NBER WORKING PAPER SERIES

SCHOOL ACCOUNTABILITY, POSTSECONDARY ATTAINMENT AND EARNINGS

David J. Deming
Sarah Cohodes
Jennifer Jennings
Christopher Jencks

Working Paper 19444
http://www.nber.org/papers/w19444

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
September 2013

We wish to thank Dick Murnane, Dan Koretz, Felipe Barrera-Osorio, Andrew Ho, Marty West, Todd Rogers, Kehinde Ajayi, Josh Goodman, David Figlio, Jonah Rockoff, Raj Chetty, John Friedman and seminar participants at the NBER Summer Institute, CESifo, Harvard Economics Labor Lunch and the QPAE lunch in the Graduate School of Education 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 rest of the staff at the University of Texas at Dallas Education Research Center for making this research possible. The conclusions of this research do not necessarily reflect the opinions or official positions of the Texas Education Agency, the Texas Higher Education Coordinating Board, or the State of Texas. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

School Accountability, Postsecondary Attainment and Earnings
David J. Deming, Sarah Cohodes, Jennifer Jennings, and Christopher Jencks
NBER Working Paper No. 19444
September 2013
JEL No. I20,I24

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). We find that high 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 - and large declines in attainment and earnings for low-scoring students - of pressure to achieve a higher accountability rating.

David J. Deming
Harvard Graduate School of Education
Gutman 411
Appian Way
Cambridge, MA 02139
and NBER
david_deming@gse.harvard.edu

Sarah Cohodes
John F. Kennedy School of Government
Harvard University
79 JFK Street
Cambridge, MA 02138
cohodes@fas.harvard.edu

Jennifer Jennings
New York University
Department of Sociology
295 Lafayette Street, 4th Floor
New York, New York 10012
jj73@nyu.edu

Christopher Jencks
Kennedy School
Harvard University
79 JFK Street
Cambridge, MA 02138
christopher_jencks@harvard.edu

An online appendix is available at:
http://www.nber.org/data-appendix/w19444

Today's schools must offer a rigorous academic curriculum to prepare students for the rising skill demands of the modern economy (Levy and Murnane, 2004). 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 to schools on the basis of their students' performance on the exams.

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, Linn 2000, 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 2011). 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 2012, 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 empirical strategy compares students within the same school, but across cohorts which faced different degrees of accountability pressure. Variation in accountability pressure arises because of changes in the ratings thresholds over time, as well as changes in the 8th grade test scores and demographics of each school's incoming 9th grade cohort. The focus on grade cohorts is possible because high schools gave high-stakes tests only in 10th grade. Our results are valid causal estimates of the impact of accountability pressure only if changes in predicted accountability pressure *across grade cohorts in a high school* are uncorrelated (conditional on covariates) with unobserved factors that also affect 10th grade scores and long-run outcomes. We test this assumption in a variety of ways, such as investigating trends in time-varying resources and pre-accountability test scores by school and subgroup. We conduct falsification tests using test scores at younger ages. We also estimate a specification that only uses policy variation in the passing standard for identification. Our results are robust to these specification checks, as well as to alternative sample restrictions and definitions of accountability pressure.

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 impacts are larger for students with low baseline achievement. Later in life, these students are more likely to attend and graduate from a four-year college, and they have higher earnings at age 25. We also find greater gains for poor and minority students with low baseline achievement, whose scores are most likely to "count" toward a higher accountability rating. However, schools that were close to receiving a rating above the "Acceptable" threshold (called "Recognized") responded differently. Instead of improving math

achievement, they classified more low-scoring students as eligible for special education and thus exempt from the "accountable" pool of test-takers. Perhaps as a consequence, we find relatively large declines in educational attainment and earnings among low-scoring students in these schools.

