45.0%	<50.0%	<50.0%	<55.0%**
DROPOUT RATE STANDARDS (GR. 7-12) ***[for all students and each student group]***								
Exemplary	<=1.0%	<=1.0%	<=1.0%	<=1.0%	<=1.0%	<=1.0%	<=1.0%	<=1.0%
Recognized	<=3.5%	<=3.5%	<=3.5%	<=3.5%	<=3.5%	<=3.5%	<=3.0%	<=2.5%
Academically Acceptable * */ Acceptable*	n / a	< = 6.0%	< = 6.0%	< = 6.0%	< = 6.0%	< = 6.0%	< = 5.5%	< = 5.0%
Academically Unacceptable * */ Low-performing*	n / a	>6.0% ☆	>6.0% ☆	>6.0% ☆	>6.0% ☆	>6.0% ☆	>5.5% ☆	>5.0% ☆
ATTENDANCE RATE STANDARD (GR. 1-12) †	>=94.0%	>=94.0%	>=94.0%	>=94.0%	>=94.0%	>=94.0%	n / a	n / a
AT WHAT LEVELS OF PERFORMANCE REQUIRED IMPROVEMENT IS ANALYZED ***[for all students and each student group]***								
To Be Rated *Recognized:* *TAAS Reading, Mathematics, and Writing*	70.0% - 79.9%	70.0% - 79.9%	75.0% - 79.9%	n / a	n / a	n / a	n / a	n / a
To Avoid *Academically Unacceptable / Low-performing*								
TAAS Reading, Mathematics, and Writing	< 25.0%	< 30.0%	< 35.0%	< 40.0%	< 45.0%	n / a	n / a	n / a
Dropout Rate	> 6.0%	> 6.0%	> 6.0%	> 6.0%	> 6.0%	n / a	n / a	n / a

☆ Special conditions for a single dropout rate exceeding the *Acceptable* standard apply.

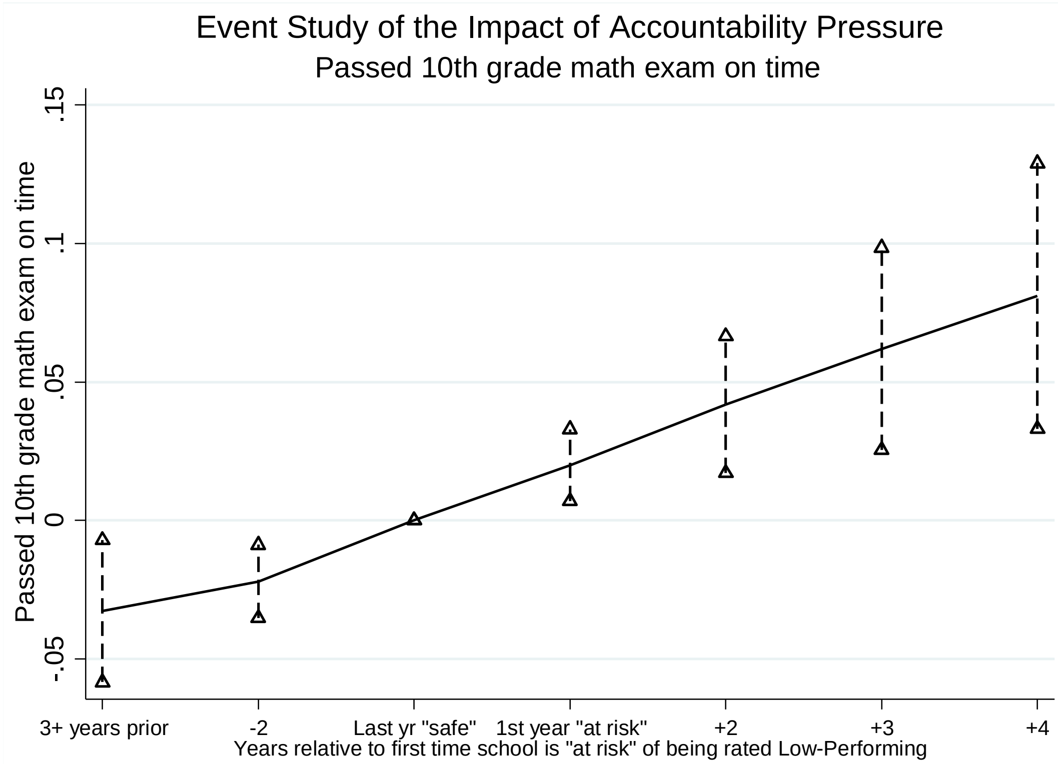
† The attendance rate standard was waived for the *Academically Acceptable / Acceptable* rating if failure to meet that standard would be the sole reason that the school would be *Low-performing* or the district *Academically Unacceptable*.

* In 1995 and 1996, the district ratings used were: *Exemplary, Recognized, Accredited*, and *Accredited Warned*. A statutory change in 1997 resulted in use of the current label.

** Social Studies has been added in 2002. The *Academically Acceptable/Acceptable* accountability for Social Studies in 2002 is >= 50% and for *Academically Unacceptable/Low performing* is <50% for the "all students" level. Social Studies is not evaluated at the student group level in 2002.

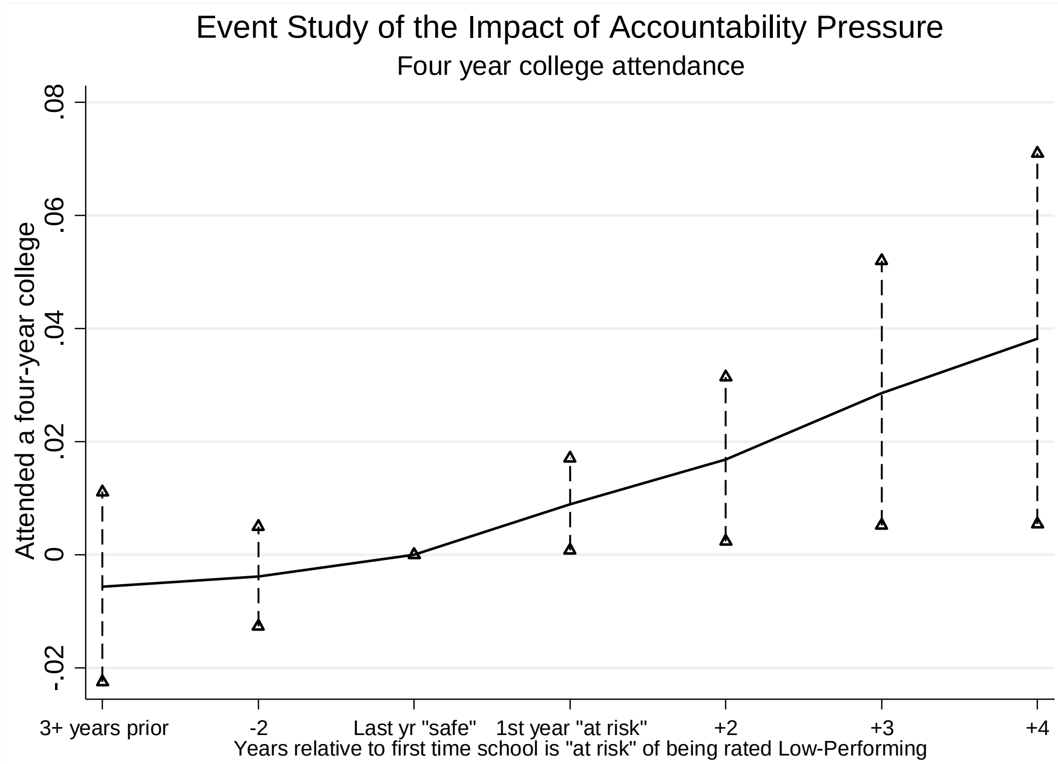
PART 2: Section XIV –Accountability System Facts — *2002 ACCOUNTABILITY MANUAL* 151

Figure 3



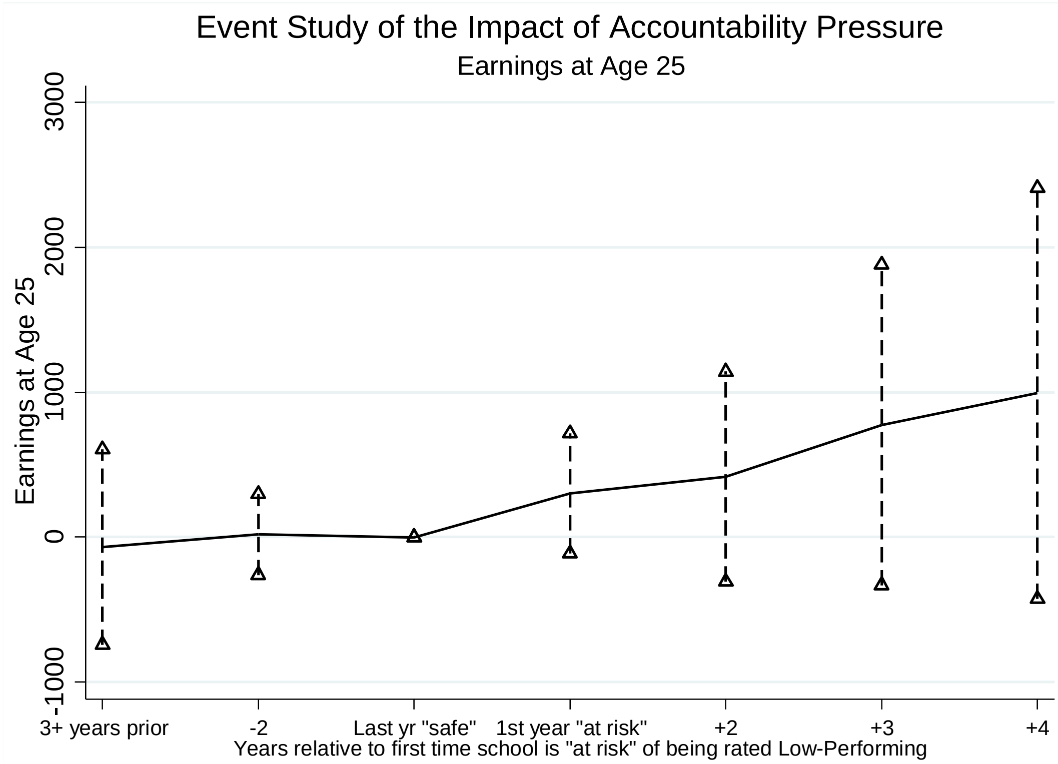
Notes: The figure presents point estimates and 95 percent confidence intervals from equation (2) in the paper, where the outcome is whether a student passed the 10^{th} grade math exam on time (defined as 1 year after the first time a student enters 9^{th} grade). We estimate the "risk" of a high school being rated Low-Performing based on the demographics and 8^{th} grade test scores of the grade cohort, combined with policy variation over time in the passing standards shown in Figure 2 – see the text for details. We then define grade cohorts according to the first year each school was "at risk" of being rated Low-Performing, with the prior year as the baseline category. The regression includes school fixed effects.

Figure 4



Notes: The figure presents point estimates and 95 percent confidence intervals from equation (2) in the paper, where the outcome is whether a student attended a four-year college in Texas within 11 years of the first time they entered 9^{th} grade. We estimate the "risk" of a high school being rated Low-Performing based on the demographics and 8^{th} grade test scores of the grade cohort, combined with policy variation over time in the passing standards shown in Figure 2 – see the text for details. We then define grade cohorts according to the first year each school was "at risk" of being rated Low-Performing, with the prior year as the baseline category. The regression includes school fixed effects.

Figure 5



Notes: The figure presents point estimates and 95 percent confidence intervals from equation (2) in the paper, where the outcome is earnings in the age 25 year, defined as the 11th year after the first time a student enters 9th grade. Students with zero reported earnings are included in the calculation. We estimate the "risk" of a high school being rated Low-Performing based on the demographics and 8th grade test scores of the grade cohort, combined with policy variation over time in the passing standards shown in Figure 2 – see the text for details. We then define grade cohorts according to the first year each school was "at risk" of being rated Low-Performing, with the prior year as the baseline category. The regression includes school fixed effects.

Table 1 - Descriptive Statistics

	Overall	Black	Latino	FRPL	Passed 8th Grade Exams	Failed an 8th Grade Exam
8th grade covariates	(1)	(2)	(3)	(4)	(5)	(6)
White / Other	0.52			0.20	0.64	0.33
Black	0.14			0.19	0.09	0.21
Latino	0.34			0.61	0.27	0.46
Free Lunch	0.38	0.54	0.68		0.29	0.55
Passed 8th math (TLI>=70)	0.67	0.48	0.56	0.53		
Passed 8th reading	0.79	0.66	0.69	0.66		
High school outcomes						
10th grade math score	78.2	72.6	75.6	74.6	83.2	66.3
Passed 10th math on time	0.76	0.59	0.67	0.64	0.90	0.40
Ever Passed 10th math	0.81	0.74	0.76	0.72	0.92	0.62
Passed 10th reading on time	0.88	0.75	0.77	0.75	0.95	0.51
Special Ed in 10th, not 8th	0.01	0.01	0.01	0.01	0.00	0.02
Total Math Credits	1.93	1.78	1.73	1.65	2.29	1.33
Graduated from high school	0.74	0.69	0.69	0.65	0.82	0.59
Later Outcomes						
Attended any college	0.54	0.46	0.45	0.39	0.65	0.35
Attended 4 year college	0.28	0.24	0.19	0.15	0.39	0.10
BA degree	0.13	0.09	0.09	0.07	0.18	0.05
Age 25 Earnings (in 1000s)	17.7	13.6	16.1	14.6	19.8	14.0
No earnings/college, all yrs	0.13	0.17	0.15	0.15	0.12	0.15
Sample Size	887,713	121,508	302,720	339,279	560,872	326,841

Notes: The sample consists of five cohorts of first-time rising 9th graders in public high schools in Texas, from years 1995 to 1999. Postsecondary attendance data include all public institutions and, from 2003 onward, all not-for-profit institutions in the state of Texas. Earnings data are drawn from quarterly unemployment insurance records from the state of Texas. Column 6 shows students who received a passing score on both the 8th grade math and reading exams. Column 7 shows descriptive statistics for students who failed either exam. Students who are first time 9th graders in year T and who pass a 10th grade exam in year T+1 are considered to have passed "on time". Math credits are defined as the sum of indicators for passing Algebra I, Geometry, Algebra II and Pre-calculus, for a total maximum value of four. "Idle" is defined as having zero recorded earnings and no postsecondary enrollment.

Table 2: Impact of Accountability Pressure - only policy variation in the prediction model

	10th Grade Math		Four Year College		Earnings
	Passed Test	Scale Score	Ever Attend	BA	Age 25
Panel A	(1)	(2)	(3)	(4)	(5)
Risk of Low Performing Rating	0.008**	0.300**	0.006*	0.0037**	141
	[0.003]	[0.096]	[0.002]	[0.0013]	[97]
Risk of Recognized Rating	0.006	0.115	-0.007	-0.0028	-232
	[0.004]	[0.132]	[0.004]	[0.0027]	[155]
Panel B					
Risk of Low Performing Rating					
* Failed an 8th grade exam	0.047**	1.362**	0.019**	0.0127**	298*
	[0.005]	[0.147]	[0.002]	[0.0015]	[122]
* Passed 8th grade exams	-0.007*	-0.125	-0.005	-0.0015	76
	[0.003]	[0.092]	[0.003]	[0.0017]	[122]
Risk of Recognized Rating					
* Failed an 8th grade exam	-0.004	-0.117	-0.018**	-0.0070*	-748**
	[0.008]	[0.209]	[0.005]	[0.0032]	[227]
* Passed 8th grade exams	0.008*	0.169	-0.002	-0.0015	112
	[0.004]	[0.128]	[0.005]	[0.0031]	[200]
Sample Size	697,728	697,728	887,713	887,713	887,713

Notes: Each column is a single regression of the indicated outcome on the set of variables from equations (3) in the paper, which includes controls for math and reading scores, demographics, and year and school fixed effects. Standard errors are block bootstrapped at the school level. Each coefficient gives the impact of being in a grade cohort that has a positive estimated risk of being rated Low-Performing or Recognized. We estimate a single risk prediction for each school, thereby using only yearly changes in the passing standard to identify cross-cohort changes in accountability pressure. See the text for details. A one standard deviation change in the math score is equal to about 7 scale score points. College attendance outcomes are measured within an 8 year time window beginning with the student's first-time 9th grade cohort. The outcome in Column 5 is annual earnings in the 11th year after the first time a student enters 9th grade (which we refer to as the age 25 year), including students with zero reported earnings. * = sig. at 5% level; ** = sig. at 1% level or less.

Table 3: Impact of Accountability Pressure - all variation in the prediction model

	10th Grade Math		Four Year College		Earnings
	Passed Test	Scale Score	Ever Attend	BA	Age 25
Panel A	(1)	(2)	(3)	(4)	(5)
Risk of Low Performing Rating	0.007**	0.265**	0.012**	0.0043**	172
	[0.003]	[0.080]	[0.002]	[0.0011]	[97]
Risk of Recognized Rating	-0.001	-0.238	-0.005	-0.0041	-121
	[0.003]	[0.127]	[0.004]	[0.0037]	[198]
Panel B					
Risk of Low Performing Rating					
Failed an 8th grade exam	0.015**	0.435**	0.014**	0.0060**	194*
	[0.006]	[0.125]	[0.002]	[0.0016]	[89]
Passed 8th grade exams	0.004	0.181*	0.010**	0.0032*	153
	[0.002]	[0.075]	[0.003]	[0.0015]	[99]
Risk of Recognized Rating					
Failed an 8th grade exam	-0.008	-0.395*	-0.028**	-0.0129**	-707**
	[0.009]	[0.173]	[0.006]	[0.0045]	[212]
Passed 8th grade exams	-0.007	-0.215	0.002	-0.0018	49
	[0.003]	[0.121]	[0.005]	[0.0039]	[155]
Sample Size	697,728	697,728	887,713	887,713	887,713

Notes: Each column is a single regression of the indicated outcome on the set of variables from equation (3) in the paper, which includes controls for math and reading scores, demographics, and year and school fixed effects. Standard errors are block boostrapped at the school level. Each coefficient gives the impact of being in a grade cohort that has a positive estimated risk of being rated Low-Performing or Recognized. See the text for details on the construction of the ratings prediction. A one standard deviation change in the math score is equal to about 7 scale score points. College attendance outcomes are measured within an 8 year time window beginning with the student's first-time 9th grade cohort, and measure attendance at any public (and after 2003, any private) institution in the state of Texas. The outcome in Column 5 is annual earnings in the 11th year after the first time a student enters 9th grade (which we refer to as the age 25 year), including students with zero reported earnings. * = sig. at 5% level; ** = sig. at 1% level or less.

Table 4: Impact of Accountability Pressure on High School Outcomes

	Special Education In 10th Grade	Graduated High School	Total Math Credits
Panel A	(1)	(2)	(3)
Risk of Low Performing Rating	-0.001	0.009**	0.060**
	[0.001]	[0.002]	[0.015]
Risk of Recognized Rating	0.002	-0.009*	0.011
	[0.001]	[0.004]	[0.016]
Panel B			
Risk of Low Performing Rating			
Failed an 8th grade exam	-0.003**	0.010**	0.073**
	[0.001]	[0.003]	[0.016]
Passed 8th grade exams	0.000	0.009**	0.051**
	[0.000]	[0.002]	[0.017]
Risk of Recognized Rating			
Failed an 8th grade exam	0.024**	0.013	-0.106**
	[0.004]	[0.007]	[0.023]
Passed 8th grade exams	-0.005**	-0.016**	0.044*
	[0.001]	[0.004]	[0.018]
Sample Size	887,713	887,713	887,713

Notes: Each column is a single regression of the indicated outcome on the set of variables from equation (3) in the paper, which includes controls for math and reading scores, demographics, and year and school fixed effects. Standard errors are block bootstrapped at the school level. Each coefficient gives the impact of being in a grade cohort that has a positive estimated risk of being rated Low-Performing or Recognized. See the text for details on the construction of the ratings prediction. The outcome is Column 1 is the share of students who are classified as eligible to receive special education services in 10th grade, conditional on not having been eligible in 8th grade. High school graduation is defined within an 8 year window beginning in the year a student first enters 9th grade. Math credits are defined as the sum of indicators for passing Algebra I, Geometry, Algebra II and Pre-calculus, for a total maximum value of four. * = sig. at 5% level; ** = sig. at 1% level or less. * = sig. at 5% level; ** = sig. at 1% level or less.

NOT FOR PUBLICATION

Data Appendix for "School Accountability, Postsecondary Attainment and Earnings"

Table A1 - Descriptive Statistics by School Ratings

	Percent Black	Percent Latino	Percent Free Lunch	% Passed 8th Math	% Passed 8th Reading	Avg. Cohort Size	Number of Students
	(1)	(2)	(3)	(4)	(5)	(7)	(8)
Rated Low-Performing at least once	0.182	0.394	0.471	0.612	0.735	333	263,657
Rated Acceptable in every year	0.136	0.414	0.426	0.641	0.768	416	362,780
Rated Recognized at least once	0.048	0.215	0.270	0.751	0.839	274	155,406
Rated Exemplary at least once	0.038	0.119	0.171	0.825	0.892	292	105,870

Notes: This table presents descriptive statistics across schools that are categorized according to the distribution of the accountability ratings that they received over the five year period from 1996 to 2000. The five categories are mutually exclusive and collectively exhaustive.