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 2012; Neal and Johnson 1996, Currie and Thomas 2001, Chetty et al 2011). We also find impacts on math course credit accumulation that mirror both the positive and negative impacts of accountability pressure, which is consistent with changes in mathematics coursework or knowledge as a plausible mechanism (Levine and Zimmerman 1995, Rose and Betts 2004, Goodman 2012). Broadly, our results indicate that school accountability led to long-run gains for students 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. Nonetheless, we caution that our results are specific to a particular time period, state and grade level. 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 do not imply that these strategic responses do not occur, or 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. NCLB mandated annual reading and math testing in grades 3 through 8 and at least once in high school, and required states to rate schools on the basis of test performance overall and for key subgroups, and to sanction schools that failed to make adequate yearly progress (AYP) toward state goals (e.g. Hanushek and Raymond 2005).

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 on the impact of high-stakes accountability on low-stakes test performance 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, Linn 2000, Klein et al 2000, Jacob 2005). Researchers studying the implementation of accountability in Texas and other settings have found evidence 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, at the expense of actions that do not increase test scores but that may be 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, what is the correlation between the actions that they take to raise test scores and the resulting changes in attainment, earnings and other long-run outcomes?[2]

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, Wong, Cook and Steiner 2011, Allen and Burgess 2012).[3]

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 2011, Dee and Jacob 2012). Increasing achievement is only one of many possible ways that schools and teachers may affect students (Chetty, Friedman and Rockoff 2012, 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

[2] For example, suppose some schools were not teaching students very much at all prior to accountability. It could simultaneously be true that while these schools are engaging in a variety of strategic behaviors, students are nonetheless better off in the long-run. On the other hand, if these schools were already doing a good job prior to accountability, perhaps because they already felt accountable to parents, test-based accountability could lead purely to a wasteful increase in compliance behavior, with no long-run impact on students or even negative impacts. The key point is that the existence of strategic behavior, and the variety of margins along which schools adjust, tells us very little about the net impact of accountability on students' long-run outcomes.

[3] The balance of the evidence suggests that gains from accountability are generally larger in math than in reading, and are concentrated among schools and students at the lower end of the achievement distribution (Carnoy and Loeb 2002, Jacob 2005, Figlio and Rouse 2006, Ladd and Lauen 2010, Figlio and Loeb 2011).

et al 2006, Cullen, Jacob and Levitt 2006, Kemple and Willner 2008, Booker et al 2009, Deming 2009, Chetty et al 2011, Deming 2011, Deming, Hastings, Kane and Staiger 2013).

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, and their records include scores on high-stakes exams in math, reading and writing, as well as measures of attendance, graduation, transfer and dropout. 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).

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). Students were allowed up to eight chances to retake the exam between the spring of the 10th grade year and the end of 12th grade (Martorell and Clark 2010). Due to the number of allowed retakes, and the fact that retakes were often scaled and administered differently, we primarily use students' scores from the first time they take the 10th grade exams. 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 whom they expect might fail the exam. 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. In 2003, the THECB began collecting similar information from private not-for-profit colleges in Texas.[4] 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.[5] 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.[6]

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, or that employment in another state.

[4] The addition of new postsecondary data in 2003 causes some cohort differences in base attendance and degree completion rates. However, the inclusion of year fixed effects in our analysis should difference out any overall cohort trends in college attendance, including those that result from differences in data coverage. Still, as a check, we rerun our main results with only public institutions included and find very similar results.

[5] 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.

[6] 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. Attendance and degree receipt at 2-year college after that time was considerably more common, which is one reason that we focus on four-year colleges in our main analysis.

We augment the TSP data with archival records of school and subgroup-specific exit exam pass rates from the pre-accountability period, back to 1991.[7] These data allow us to test for differential pre-accountability trends at the school and subgroup level.

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 this 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.

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 try, 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, and almost all did so on the first try.

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,

[7] We obtained these records by pulling them off the TEA website, where they are posted in .html format. School codes and subgroup pass rates allow us to clean the data and merge them onto the TSP records.

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 5 measures by 4 subgroups = 20 total performance indicators. 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 mean that math scores were almost always the main obstacle to improving a school's rating.[8] 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.[9]

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). 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. Schools rated as Low-Performing were also forced to

[8] Schools that achieved 25 percent pass rates on the TAAS received Acceptable ratings even if they failed the attendance and dropout provisions. Moreover, dropouts were counted "affirmatively" so that students who simply disappeared from the high school did not count against the rating. This led to unrealistically low dropout rates (e.g. Haney 2000, Vasquez Heilig and Darling-Hammond 2008).

[9] Appendix Table A1 presents descriptive statistics for high schools by the accountability ratings they received over our sample period. About 30 percent of schools received a Low-Performing rating at least once in five years, while 40 percent of schools were rated Acceptable in all five years. Not surprisingly, schools that received higher ratings (Recognized or Exemplary) had very few minority students and had high average test scores. 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. The often large distance from the 45 degree line shows that some schools have high overall pass rates yet still face pressure because of low scores among particular subgroups. Note that the disparity between school-wide and subgroup pass rates (reflected by the slope of the dashed lines in each picture) shrinks over time, suggesting that lower-scoring subgroups improve relatively more in later years.

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 sanctions were rarely used, dismissal or transfer of teachers and principals in Low-Performing schools was somewhat more common (Evers and Walberg 2002, Lemons, Luschei and Siskin 2004, Mintrop and Trujillo 2005).

In many ways, the Texas accountability system was 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. Some other states, particularly in the South, began implementing Texas-style provisions in the 1990s. When NCLB passed, 45 states already published report cards on schools and 27 rated or identified low-performing schools (Figlio and Loeb 2011). Accountability policies have changed somewhat since 2002, with a handful of states moving toward test score growth models (using a "value-added" approach). However, these changes require a Federal waiver of NCLB requirements. 26 states currently have high school exit exams. Approximately 18 states had such exams during the period covered by our study, with Texas falling roughly in the middle in terms of the exams' difficulty (Dee and Jacob 2007, Education Week 2012).

IV. Empirical Strategy

Our empirical strategy hinges on three critical facts about school accountability in Texas. First, schools faced strong incentives to improve TAAS pass rates, particularly in math. Since the passing standard required to avoid a Low-Performing rating rose by five percentage points per year from 1995 to 2000, even schools that were far from the threshold in earlier years often faced pressure in later years. Second, in any given year, schools could receive an Acceptable rating for a wide range of possible pass rates (e.g. between 30 and 70 percent in 1996, between 45 and 80 percent in 1999). This suggests that schools might feel more "safe" in some years than others, depending on the baseline levels of achievement in particular cohorts. Third, the policy of assigning ratings based on the lowest scoring subgroup created strong incentives for schools to focus on improving the test scores of poor and minority students with low baseline levels of achievement.[10]

[10] Neal and Schanzenbach (2010) find that high-stakes testing in Chicago led to increases in scores for "bubble" students in the middle of the prior test score distribution, but found no gains for the lowest-scoring students. However, only 6 percent of students in the bottom decile in Chicago passed the exam the following year. In our data, more than 20 percent of students in the bottom decile of the 8th grade math exam still passed the 10th grade exam. For this reason, we do not distinguish between low-scoring students and "bubble" students.

Our identification strategy uses these policy details to compare similar students in schools and cohorts that faced differing degrees of accountability pressure. In particular, we hypothesize that schools feel less pressure to increase achievement when the lowest-scoring subgroup is very likely to land in the middle of the "Acceptable" range. Moreover, we can use the yearly variation in passing standards to make within-school, across-cohort comparisons. The key challenge is modeling the accountability pressure faced by high schools in each year.

We develop a specification that predicts the share of students in a grade cohort and subgroup that will pass the 10th grade exit exam, using students' demographics and achievement prior to high school. Our approach is similar in spirit to Reback, Rockoff and Schwartz (2011), 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.[11]

We begin by estimating logistic regressions of indicator variables for passing each of the 10th grade exit exams on demographic characteristics, fully interacted with a third order polynomial in 8th grade reading and math scores, using the full analysis sample. We also include year fixed effects in each model, which explicitly differences out common yearly trends in prior student achievement. We then aggregate these individual predictions up to the school-subgroup-test level to estimate the "risk" that schools will receive a particular rating. Our calculation of the school rating prediction proceeds in three steps:

1. We form mean pass rates and standard errors at the year-school-subgroup-test subject level, using the predicted values from the student level regressions discussed above.
2. Using the yearly ratings thresholds, we integrate over the distribution of test-by-subgroup pass rates to get predicted accountability ratings for each subgroup and test within a school and year.
3. Finally, since Texas' accountability rating system specifies an overall school rating that is based on the lowest subgroup-test pass rate, we can multiply the conditional probabilities for each subgroup and test together to get a distribution of possible ratings at the school level.[12] [13]

[11] The main differences between our prediction model and theirs are 1) we use individual student data rather than school-year aggregates; 2) their policy variation is across states, while ours is within the same state across different years.