Table A2: Impact of Accountability Pressure, by Terciles of Predicted Rating

	10th Grade Math		Four Year College		Earnings
Risk of Low-Performing Rating	Passed Test	Scale Score	Ever Attend	BA	Age 25
School Predicted Rating is in:	(1)	(2)	(3)	(4)	(5)
Bottom Third	0.006*	0.228**	0.011**	0.0041**	141
	[0.003]	[0.076]	[0.002]	[0.0011]	[89]
Middle Third	0.014*	0.490**	0.011**	0.0047*	233
	[0.006]	[0.157]	[0.003]	[0.0020]	[130]
Top Third	0.010*	0.308	0.020**	0.0054**	326*
	[0.005]	[0.171]	[0.002]	[0.0019]	[143]
Risk of Recognized Rating					
School Predicted Rating is in:					
Bottom Third	-0.003	-0.085	-0.003	-0.0026	-168
	[0.004]	[0.119]	[0.004]	[0.0034]	[204]
Middle Third	-0.011*	-0.441*	-0.009	-0.0061	-336
	[0.005]	[0.197]	[0.007]	[0.0046]	[267]
Top Third	-0.011*	-0.478**	-0.008	-0.0065	51
	[0.005]	[0.161]	[0.005]	[0.0045]	[226]
Sample Size	697,728	697,728	887,713	887,713	887,713

Notes: Each column is a single regression of the indicated outcome on the variables from equation (3) in the paper, which includes controls for cubics in 8th grade math and reading scores, dummies for male, black, Hispanic, and free/reduced lunch, each student's percentile rank on the 8th grade exams within their incoming 9th grade cohort, year fixed effects, and school fixed effects. Standard errors are block bootstrapped at the school level. Each coefficient gives the impact of being in a cohort that has a positive estimated risk of being rated either Low-Performing or Recognized. The estimates are also allowed to vary by terciles (low/middle/high) of the ratings prediction. The reference category is grade cohorts for whom the estimated risk of receiving an Acceptable rating rounds up to 100 percent. See the text for details on the construction of the ratings prediction. Students who are first time 9th graders in year T and who pass the 10th grade math exam in year T+1 are considered to have passed "on time". A one standard deviation change in the math score is equal to about 7 scale score points. College attendance outcomes are measured within an 8 year time window beginning with the student's first-time 9th grade cohort, and measure attendance at any public (and after 2003, any private) institution in the state of Texas. The outcome in Column 5 is annual earnings in the 11th year after the first time a student enters 9th grade (which we refer to as the age 25 year), including students with zero reported earnings. * = sig. at 5% level; ** = sig. at 1% level or less.

Table A3: Results by lowest scoring subgroup's pass rate relative to the yearly threshold

	10th Grade Math		Four Year College		Earnings
8th Grade Pass Rate of lowest-scoring subgroup and test, relative to yearly threshold, is:	Passed Test	Scale Score	Ever Attend	BA	Age 25
	(1)	(2)	(3)	(4)	(5)
More than 10 points below	0.034**	0.717**	0.015**	0.0053*	1043**
	[0.007]	[0.229]	[0.004]	[0.0026]	[184]
5 to 10 points below	0.037**	0.576**	0.018**	0.0056*	707**
	[0.006]	[0.179]	[0.004]	[0.0025]	[172]
0 to 5 points below	0.022**	0.401**	0.017**	0.0025	841**
	[0.005]	[0.151]	[0.003]	[0.0021]	[154]
0 to 5 points above	0.018**	0.304**	0.009**	0.0011	520**
	[0.005]	[0.125]	[0.003]	[0.0019]	[126]
5 to 10 points above	0.011**	0.250*	0.010**	0.0032	438**
	[0.004]	[0.098]	[0.002]	[0.0018]	[120]
10 to 15 points above	0.007	0.083	0.009**	0.0031*	89
	[0.004]	[0.095]	[0.002]	[0.0015]	[109]
25 to 30 points above	-0.008	-0.030	-0.006*	-0.0039*	-247
	[0.005]	[0.118]	[0.003]	[0.0019]	[169]
30 to 35 points above	-0.006	-0.146	-0.009**	-0.0043	-79
	[0.005]	[0.112]	[0.003]	[0.0024]	[153]
35 to 40 points above	-0.006	-0.333*	-0.013**	-0.0044	-305
	[0.006]	[0.137]	[0.004]	[0.0027]	[240]
More than 40 points above	-0.016*	-0.197	-0.017**	-0.0063*	-317
	[0.006]	[0.150]	[0.004]	[0.0029]	[216]
Sample Size	697,728	697,728	887,713	887,713	887,713

Notes: Each column is a single regression of the indicated outcome on controls for cubics in 8th grade math and reading scores, dummies for male, black, Hispanic, and free/reduced price lunch, each student's percentile rank on the 8th grade exams within their incoming 9th grade cohort, year fixed effects, school fixed effects, and 5 percentage point bins of each school and grade cohort's lowest 8th grade test-subgroup pass rate, minus the yearly passing threshold for an Acceptable rating. 15 to 25 percentage points above the threshold is the left-out category, because nearly all schools in this group would be rated as "safe" using the ratings prediction from our main results. See text for details. Standard errors are block bootstrapped at the school level. Students who are first time 9th graders in year T and who pass the 10th grade math exam in year T+1 are considered to have passed "on time". A one standard deviation change in the math score is equal to about 7 scale score points. College attendance outcomes are measured within an 8 year time window beginning with the student's first-time 9th grade cohort, and measure attendance at any public (and after 2003, any private) institution in the state of Texas. The outcome in Column 5 is annual earnings in the 11th year after the first time a student enters 9th grade (which we refer to as the age 25 year), including students with zero reported earnings. * = sig. at 5% level; ** = sig. at 1% level or less.

Table A4: Transition Matrix for Predicted Ratings Categories

Predicted Rating in Year T	Predicted Rating in Year T+1			
	LP	Safe A	R	Total
Pr(Low-Performing)>0	0.589	0.389	0.021	1,035
Pr(Acceptable) => 100%	0.261	0.634	0.105	1,512
Pr(Recognized)>0	0.043	0.170	0.787	737

	highLP	midLP	lowLP	safeA	lowR	midR	highR	Total
Low-Performing (high)	0.227	0.128	0.370	0.270	0.005	0.000	0.000	211
Low-Performing (mid)	0.157	0.126	0.384	0.327	0.006	0.000	0.000	159
Low-Performing (low)	0.123	0.081	0.323	0.442	0.024	0.002	0.005	665
Pr(Acceptable) => 100%	0.034	0.033	0.194	0.634	0.078	0.013	0.015	1,512
Recognized (low)	0.003	0.008	0.045	0.229	0.416	0.156	0.142	353
Recognized (mid)	0.015	0.000	0.031	0.146	0.292	0.231	0.285	130
Recognized (high)	0.000	0.000	0.024	0.098	0.165	0.094	0.618	254

Notes: The top panel presents a transition matrix of schools across three ratings categories, while the bottom panel gives a similar transition matrix where we break the Low-Performing and Recognized categories into three terciles each. Low is a probabiliy greater than zero and less than or equal to 33 percent, Mid is 33 to 67 percent, and High is 67 to 100 percent. Each cell gives the share of schools in the indicated row category in year T that are included in the indicated column category in year T+1. Rows may not sum exactly to one due to rounding error. See the text for details on the construction of predicted ratings.

Table A5: Main results compared to "placebo" schools with subgroups too small to qualify

	10th Grade Math		Four Year College		Earnings
	Passed Test	Scale Score	Ever Attend	BA	Age 25
Panel A	(1)	(2)	(3)	(4)	(5)
Risk of Low Performing Rating	0.008**	0.318**	0.014**	0.0053**	218*
	[0.003]	[0.090]	[0.002]	[0.0013]	[93]
Placebo	0.002	0.111	0.003	0.0017	74
	[0.003]	[0.077]	[0.002]	[0.0012]	[101]
Risk of Recognized Rating	-0.004	-0.180	-0.007	-0.0048	5
	[0.003]	[0.124]	[0.004]	[0.0031]	[184]
Placebo	-0.013*	-0.394*	-0.006	-0.0032	-49
	[0.006]	[0.180]	[0.006]	[0.0035]	[279]
Panel B					
Risk of Low Performing Rating					
Failed an 8th grade exam	0.023**	0.595**	0.009**	0.0046*	144
	[0.006]	[0.174]	[0.003]	[0.0019]	[119]
Placebo	0.013*	0.319*	-0.008*	-0.0025	-74
	[0.005]	[0.139]	[0.003]	[0.0021]	[101]
Passed 8th grade exams	0.003	0.203*	0.015**	0.0053**	248*
	[0.003]	[0.082]	[0.003]	[0.0019]	[115]
Placebo	-0.002	0.042	0.009**	0.0038*	147
	[0.003]	[0.069]	[0.003]	[0.0016]	[113]
Risk of Recognized Rating					
Failed an 8th grade exam	-0.017*	-0.544**	-0.031**	-0.0144**	-566*
	[0.008]	[0.190]	[0.005]	[0.0037]	[220]
Placebo	0.041*	0.752*	0.021	0.0118**	-97
	[0.019]	[0.380]	[0.011]	[0.0087]	[378]
Passed 8th grade exams	-0.002	-0.098	-0.001	-0.0025	167
	[0.035]	[0.094]	[0.004]	[0.0038]	[170]
Placebo	-0.017**	-0.375	-0.003	-0.0005	-59
	[0.006]	[0.194]	[0.004]	[0.0039]	[264]
Sample Size	697,728	697,728	887,711	887,711	887,711

Notes: Within Panels A and B, each column is a single regression of the indicated outcome on the set of variables from equation (3) in the paper, which includes controls for cubics in 8th grade math and reading scores, dummies for male, black, hispanic, and free/reduced price lunch, each student's percentile rank on the 8th grade exams within their incoming 9th grade cohort, year fixed effects, and school fixed effects. Standard errors are block boostrapped at the school level. Each coefficient gives the impact of being in a grade cohort that has a positive estimated risk of being rated Low-Performing or Recognized, for either all students in the grade cohort (Panel A) or students who failed one / passed both 8th grade exams (Panel B). The reference category is grade cohorts for whom the estimated risk of receiving an Acceptable rating rounds up to 100 percent. See the text for details on the construction of the ratings prediction. Placebos are grade cohorts that would have been "at risk" of being rated Low-Performing or Recognized, except the lowest-scoring subgroup comprised 8 percent or less of the grade cohort, which is below the minimum size threshold of 10 percent. See the text for details. Students who are first time 9th graders in year T and who pass the 10th grade math exam in year T+1 are considered to have passed "on time". A one standard deviation change in the math scale score is equal to about 7 scale score points. College attendance outcomes are measured within an 8 year time window beginning with the student's first-time 9th grade cohort, and measure attendance at any public (and after 2003, any private) institution in the state of Texas. The outcome in Column 5 is annual earnings in the 11th year after the first time a student enters 9th grade (which we refer to as the age 25 year), including students with zero reported earnings. * = sig. at 5% level; ** = sig. at 1% level or less.

Table A6: Determinants of Schools' Predicted Ratings

Outcome is prob(Low-Performing)>0

	(1)	(2)	(3)	(4)	(5)	(6)
Percent Black	0.201*	0.390	0.155	0.080	0.647*	0.783
	[0.073]	[0.285]	[0.091]	[0.476]	[0.288]	[0.402]
Percent Latino	-0.197**	-0.052	-0.123	-0.189	0.165	-0.012
	[0.062]	[0.213]	[0.078]	[0.355]	[0.216]	[0.339]
Percent Free Lunch	0.096	-0.213	0.013	0.024	-0.142	-0.253
	[0.084]	[0.120]	[0.102]	[0.202]	[0.119]	[0.181]
8th Gd. Math Pass Rate	-0.413**	-0.456**	-0.360**	-0.412**	-0.562**	-0.367**
	[0.058]	[0.070]	[0.063]	[0.095]	[0.104]	[0.107]
First-time 9th grade in 1996	0.029	0.029			0.031	
	[0.018]	[0.020]			[0.020]	
First-time 9th grade in 1997	0.065**	0.073**			0.111**	0.035
	[0.020]	[0.023]			[0.023]	[0.028]
First-time 9th grade in 1998	0.099**	0.113**	0.057**	0.050*	0.184**	0.057
	[0.022]	[0.024]	[0.021]	[0.026]	[0.026]	[0.039]
First-time 9th grade in 1999	0.066**	0.082**	0.052*	0.034	0.180**	
	[0.024]	[0.028]	[0.025]	[0.032]	[0.030]	
Teacher Yrs of Experience			0.015	-0.005		
			[0.022]	[0.027]		
Changed Principals			0.012*	-0.003		
			[0.006]	[0.014]		
Average Teacher Pay (in $1000s)			-0.024**	-0.013		
			[0.006]	[0.007]		
8th Grade Math Pass Rate - Black					-0.142	
					[0.098]	
8th Grade Math Pass Rate - Latino					0.353**	
					[0.113]	
8th Grade Math Pass Rate - Ec. Disadv.					-0.534**	
					[0.120]	
Lag of 8th Grade Math Pass Rate						0.110
						[0.104]
Lead of 8th Grade Math Pass Rate						0.014
						[0.092]
School Fixed Effects	No	Yes	No	Yes	Yes	Yes
F(demographics = 0)	0.000	0.135	0.001	0.938	0.105	0.091
F (school vars = 0)			0.001	0.300		
F (cohort effects = 0)	0.000	0.000	0.020	0.142	0.000	0.328
F(lag and lead = 0)						0.561
R-Squared	0.055	0.480	0.055	0.590	0.495	0.586
Sample Size	4,506	4,506	2,618	2,618	4,506	2,693

Notes: Each column represents a single regression of the probability that a grade cohort will be rated "Low-Performing" on the indicated set of time-varying school characteristics. The teacher and principal variables are measured as of each cohort's 9th grade year, and are only available from 1997 onward. The subgroup math pass rates in Column 5 are given a value of zero in schools with too few students to count, and we also include a dummy variable that is equal to one if the group is missing. The lag and lead variables in Column 6 are the average math pass rates of the grade cohorts immediately before and after the one in question, and thus are only available for grade cohorts 1996, 1997, and 1998. See the text in Section V for a description of how schools' predicted ratings were constructed. * = sig. at 5% level; ** = sig. at 1% level or less.

Table A7 - Impact of Pre-Accountability Test Score Trends on Predicted Rating

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
8th grade scores - all	-0.290**	-0.277**	-0.368**	-0.470**	-0.474**	-0.473**	-0.485**
	[0.057]	[0.058]	[0.053]	[0.070]	[0.071]	[0.072]	[0.072]
8th grade scores - black	-0.029	-0.031	-0.053**	-0.038	-0.038	-0.037	-0.035
	[0.017]	[0.017]	[0.019]	[0.021]	[0.021]	[0.021]	[0.021]
8th grade scores - Latino	-0.009	-0.008	-0.290**	-0.074**	-0.074**	-0.074**	-0.076**
	[0.024]	[0.024]	[0.075]	[0.024]	[0.024]	[0.024]	[0.024]
8th grade scores - FRPL	-0.999**	-0.995**	-0.859**	-0.597**	-0.598**	-0.599**	-0.598**
	[0.074]	[0.074]	[0.063]	[0.065]	[0.065]	[0.065]	[0.065]
Linear Trend		0.105**	0.094**				
		[0.013]	[0.016]				
1994 Pass Rate - all		-0.019	-0.364				
		[0.114]	[0.226]				
Trend*1994 pass rate		-0.025	0.008		-0.011	-0.009	0.011
		[0.023]	[0.036]		[0.025]	[0.038]	[0.040]
Trend*1993 pass rate			0.017				0.007
			[0.029]				[0.032]
Trend*1992 pass rate			0.017				0.021
			[0.032]				[0.034]
Trend*1991 pass rate			-0.029				-0.035
			[0.023]				[0.025]
Trend * 1994 subgroup pass rates	no	no	yes	no	no	yes	yes
Trend * 1991-1993 subgroup pass rates	no	no	yes	no	no	no	yes
School Fixed Effects	no	no	no	yes	yes	yes	yes
Number of trend interactions	0	1	4	16	1	4	16
F (Trends = 0)		0.000	0.000		0.656	0.960	0.482
R-squared	0.277	0.278	0.350	0.618	0.618	0.618	0.621
Sample size	4,253	4,253	4,253	4,253	4,253	4,253	4,253

Notes: Each column represents a single regression of the probability that a grade cohort will be rated "Low-Performing" on the indicated set of time-varying school characteristics. The models in Columns 1 through 3 include a linear trend indexed by cohort, mathematics pass rates overall and by subgroup (black, Latino, free lunch) for grade cohorts 1991 through 1994, and the interaction between them. Columns 4 through 7 only include the pass rate by trend interactions, since only these are identified after controlling for school fixed effects. Subgroup pass rates are given a value of zero in schools with too few students to count, and we also include a dummy variable that is equal to one if the group is missing. See the text for details on the construction of the ratings prediction. * = sig. at 5% level; ** = sig. at 1% level or less.

Table A8: Main Results with controls for pre-accountability test score trend interactions

	10th Grade Math		Four Year College		Earnings
	Passed Test	Scale Score	Ever Attend	BA	Age 25
Panel A	(1)	(2)	(3)	(4)	(5)
Risk of Low Performing Rating	0.007**	0.254**	0.011**	0.0044**	167*
	[0.003]	[0.085]	[0.002]	[0.0012]	[81]
Risk of Recognized Rating	-0.005	-0.212	-0.004	-0.0040	-96
	[0.003]	[0.122]	[0.004]	[0.0032]	[176]
Panel B					
Risk of Low Performing Rating					
Failed an 8th grade exam	0.016**	0.428**	0.014**	0.0061**	186
	[0.005]	[0.148]	[0.003]	[0.0015]	[93]
Passed 8th grade exams	0.003	0.169*	0.009**	0.0032*	133
	[0.003]	[0.080]	[0.003]	[0.0016]	[104]
Risk of Recognized Rating					
Failed an 8th grade exam	-0.009	-0.408*	-0.028**	-0.0131**	-642**
	[0.008]	[0.199]	[0.006]	[0.0042]	[218]
Passed 8th grade exams	-0.006	-0.182	0.002	-0.0015	61
	[0.004]	[0.124]	[0.005]	[0.0034]	[183]
Sample Size	697,728	697,728	887,713	887,713	887,713

Notes: Within Panels A and B, each column is a single regression of the indicated outcome on the set of variables from equations (1) (Panel A) or (2) (Panel B) in the paper, which includes controls for cubics in 8th grade math and reading scores, dummies for male, black, Hispanic, and free/reduced price lunch, each student's percentile rank on the 8th grade exams within their incoming 9th grade cohort, year fixed effects, school fixed effects, and interactions between a linear trend and overall and subgroup-specific math and reading pass rates for the high school for the four years (1991-1994) prior to the cohorts used in our sample. Standard errors are block bootstrapped at the school level. Each coefficient gives the impact of being in a grade cohort that has a positive estimated risk of being rated Low-Performing or Recognized, for either all students in the grade cohort (Panel A) or students who failed one / passed both 8th grade exams (Panel B). The reference category is grade cohorts for whom the estimated risk of receiving an Acceptable rating rounds up to 100 percent. See the text for details on the construction of the ratings prediction. Students who are first time 9th graders in year T and who pass the 10th grade math exam in year T+1 are considered to have passed "on time". A one standard deviation change in the math score is equal to about 7 scale score points. College attendance outcomes are measured within an 8 year time window beginning with the student's first-time 9th grade cohort, and measure attendance at any public (and after 2003, any private) institution in the state of Texas. The outcome in Column 5 is annual earnings in the 11th year after the first time a student enters 9th grade (which we refer to as the age 25 year), including students with zero reported earnings. * = sig. at 5% level; ** = sig. at 1% level or less.

Table A9: Sensitivity of Earnings Results to Imputation

	Annual Earnings at Age 25			
	Missing = Zero	Impute Mean	Minus 1 SD	Plus 1 SD
	(1)	(2)	(3)	(4)
Risk of Low Performing Rating	172	240**	332**	149*
	[97]	[66]	[84]	[69]
Risk of Recognized Rating	-121	1032**	102	1,962**
	[198]	[132]	[153]	[150]
Sample Size	887,713	887,713	887,713	887,713

Notes: Each column is a single regression of the indicated outcome on the set of variables from equations (1) in the paper, which includes controls for cubics in 8th grade math and reading scores, dummies for male, black, Hispanic, and free/reduced price lunch, each student's percentile rank on the 8th grade exams within their incoming 9th grade cohort, year fixed effects, and school fixed effects. Standard errors are block bootstrapped at the school level. Each coefficient gives the impact of being in a grade cohort that has a positive estimated risk of being rated Low-Performing or Recognized. The reference category is grade cohorts for whom the estimated risk of receiving an Acceptable rating rounds up to 100 percent. See the text for details on the construction of the ratings prediction. The outcomes in Columns 1 through 4 are annual earnings in the 11th years after the first time a student enters 9th grade (which we refer to as the age 25 year). Column 1 replicates the main results from Table 3. Column 2 replaces missing earnings with the mean value of earnings for students in the grade cohort and school ratings category. Columns 3 and 4 subtract and add 1 standard deviation from that mean value, respectively. * = sig. at 5% level; ** = sig. at 1% level or less.