[12] Consider the following example for a particular high school and year. 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 conditional probability of an R is zero. The probability of an A

It is important to keep in mind that the calculation is entirely *ex ante* – that is, it is based only the 8th grade characteristics of first-time 9th grade students in each cohort. It is not intended to predict the school's actual accountability rating. Rather, we think of it as modeling a process in which high schools observe the demographic composition and academic skills of their incoming 9th grade cohorts, make a prediction about how close they will be to the boundary between two ratings categories, and decides on a set of actions that they hope will achieve a higher rating. Thus, we are modeling the school's perception of accountability pressure rather than constructing an accurate prediction. For example, if our model predicts that a school is very likely to receive a Low Performing rating but it ends up with an Acceptable rating, this might be because the school is particularly effective overall, or that it responded to accountability pressure by increasing student achievement.

Appendix Figure A2 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. Appendix Table A2 presents a transition matrix that shows the relationship between schools' predicted ratings in year T and year T+1. In any given year, almost half of all schools have an estimated 100 percent risk of being rated Acceptable. We refer to these schools as "safe". About 63 percent of the schools that are "safe" in year T are also "safe" in year T+1. 26 percent have some estimated risk of being rated Low-Performing the following year. The remaining 11 percent are close to achieving a Recognized rating. Thus, while ratings in one year predict the following year with some accuracy, there is plenty of year-to-year variation in predicted ratings across cohorts in the same school.

Our main results come from regressions of the form:

$$Y_{isc} = \propto + \delta I[pr(LP)_{sc} > 0] + \theta I[pr(R)_{sc} > 0] + \beta X_{isc} + \gamma_c + \eta_s + \varepsilon_{isc} \quad (1)$$

$$Y_{ifsc} = \propto + \sum_{f=0}^{1} \delta_f I[pr(LP)_{sc} > 0] + \sum_{f=0}^{1} \theta_f I[pr(R)_{sc} > 0] + \beta X_{ifsc} + \gamma_c + \eta_s + \varepsilon_{ifsc} \quad (2)$$

---

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 most cases math is the only relevant test since scores are so much lower than reading and writing.

[13] 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 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 A3 shows that the results are very similar when we use a simpler prediction model that only controls for the polynomial in 8th grade scores, and does not explicitly model the Required Improvement provision. Schools could also receive higher ratings after the fact through appeals and waivers, and we do not model this explicitly

We examine the impact of accountability pressure on outcomes Y for student i in 8th grade math score in school s and grade cohort c. Equation 2 allows the impact to vary by an indicator f for whether a student failed a prior 8th grade exam in either math or reading. Our main independent variables of interest are indicators for being in a *school and grade cohort* that has some positive probability being rated Low Performing or Recognized/Exemplary, with a (rounded) 100 percent probability of an Acceptable rating as the left-out category. Note that the probabilities come directly out of the procedure described above, and as such they should not be thought of as a school's actual probability of receiving a rating – we refer to these calculations from here onward as the "risk" of receiving a particular rating. The X vector includes controls for a third order polynomial in students' 8th grade reading and math scores plus demographic covariates such as race, gender and eligibility for free or reduced price lunch.[14] We also control for cohort fixed effects (γ_c) and school fixed effects (η_s). Since our main independent variables are nonlinear functions of generated regressors, we adjust the standard errors by block bootstrapping at the school level.[15]