Table A10: Main Results by high school share of out-of-state college attendees

	10th Grade Math		Four Year College		Earnings
	Passed Test	Scale Score	Ever Attend	BA	Age 25
	(1)	(2)	(3)	(4)	(5)
Risk of Low Performing Rating	0.007**	0.260**	0.012**	0.0045**	188*
	[0.003]	[0.090]	[0.002]	[0.0012]	[87]
*>10% attend out-of-state	0.015	0.500	-0.010	-0.0044	-383
	[0.009]	[0.267]	[0.008]	[0.0047]	[370]
Risk of Recognized Rating	-0.005	-0.226	-0.006	-0.0050	-151
	[0.004]	[0.129]	[0.004]	[0.0033]	[190]
*>10% attend out-of-state	-0.007	0.034	0.003	0.0052	178
	[0.008]	[0.260]	[0.014]	[0.123]	[613]
Sample Size	697,728	697,728	887,713	887,713	887,713

Notes: Each column is a single regression of the indicated outcome on the set of variables from equations (1) in the paper, which includes controls for cubics in 8th grade math and reading scores, dummies for male, black, Hispanic, and free/reduced price lunch, each student's percentile rank on the 8th grade exams within their incoming 9th grade cohort, year fixed effects, and school fixed effects. The main treatment variables are interacted with indicators that are equal to one if a high school sends more than 10 percent of college-bound seniors to out-of-state institutions (based on a match of 2008/2009 graduating classes to the National Student Clearinghouse - see text for details.) Standard errors are block bootstrapped at the school level. Each coefficient gives the impact of being in a grade cohort that has a positive estimated risk of being rated Low-Performing or Recognized. The reference category is grade cohorts for whom the estimated risk of receiving an Acceptable rating rounds up to 100 percent. See the text for details on the construction of the ratings prediction. Students who are first time 9th graders in year T and who pass the 10th grade math exam in year T+1 are considered to have passed "on time". College attendance outcomes are measured within an 8 year time window beginning with the student's first-time 9th grade cohort, and measure attendance at any public (and after 2003, any private) institution in the state of Texas. The outcome in Column 5 is annual earnings in the 11th year after the first time a student enters 9th grade (which we refer to as the age 25 year), including students with zero reported earnings. * = sig. at 5% level; ** = sig. at 1% level or less.

Table A11: Main Results restricted to non-consecutive cohorts (1995, 1997 and 1999)

	10th Grade Math		Four Year College		Earnings
	Passed Test	Scale Score	Ever Attend	BA	Age 25
Panel A	(1)	(2)	(3)	(4)	(5)
Risk of Low Performing Rating	0.015**	0.450**	0.014**	0.0059**	202
	[0.004]	[0.107]	[0.003]	[0.0018]	[132]
Risk of Recognized Rating	-0.005	-0.292*	-0.007	-0.0010	-17
	[0.005]	[0.147]	[0.005]	[0.0036]	[251]
Panel B					
Risk of Low Performing Rating					
Failed an 8th grade exam	0.033**	0.851**	0.013**	0.0071**	280
	[0.007]	[0.179]	[0.004]	[0.0021]	[156]
Passed 8th grade exams	0.007*	0.271**	0.014**	0.0052*	148
	[0.00]	[0.104]	[0.003]	[0.0023]	[152]
Risk of Recognized Rating					
Failed an 8th grade exam	-0.005	-0.434	-0.029**	-0.0099*	-610
	[0.009]	[0.224]	[0.007]	[0.0043]	[345]
Passed 8th grade exams	-0.006	-0.298*	-0.001	0.0011	132
	[0.005]	[0.151]	[0.005]	]0.0039]	[262]
Sample Size	415,731	415,731	528,830	528,830	528,830

Notes: The 1996 and 1998 grade cohorts are excluded from this sample. Within Panels A and B, each column is a single regression of the indicated outcome on the set of variables from equations (1) (Panel A) or (2) (Panel B) in the paper, which includes controls for cubics in 8th grade math and reading scores, dummies for male, black, Hispanic, and free/reduced price lunch, each student's percentile rank on the 8th grade exams within their incoming 9th grade cohort, year fixed effects, and school fixed effects. Standard errors are block bootstrapped at the school level. Each coefficient gives the impact of being in a grade cohort that has a positive estimated risk of being rated Low-Performing or Recognized, for either all students in the grade cohort (Panel A) or students who failed one / passed both 8th grade exams (Panel B). The reference category is grade cohorts for whom the estimated risk of receiving an Acceptable rating rounds up to 100 percent. See the text for details on the construction of the ratings prediction. Students who are first time 9th graders in year T and who pass the 10th grade math exam in year T+1 are considered to have passed "on time". A one standard deviation change in the math score is equal to about 7 scale score points. College attendance outcomes are measured within an 8 year time window beginning with the student's first-time 9th grade cohort, and measure attendance at any public (and after 2003, any private) institution in the state of Texas. The outcome in Column 5 is annual earnings in the 11th year after the first time a student enters 9th grade (which we refer to as the age 25 year), including students with zero reported earnings. * = sig. at 5% level; ** = sig. at 1% level or less.

Table A12: Main Results restricted to non-overlapping cohorts (1995 and 1999)

	10th Grade Math		Four Year College		Earnings
	Passed Test	Scale Score	Ever Attend	BA	Age 25
Panel A	(1)	(2)	(3)	(4)	(5)
Risk of Low Performing Rating	0.011	0.383*	0.018**	0.0077**	391*
	[0.006]	[0.177]	[0.004]	[0.0026]	[177]
Risk of Recognized Rating	-0.012	-0.548*	0.006	0.0026	54
	[0.008]	[0.247]	[0.007]	[0.0055]	[406]
Panel B					
Risk of Low Performing Rating					
Failed an 8th grade exam	0.033**	0.874**	0.018**	0.0099**	340
	[0.010]	[0.267]	[0.005]	[0.0029]	[201]
Passed 8th grade exams	0.002	0.170	0.019**	0.0064*	424*
	[0.005]	[0.166]	[0.005]	[0.0031]	[191]
Risk of Recognized Rating					
Failed an 8th grade exam	-0.003	-0.539	-0.021*	-0.0066	-623
	[0.011]	[0.319]	[0.009]	[0.0062]	[467]
Passed 8th grade exams	-0.015*	-0.599*	0.012	0.0044	269
	[0.008]	[0.246]	[0.007]	[0.0058]	[394]
Sample Size	273,177	273,177	348,375	348,375	348,375

Notes: The 1996, 1997 and 1998 grade cohorts are excluded from this sample. Within Panels A and B, each column is a single regression of the indicated outcome on the set of variables from equations (1) (Panel A) or (2) (Panel B) in the paper, which includes controls for cubics in 8th grade math and reading scores, dummies for male, black, Hispanic, and free/reduced price lunch, each student's percentile rank on the 8th grade exams within their incoming 9th grade cohort, year fixed effects, and school fixed effects. Standard errors are block bootstrapped at the school level. Each coefficient gives the impact of being in a grade cohort that has a positive estimated risk of being rated Low-Performing or Recognized, for either all students in the grade cohort (Panel A) or students who failed one / passed both 8th grade exams (Panel B). The reference category is grade cohorts for whom the estimated risk of receiving an Acceptable rating rounds up to 100 percent. See the text for details on the construction of the ratings prediction. Students who are first time 9th graders in year T and who pass the 10th grade math exam in year T+1 are considered to have passed "on time". A one standard deviation change in the math score is equal to about 7 scale score points. College attendance outcomes are measured within an 8 year time window beginning with the student's first-time 9th grade cohort, and measure attendance at any public (and after 2003, any private) institution in the state of Texas. The outcome in Column 5 is annual earnings in the 11th year after the first time a student enters 9th grade (which we refer to as the age 25 year), including students with zero reported earnings. * = sig. at 5% level; ** = sig. at 1% level or less.

Table A13: Main results when controlling for new special education classification

	Four Year College				Earnings	
	Ever Attend		BA		Age 25	
	(1)	(2)	(3)	(4)	(5)	(6)
Special Education in 10th grade		-0.049**		-0.034**		-1,740**
		[0.005]		[0.002]		[187]
Risk of Low Performing Rating	0.011**	0.012**	0.0043**	0.0044**	172	178*
	[0.002]	[0.003]	[0.0011]	[0.0012]	[97]	[77]
Risk of Recognized Rating	-0.006	-0.003	-0.0041	-0.0043	-121	7
	[0.004]	[0.004]	[0.0037]	[0.0031]	[98]	[169]
Sample Size	887,711	887,711	887,711	887,711	887,711	887,711

Notes: The 1996 and 1998 grade cohorts are excluded from this sample. Within Panels A and B, each column is a single regression of the indicated outcome on the set of variables from equations (1) (Panel A) or (2) (Panel B) in the paper, which includes controls for cubics in 8th grade math and reading scores, dummies for male, black, Hispanic, and free/reduced price lunch, each student's percentile rank on the 8th grade exams within their incoming 9th grade cohort, year fixed effects, and school fixed effects. Standard errors are block bootstrapped at the school level. Each coefficient gives the impact of being in a grade cohort that has a positive estimated risk of being rated Low-Performing or Recognized, for either all students in the grade cohort (Panel A) or students who failed one / passed both 8th grade exams (Panel B). The reference category is grade cohorts for whom the estimated risk of receiving an Acceptable rating rounds up to 100 percent. See the text for details on the construction of the ratings prediction. Students who are first time 9th graders in year T and who pass the 10th grade math exam in year T+1 are considered to have passed "on time". College attendance outcomes are measured within an 8 year time window beginning with the student's first-time 9th grade cohort, and measure attendance at any public (and after 2003, any private) institution in the state of Texas. The outcome in Column 5 is annual earnings in the 11th year after the first time a student enters 9th grade (which we refer to as the age 25 year), including students with zero reported earnings. * = sig. at 5% level; ** = sig. at 1% level or less.

Table A14: Main results when controlling for total math credits

	Four Year College				Earnings	
	Ever Attend		BA		Age 25	
	(1)	(2)	(3)	(4)	(5)	(6)
Math Credits		0.109**		0.055**		2,671**
		[0.002]		[0.001]		[37]
Risk of Low Performing Rating	0.011**	0.006**	0.0043**	0.0012	172	22
	[0.002]	[0.002]	[0.0011]	[0.0014]	[97]	[79]
Risk of Recognized Rating	-0.006	-0.007	-0.0041	-0.0048	-121	-13
	[0.004]	[0.004]	[0.0037]	[0.0031]	[98]	[181]
Sample Size	887,711	887,711	887,711	887,711	887,711	887,711

Notes: The 1996 and 1998 grade cohorts are excluded from this sample. Within Panels A and B, each column is a single regression of the indicated outcome on the set of variables from equations (1) (Panel A) or (2) (Panel B) in the paper, which includes controls for cubics in 8th grade math and reading scores, dummies for male, black, Hispanic, and free/reduced price lunch, each student's percentile rank on the 8th grade exams within their incoming 9th grade cohort, year fixed effects, and school fixed effects. Standard errors are block bootstrapped at the school level. Each coefficient gives the impact of being in a grade cohort that has a positive estimated risk of being rated Low-Performing or Recognized, for either all students in the grade cohort (Panel A) or students who failed one / passed both 8th grade exams (Panel B). The reference category is grade cohorts for whom the estimated risk of receiving an Acceptable rating rounds up to 100 percent. See the text for details on the construction of the ratings prediction. Students who are first time 9th graders in year T and who pass the 10th grade math exam in year T+1 are considered to have passed "on time". College attendance outcomes are measured within an 8 year time window beginning with the student's first-time 9th grade cohort, and measure attendance at any public (and after 2003, any private) institution in the state of Texas. The outcome in Column 5 is annual earnings in the 11th year after the first time a student enters 9th grade (which we refer to as the age 25 year), including students with zero reported earnings. * = sig. at 5% level; ** = sig. at 1% level or less.

Table A15A: Impact of Accountability Pressure on Additional Outcomes

Panel A	Reading Scale Score (1)	Took 10th Math On Time (2)	Passed 10th Writing On Time (3)	10th Gd. Absences (4)	Same School T+1 (5)	Same Schl, On Time (6)	Still in TX in T+1 (7)	Transfer to Alt. School (8)
Risk of Low Performing Rating	0.288**	0.005	0.005	-0.123	0.005	0.013**	0.003*	-0.005*
	[0.062]	[0.003]	[0.003]	[0.089]	[0.005]	[0.005]	[0.001]	[0.002]
Risk of Recognized Rating	-0.089	-0.009	-0.008	-0.189	-0.001	-0.005	0.001	-0.003
	[0.081]	[0.006]	[0.006]	[0.139]	[0.006]	[0.006]	[0.002]	[0.004]
Panel B								
Risk of Low Performing Rating								
Failed an 8th grade exam	0.321**	0.009	0.008	-0.063	0.010*	0.020**	0.005**	-0.011**
	[0.104]	[0.005]	[0.004]	[0.130]	[0.004]	[0.005]	[0.001]	[0.003]
Passed 8th grade exams	0.270**	0.002	0.003	-0.157	0.002	0.008	0.003*	-0.001
	[0.060]	[0.004]	[0.004]	[0.091]	[0.005]	[0.006]	[0.001]	[0.002]
Risk of Recognized Rating								
Failed an 8th grade exam	-0.170	0.015	0.006	-0.924**	-0.007	0.019*	-0.000	-0.022**
	[0.141]	[0.008]	[0.008]	[0.195]	[0.006]	[0.008]	[0.003]	[0.006]
Passed 8th grade exams	-0.072	-0.016**	-0.012*	-0.049	0.001	-0.013*	0.001	0.004
	[0.087]	[0.006]	[0.006]	[0.142]	[0.006]	[0.006]	[0.002]	[0.005]
Sample Size	697,404	887,713	887,713	543,744	887,713	887,713	887,713	887,713

Notes: Within Panels A and B, each column is a single regression of the indicated outcome on the set of variables from equations (1) (Panel A) or (2) (Panel B) in the paper, which includes controls for cubics in 8th grade math and reading scores, dummies for male, black, Hispanic, and free/reduced price lunch, each student's percentile rank on the 8th grade exams within their incoming 9th grade cohort, year fixed effects, and school fixed effects. Standard errors are block bootstrapped at the school level. Each coefficient gives the impact of being in a grade cohort that has a positive estimated risk of being rated Low-Performing or Recognized, for either all students in the grade cohort (Panel A) or students who failed one / passed both 8th grade exams (Panel B). The reference category is grade cohorts for whom the estimated risk of receiving an Acceptable rating rounds up to 100 percent. See the text for details on the construction of the ratings prediction. Students who are first time 9th graders in year T and who are in 10th grade and/or pass the 10th grade math exam in year T+1 are considered to be or to have passed "on time". Data on absences (Column 3) are available only beginning in 1998. Alternative schools are generally (although not always) intended for students who have behavior problems. * = sig. at 5% level; ** = sig. at 1% level or less.

Table A15B: Impact of Accountability Pressure on Additional Outcomes

Panel A	Pass Algebra I (8)	Pass Geometry (9)	Pass Algebra II (10)	Pass Pre-Calc (11)	Attend 2 yr coll (12)	AA (13)	Attend Flagship (14)
Risk of Low Performing Rating	0.002	0.021**	0.021**	0.016**	0.008**	0.0011*	0.0030**
	[0.008]	[0.006]	[0.004]	[0.003]	[0.002]	[0.0005]	[0.0008]
Risk of Recognized Rating	0.028*	-0.001	-0.004	-0.012	-0.001	0.0002	-0.0032
	[0.012]	[0.010]	[0.005]	[0.006]	[0.006]	[0.0015]	[0.0023]
Panel B							
Risk of Low Performing Rating							
Failed an 8th grade exam	0.019*	0.028**	0.017**	0.008*	0.003	0.0018**	-0.0016
	[0.009]	[0.006]	[0.005]	[0.004]	[0.004]	[0.0006]	[0.0016]
Passed 8th grade exams	-0.010	0.016*	0.023**	0.021**	0.011**	0.0007	0.0061**
	[0.009]	[0.007]	[0.006]	[0.004]	[0.003]	[0.0007]	[0.0015]
Risk of Recognized Rating							
Failed an 8th grade exam	0.029	-0.047**	-0.058**	-0.030**	0.021	0.0031	-0.0203**
	[0.015]	[0.012]	[0.008]	[0.009]	[0.011]	[0.0017]	[0.0043]
Passed 8th grade exams	0.026*	0.011	0.012*	-0.005	-0.007	-0.0007	0.0023
	[0.012]	[0.011]	[0.006]	[0.007]	[0.007]	[0.0016]	[0.0030]
Sample Size	887,713	887,713	887,713	887,713	887,713	887,713	887,713

Notes: Within Panels A and B, each column is a single regression of the indicated outcome on the set of variables from equations (1) (Panel A) or (2) (Panel B) in the paper, which includes controls for cubics in 8th grade math and reading scores, dummies for male, black, Hispanic, and free/reduced price lunch, each student's percentile rank on the 8th grade exams within their incoming 9th grade cohort, year fixed effects, and school fixed effects. Standard errors are block bootstrapped at the school level. Each coefficient gives the impact of being in a grade cohort that has a positive estimated risk of being rated Low-Performing or Recognized, for either all students in the grade cohort (Panel A) or students who failed one / passed both 8th grade exams (Panel B). The reference category is grade cohorts for whom the estimated risk of receiving an Acceptable rating rounds up to 100 percent. See the text for details on the construction of the ratings prediction. The math courses in rows 8 through 11 are state-standardized courses - students are considered to have passed if they received at least one course credit at any point in their high school career. College attendance outcomes are measured within an 8 year time window beginning with the student's first-time 9th grade cohort, and measure attendance at any public (and after 2003, any private) institution in the state of Texas. Flagship institutions are UT-Austin and Texas A&M. * = sig. at 5% level; ** = sig. at 1% level or less.

Table A16: Impact of Differential Accountability Pressure for Targeted Subgroups

	10th Grade Math		Four Year College		Earnings
	Passed Test	Scale Score	Ever Attend	BA	Age 25
Risk of Low-Performing Rating	(1)	(2)	(3)	(4)	(5)
Targeted Subgroup, Failed 8th Grade Exam	0.011*	0.279*	0.012*	0.008**	579**
	[0.005]	[0.134]	[0.005]	[0.002]	[141]
Risk of Recognized Rating					
Targeted Subgroup, Failed 8th Grade Exam	0.009	-0.370	-0.012	-0.006	-193
	[0.020]	[0.422]	[0.012]	[0.008]	[586]
Sample Size	618,721	618,721	797,703	797,703	797,703

Notes: Each column is a single regression of the indicated outcome on the set of variables from equation (4) in the paper, which includes controls for cubics in 8th grade math and reading scores, dummies for an exhaustive set of race (black/Latino vs. white/other) by poverty by prior test score (failed either or passed both 8th grade exams) categories, each student's percentile rank on the 8th grade exams within their incoming 9th grade cohort, and school-by-year fixed effects. Standard errors are block bootstrapped at the school level. Each coefficient gives the difference in outcomes between students in a targeted subgroup (i.e. poor black or Latino students with low 8th grade test scores) and all other students, within a grade cohort and school that has a positive estimated risk of being rated either Low-Performing or Recognized. The reference category is the difference between targeted subgroups and all other students in grade cohorts for whom the estimated risk of receiving an Acceptable rating rounds up to 100 percent. See the text for details on the construction of the ratings prediction. Students who are first time 9th graders in year T and who pass the 10th grade math exam in year T+1 are considered to have passed "on time". A one standard deviation change in the math score is equal to about 7 scale score points. College attendance outcomes are measured within an 8 year time window beginning with the student's first-time 9th grade cohort, and measure attendance at any public (and after 2003, any private) institution in the state of Texas. The outcome in Column 5 is annual earnings in the 11th year after the first time a student enters 9th grade (which we refer to as the age 25 year), including students with zero reported earnings. * = sig. at 5% level; ** = sig. at 1% level or less.

Table A17: Main Results by gender

	10th Grade Math		Four Year College		Earnings
	Passed Test	Scale Score	Ever Attend	BA	Age 25
	(1)	(2)	(3)	(4)	(5)
Risk of Low Performing Rating	0.007*	0.247**	0.008**	0.0013	79
	[0.003]	[0.088]	[0.002]	[0.0015]	[120]
*Male	0.001	0.038	0.008**	0.0062**	189
	[0.002]	[0.046]	[0.002]	[0.0017]	[196]
Risk of Recognized Rating	-0.001	-0.132	-0.008	-0.0142**	-286
	[0.004]	[0.118]	[0.004]	[0.0038]	[226]
*Male	-0.011**	-0.215**	0.006	0.0203**	333
	[0.002]	[0.055]	[0.004]	[0.0038]	[312]
Sample Size	697,728	697,728	887,713	887,713	887,713

Notes: Each column is a single regression of the indicated outcome on the set of variables from equations (1) in the paper, which includes controls for cubics in 8th grade math and reading scores, dummies for male, black, Hispanic, and free/reduced price lunch, each student's percentile rank on the 8th grade exams within their incoming 9th grade cohort, year fixed effects, and school fixed effects. The main treatment variables are interacted with indicators that are equal to one if a student is male. Standard errors are block bootstrapped at the school level. Each coefficient gives the impact of being in a grade cohort that has a positive estimated risk of being rated Low-Performing or Recognized. The reference category is grade cohorts for whom the estimated risk of receiving an Acceptable rating rounds up to 100 percent. See the text for details on the construction of the ratings prediction. Students who are first time 9th graders in year T and who pass the 10th grade math exam in year T+1 are considered to have passed "on time". A one standard deviation change in the math score is equal to about 7 scale score points. College attendance outcomes are measured within an 8 year time window beginning with the student's first-time 9th grade cohort, and measure attendance at any public (and after 2003, any private) institution in the state of Texas. The outcome in Column 5 is annual earnings in the 11th year after the first time a student enters 9th grade (which we refer to as the age 25 year), including students with zero reported earnings. * = sig. at 5% level; ** = sig. at 1% level or less.