We also estimate models that allow the impacts to vary by terciles of a school's estimated risk r of being rated Low-Performing or Recognized, which ranges from 0 to 1 because of the logit specification:

$$Y_{irsc} = \propto + \sum_{r=0}^{2} \delta_r I\left[\frac{1}{3}r < pr(LP)_{sc} \leq \frac{1}{3}(r+1)\right] + \sum_{r=0}^{2} \theta_r I\left[\frac{1}{3}r < pr(R)_{sc} \leq \frac{1}{3}(r+1)\right] + \beta X_{irsc} + \gamma_c + \eta_s + \varepsilon_{irsc} \quad (3)$$

An ideal empirical strategy 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. Neal and Schanzenbach (2010) compare students with similar prior scores before and after the implementation of test-based accountability in Chicago Public Schools.[16] While we do make use of school and subgroup aggregate data from earlier years to test for the possibility of differential pre-accountability trends, we do not have student-level data until 1994. In most states, accountability metrics are continuous but ratings are assigned based on sharp cutoffs. Several papers have exploited this policy feature to identify the impact of receiving a low school rating in

[14] Additional covariates include special education, limited English proficiency, and a measure of *ex ante* rank within high school and cohort using the average of a students' 8th grade math and reading scores. This last measure addresses concerns that students in grade cohorts with lower average test scores may do better because they have a higher relative rank.

[15] Another possible approach is to employ a parametric correction following Murphy and Topel (1985). Estimates that use either a Murphy-Topel (1985) adjustment or no adjustment are very similar to the main results.

[16] Using data from Texas, Reback (2008) finds achievement gains for students who are more likely to contribute to a school's accountability rating, compared to other students in the same school and year. By using students in the same school and grade cohort whose scores don't count toward the rating as a control group, this identification strategy identifies relative changes and intentionally differences out the level effect of accountability pressure.

a regression discontinuity (RD) framework (e.g. Figlio and Lucas 2004, Chiang 2009, Rockoff and Turner 2010, Rouse, Hannaway, Goldhaber and Figlio 2013).[17]

Our results are valid causal estimates of the impact of accountability pressure only if changes in the predicted risk of receiving a rating are uncorrelated (conditional on covariates) with unobserved factors that also affect within-school, across-cohort variation in 10th grade scores and long-run outcomes. The school fixed effects approach accounts for persistent differences across schools in unobserved factors such as parental education, wealth, or school effectiveness. However, if there are contemporaneous shocks that affect both the timing of a school's predicted rating and the relative performance of the grade cohort on tests and long-run outcomes, our results will be biased.

We test for the possibility of contemporaneous shocks in Appendix Table A4 by regressing a school's predicted risk of being rated Low-Performing on the cohort characteristics in our prediction model, school and year fixed effects, and time-varying high school inputs such as principal turnover, teacher pay and teacher experience. Appendix Table A5 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. The results in Appendix Tables A4 and A5 show that variation in accountability pressure across cohorts arises primarily from two sources – 1) changes in the ratings thresholds over time, as shown in Figure 1; and 2) fluctuations in the prior test scores and demographics of incoming grade cohorts.[18] 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.[19]

[17] We do not pursue the RD approach, for three reasons. First, in Texas the consequences of receiving multiple low-performing ratings relative to just one were not well-defined, and depended on local discretion (TEA 1994, Haney 2000, Reback 2008). Second, since the high-stakes test is given only in 10th grade in Texas, students have already finished their first year of high school before the school's rating is known. Any important response to accountability pressure that occurs prior to the end of the first year of high school would not be measured by the RD approach. Third, the RD approach (at least in Texas) suffers from the limitation that the control group also faces significant pressure to improve student achievement.

[18] A third potential source of variation is fluctuation around the minimum size requirements – at least 10 percent of the cohort - for a subgroup to "count" toward a school's accountability rating. Only 25 percent of students were in schools where the number of "accountable" subgroups changed across the five grade cohorts. Even within this minority of schools, relatively few faced a different predicted rating when the set of "accountable" subgroups changed. In Appendix Table A6, we find a very similar pattern of results when we restrict the sample to the remaining 75 percent of students in schools where the "accountable" subgroups did not change.