Table A18: Main Results by limited English proficiency

	10th Grade Math		Four Year College		Earnings
	Passed Test	Scale Score	Ever Attend	BA	Age 25
	(1)	(2)	(3)	(4)	(5)
Risk of Low Performing Rating	0.007**	0.251**	0.011**	0.0040**	188*
	[0.003]	[0.083]	[0.002]	[0.0012]	[84]
*LEP	0.011	0.363	0.012	0.0055	-351
	[0.008]	[0.226]	[0.007]	[0.0037]	[233]
Risk of Recognized Rating	-0.007*	-0.245*	-0.005	-0.0042	-98
	[0.003]	[0.116]	[0.004]	[0.0032]	[179]
*LEP	-0.006	-0.096	-0.007	0.0024	-170
	[0.018]	[0.440]	[0.016]	[0.0117]	[121]
Sample Size	697,728	697,728	887,713	887,713	887,713

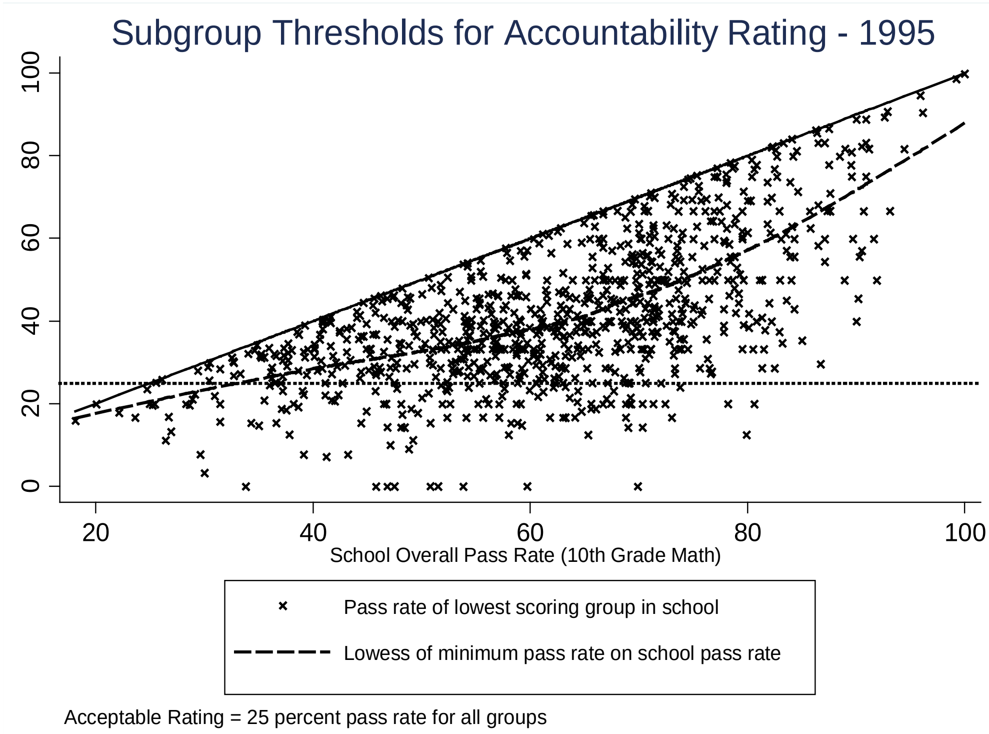
Notes: Each column is a single regression of the indicated outcome on the set of variables from equations (1) in the paper, which includes controls for cubics in 8th grade math and reading scores, dummies for male, black, Hispanic, and free/reduced price lunch, each student's percentile rank on the 8th grade exams within their incoming 9th grade cohort, year fixed effects, and school fixed effects. The main treatment variables are interacted with indicators that are equal to one if a student was designated as having limited English proficiency in 8th grade. Standard errors are block bootstrapped at the school level. Each coefficient gives the impact of being in a grade cohort that has a positive estimated risk of being rated Low-Performing or Recognized. The reference category is grade cohorts for whom the estimated risk of receiving an Acceptable rating rounds up to 100 percent. See the text for details on the construction of the ratings prediction. Students who are first time 9th graders in year T and who pass the 10th grade math exam in year T+1 are considered to have passed "on time". A one standard deviation change in the math score is equal to about 7 scale score points. College attendance outcomes are measured within an 8 year time window beginning with the student's first-time 9th grade cohort, and measure attendance at any public (and after 2003, any private) institution in the state of Texas. The outcome in Column 5 is annual earnings in the 11th year after the first time a student enters 9th grade (which we refer to as the age 25 year), including students with zero reported earnings. * = sig. at 5% level; ** = sig. at 1% level or less.

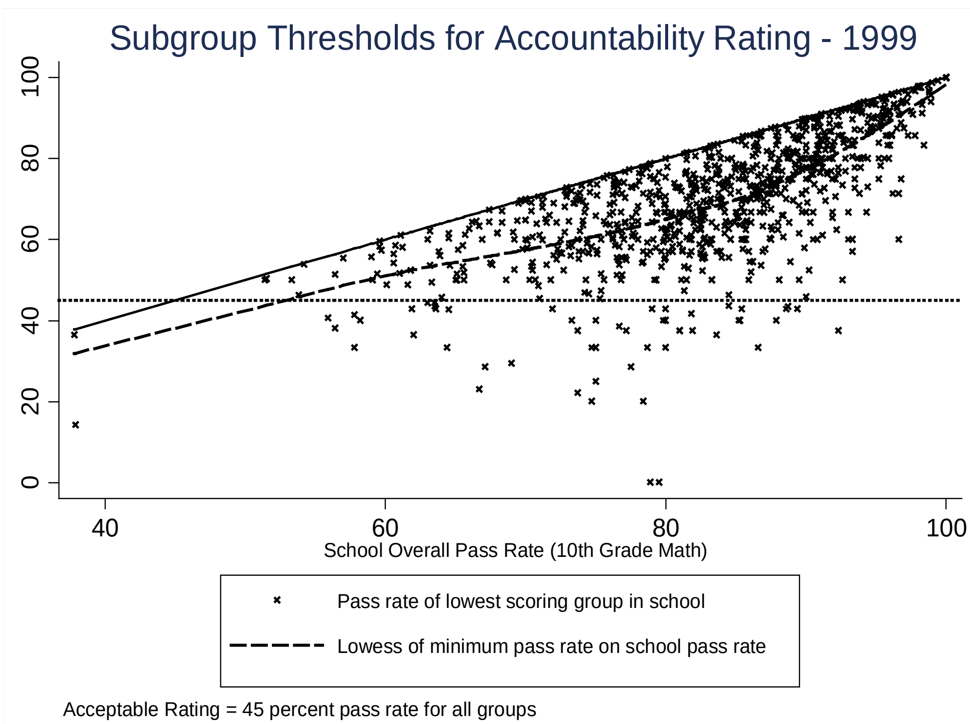
Table A19: Impacts on college enrollment, earnings and idle by year

	Enrolled in any postsecondary institution						
	Age 19	Age 20	Age 21	Age 22	Age 23	Age 24	Age 25
Panel A	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Risk of Low Performing Rating	0.010**	0.009**	0.009**	0.010**	0.004**	0.002*	0.002*
	[0.002]	[0.002]	[0.002]	[0.002]	[0.001]	[0.001]	[0.001]
Risk of Recognized Rating	-0.002	0.001	-0.004	-0.003	-0.005	0.001	-0.002
	[0.004]	[0.003]	[0.003]	[0.004]	[0003]	[0.002]	[0.002]
	Annual earnings if not enrolled in college						
Panel B	Age 19	Age 20	Age 21	Age 22	Age 23	Age 24	Age 25
Risk of Low Performing Rating	51	135*	131*	232**	279**	200*	269**
	[49]	[56]	[65]	[71]	[75]	[82]	[86]
Risk of Recognized Rating	69	10	-115	131	278	283	260
	[102]	[119]	[140]	[167]	[161]	[185]	[200]
	Idle (zero earnings, not enrolled in college)						
Panel C	Age 19	Age 20	Age 21	Age 22	Age 23	Age 24	Age 25
Risk of Low Performing Rating	-0.002	-0.002	-0.003	-0.002	0.001	-0.002	-0.000
	[0.002]	[0.002]	[0.002]	[0.002]	[0.002]	[0.002]	[0.002]
Risk of Recognized Rating	0.004	0.003	0.003	0.008*	0.009*	0.009*	0.011**
	[0.004]	[0.004]	[0.004]	[0.003]	[0.004]	[0.004]	[0.004]

Notes: Each column is a single regression of the indicated outcome on the set of variables from equations (1) in the paper, which includes controls for cubics in 8th grade math and reading scores, dummies for male, black, Hispanic, and free/reduced price lunch, each student's percentile rank on the 8th grade exams within their incoming 9th grade cohort, year fixed effects, and school fixed effects. Standard errors are block bootstrapped at the school level. Each coefficient gives the impact of being in a grade cohort that has a positive estimated risk of being rated Low-Performing or Recognized. The reference category is grade cohorts for whom the estimated risk of receiving an Acceptable rating rounds up to 100 percent. See the text for details on the construction of the ratings prediction. The outcomes in Panel A are indicator variables that are equal to one if a student was enrolled in any public (and after 2003, any private) institution in the state of Texas in the indicated year. The outcomes in Panel B are annual earnings in the 5th through 11th years after the first time a student enters 9th grade (which we refer to as the age 19 to 25 years), for all students who were not enrolled in any postsecondary institution in the indicated year. The outcomes in Panel C are indicator variables that are equal to one if a student had zero reported earnings and was not enrolled in any postsecondary institution in the indicated year. * = sig. at 5% level; ** = sig. at 1% level or less.

Figure A1





Notes: The X axis plots each high school overall pass rate on the 10th grade math exam, while the Y axis plots the same value for the lowest scoring subgroup in the school. Texas' accountability policy rates schools based on the lowest scoring subgroup. The dashed lines are locally weighted regressions.

Figure A2

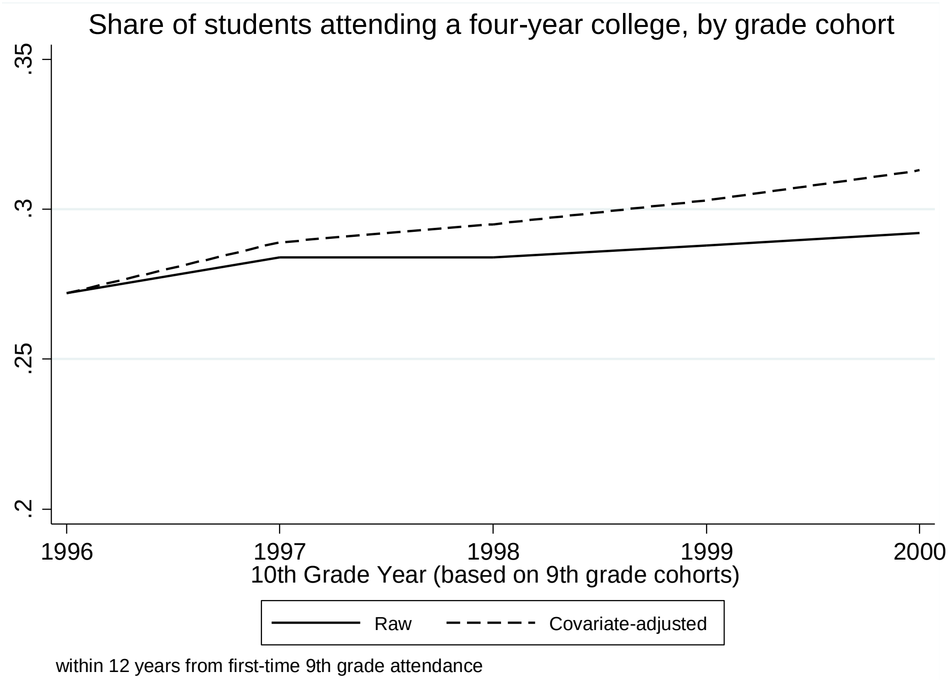
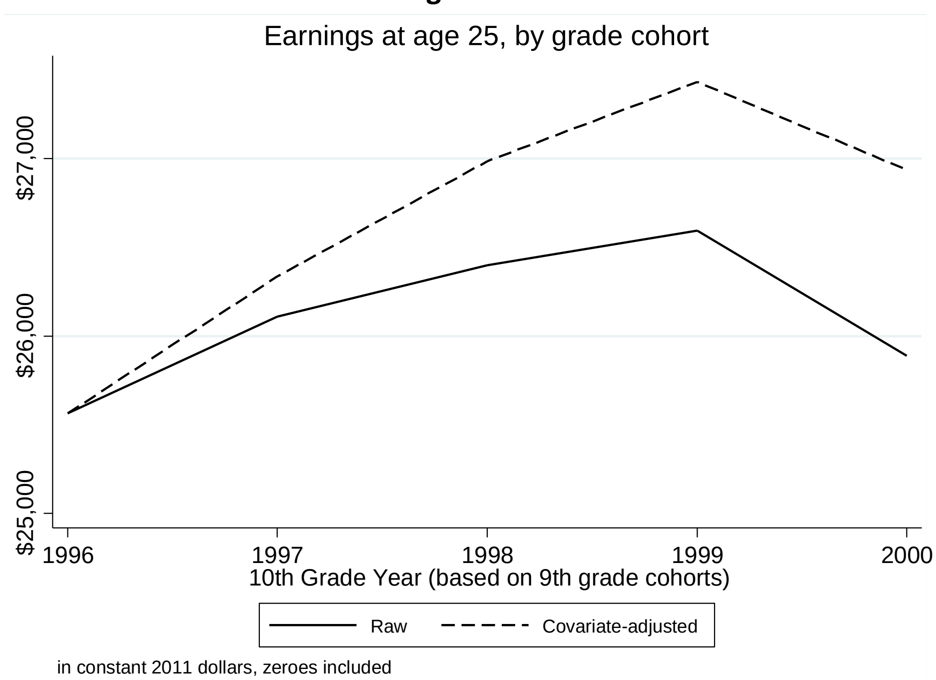
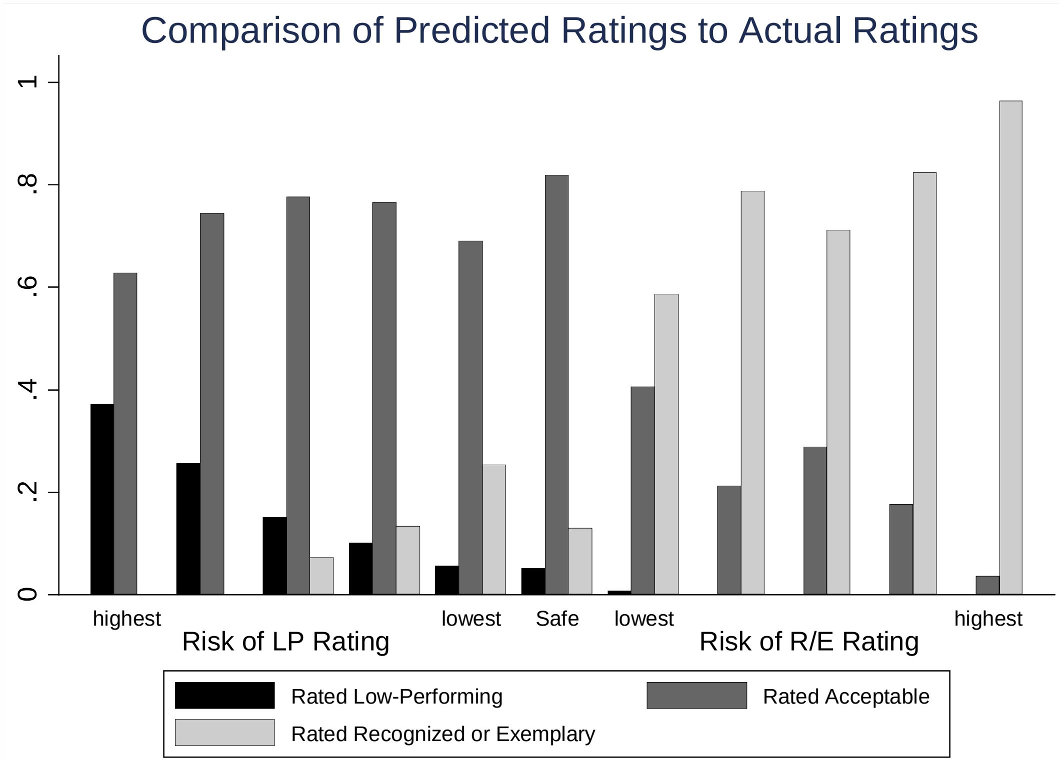


Figure A3



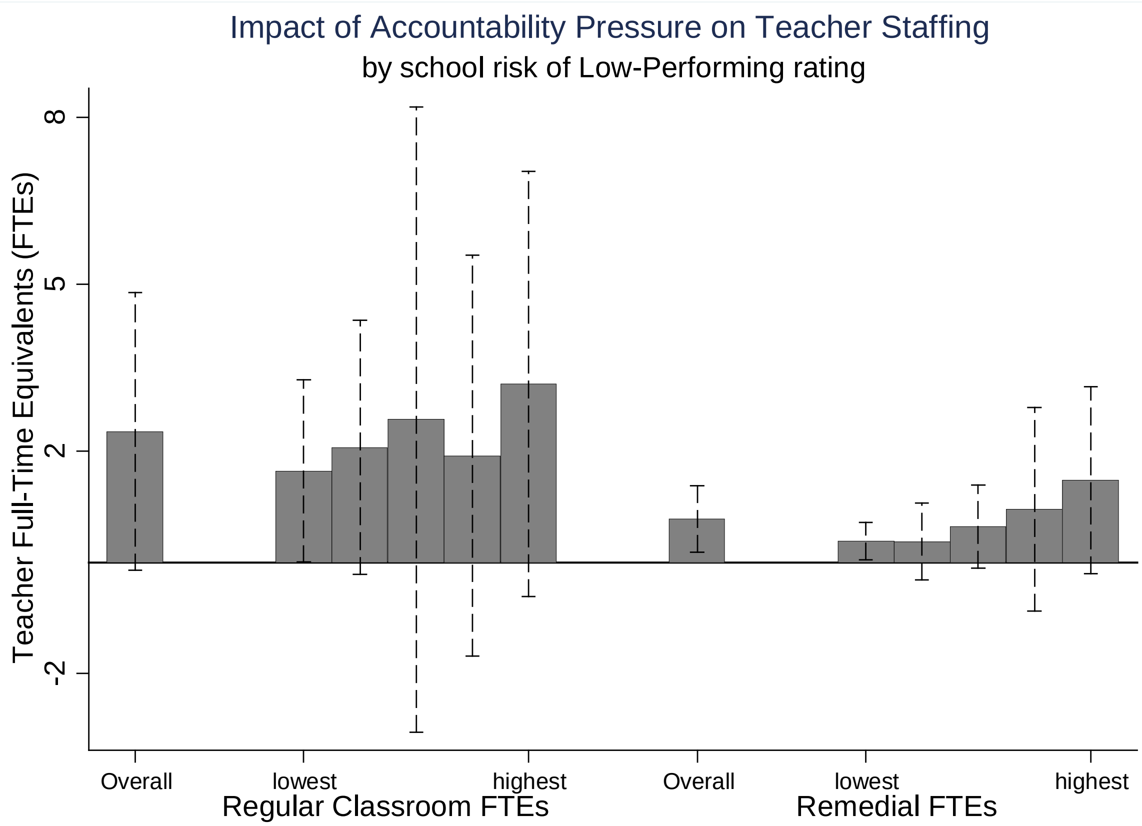
Notes: The figure above shows time trends in four-year college attendance and earnings in the state of Texas as of the 11^{th} year after the first time a student enters 9^{th} grade. Students with zero reported earnings are included in the calculation. The dashed line presents results from a regression of each outcome on year fixed effects, controlling for race, gender, free or reduced price lunch eligibility, and English Language Learner (ELL) and special education status. All covariates are measured as of 8^{th} grade.

Figure A4



Notes: This figure presents the share of school-cohorts in each predicted risk quintile that actually received the indicated accountability ratings from the Texas Education Agency (TEA). See the text for details on the construction of predicted ratings.

Figure A5



Notes: This figure presents coefficients and associated 95 percent confidence intervals from a single estimate of a modified version of equation (2) in the paper, with separate coefficients for five quintiles of a school-cohort's estimated risk of being rated Low-Performing. Since the teacher FTE allocation results vary only at the school-cohort level, these models do not include separate results by students' baseline math scores. FTE stands for Full-time Equivalent. Coefficients for schools that are on the margin of being rated Recognized are included in the model but not presented here. We also present the overall results next to each set of estimates by risk quintile.