[19] The p-value on an F-test for the joint significance of time-varying high school characteristics is 0.300. The p-value on an F-test for the joint significance of all 16 interactions (pass rates overall and for black, Latino and poor students, for 1991 through 1994) between a linear trend and pre-accountability pass rates is 0.482. Saturating the model in Appendix Table A5 with all possible trend interactions increases the R-squared from 0.28 to 0.35 (with p=0.000) without school fixed effects, but from 0.618 to only 0.621 once school fixed effects are included. As a final check, in Appendix Table A7 we show that the main results of the paper are very similar when we include the full set of trend interactions directly in the estimating equation.

A related possibility is that our estimates of accountability pressure are correlated with responses to accountability pressure in earlier grades (e.g. in feeder middle schools). This is a concern to the extent that covariates such as prior test scores and school fixed effects do not fully capture differences in student ability across cohorts, although the bias could be in either direction. We test for this possibility in Appendix Table A8 by including 7th grade math scores as an outcome in our main estimating equations. Conditional on 8th grade test scores, we find no statistically significant difference in the 7th grade math achievement of students in grade cohorts that were at risk of receiving a Low-Performing rating. Finally, in Appendix Table A9 we show that our results are very similar when we exclude from our sample schools that display a clear trend in predicted ratings, in either direction.[20]

Another potential threat to the validity of our approach comes from measurement error in a school's predicted rating. Specifically, if cohort fluctuations in 8th grade test scores are driven by random shocks that cause group underperformance on the 8th grade test, our results could be biased by mean reversion. We address the possibility of mean reversion in two ways. First, we take the predicted pass rates estimated in step 1 above, and average them across all five grade cohorts within a school. This gives us one set of predictions for all five cohorts, and isolates policy-induced changes in the pass rate thresholds over time as the only source of variation in predicted ratings.[21] This approach addresses mean reversion by eliminating any identifying variation that comes from cohort fluctuations in prior scores. Appendix Table A10 shows that our main results are generally very similar when we average predicted pass rates across cohorts. Second, in Appendix Table A11 we show that many of the main results of the paper still hold even without school fixed effects. In this case, mean reversion is not a concern because we are also making use of cross-school variation in predicted ratings. Relatedly, the pattern of results is very similar when we identify schools that face accountability pressure using only 8th grade pass rates, rather than estimating risk directly.[22]

Ultimately, our identifying assumption is not directly testable. Yet it is worth pointing out that the omission of any factor that is positively correlated with 8th grade test scores and long-run outcomes

[20] Specifically, we exclude schools that never "switch back" between prob(LP)>0 and A or prob(R)>0 and A over the five year period. For example, a school with a five-year predicted ratings pattern of LP-A-A-A-A would be excluded by this rule, but a school with a pattern of LP-A-A-A-LP would not be excluded, because it "switched back" to LP in a later year.

[21] For example, we might find that the lowest-scoring subgroup in a school has a predicted pass rate of 35 percent, averaged across all five cohorts. This places them 5 percentage points below the threshold in 1996, right at the threshold in 1997, and five points above it in 1998 (see Figure 1).

[22] We calculate 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 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 these specifications, shown in Appendix Table A12, we find a very similar patterns of results to Tables 2 through 3.

biases the estimated impact of the risk of receiving a Low-Performing rating *downward*. If a particular grade cohort within a high school has relatively lower 8th grade achievement, then all else equal that cohort is more likely to be rated Low-Performing. If the grade cohort also has a relatively lower level of parental engagement, and parental engagement improves long-run outcomes conditional on the covariates in our models, the estimated impact of being rated Low-Performing will be biased downward.

V. Results

V.1 10th Grade Test Scores and Other High School Outcomes

Table 2 estimates the impact of accountability pressure on 10th grade test scores and other outcomes in high school. Panels A and B present results from separate estimates of equations (1) and (2) respectively. Columns 1 through 3 show the impact of accountability pressure on 10th grade math exams. Since pass rates on the 10th grade math exam were almost always the key obstacle to a school's achieving a higher rating, the impacts on these outcomes are a direct test of the influence of accountability pressure. For framing purposes, it is also important to note that we do not view increases in pass rates or overall achievement on the high stakes exams as necessarily measuring true learning. Rather, we consider them as a signal that schools were responding to their incentives. As stated earlier, these gains could reflect true increases in learning, teaching to the test, cheating, or a combination of these and other mechanisms.

In Column 1, we see that 9th graders in cohorts with a risk of being rated Low-Performing were about 0.7 percentage points more likely to pass the 10th grade math exam in the following year, compared to similar students in the same school whose cohort was not at risk of being rated Low-Performing. In Panel B, we see that the impact is more than twice as large (1.5 percentage points) for students who failed a math or reading exam in 8th grade. Both impacts are statistically significant at the less than 1 percent level. In contrast, we find no impact of accountability pressure to achieve a Recognized rating. Column 2 shows the impact of accountability pressure on the probability that a student ever passes the 10th grade math exit exam. Since students are allowed to retake the exit exam up to eight times, it is not surprising that the impacts on ever passing the exam are smaller.

Column 3 shows the impact of accountability pressure on 10th grade math scale scores. We find an overall increase of about 0.27 scale score points (equivalent to about 0.04 standard deviations) for students in schools that were at risk of receiving a Low-Performing rating. The impacts are larger for students with lower baseline scores (0.44 scale score points, or about 0.07 SDs), but are still significantly greater than zero for higher-scoring students (0.18 points, or about 0.025 SDs).

We find negative and statistically significant impacts of accountability pressure on students with low baseline scores in schools that were close to achieving a Recognized rating. These students score about 0.4 scale score points *lower* overall, and are 1.9 percentage points *less* likely to ever pass the 10th grade math exit exam. We find no statistically significant impact of accountability pressure to achieve a Recognized rating on higher-scoring students.

The results in the first three columns of Table 2 show that if schools were at risk of receiving a Low-Performing rating, they responded by increasing math achievement, while schools that were close to achieving a Recognized rating did not. However, another way that high schools can improve their rating is by manipulating the pool of "accountable" students. The outcome in Column 5 is an indicator that is equal to one if a student is receiving special education services in the 10th grade year, *conditional on not receiving special education services in 8th grade.* Special education students in Texas were allowed to take the TAAS, but their scores did not count toward the school's accountability rating. They also were not required to pass the 10th 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. Panel B of Column 5 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. We also find a smaller (0.5 percentage points) but still highly significant *decrease* in special education classification for high scoring students.

Column 6 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. We also find a concomitant *decline* in high school graduation in schools that face pressure to achieve a Recognized rating. Finally, Column 7 shows results for 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 0.10 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 sum, we find a very different pattern of responses to accountability pressure along different margins. Schools that were at risk of receiving a Low-Performing rating responded by increasing the math scores of all students, with particularly large gains for students who had previously failed an 8th

grade exam. These students also had higher reading scores, were more likely to graduate from high school, and accumulated significantly more math credits. In Appendix Table A13, 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.[23]

On the other hand, schools that were close to achieving a Recognized rating responded not by improving math achievement, but by classifying more low-scoring students as eligible for special education and thus excluded from the "accountable" pool of test-takers. Our finding of a decrease in math pass rates (1.9 percentage points, Column 2, Panel B) combined with a borderline statistically significant increase in high school graduation (1.3 percentage points, Column 6, Panel B) makes sense in this light - special education students were not required to take the 10th grade exit exams to graduate, and thus did not have to learn as much math to make it through high school. This is also reflected in the decline in math course credit accumulation found in Panel B of Column 7. Overall, it appears that schools responded to pressure to achieve a Recognized rating in part by exempting marginal low-scoring students from the normal high school graduation requirements.[24]

V.2 Long-Run Outcomes - Postsecondary Attainment and Earnings

Table 3 presents results for the long-run impact of accountability pressure on postsecondary attainment and labor market earnings. As a reminder, each of these outcomes is measured for the universe of public school students in Texas, even if they transferred or left high school altogether prior to graduation. We measure college attendance within an eight year window following the first time a student enters 9th grade, and we can measure degree receipt through age 25 even for the youngest cohort of students.

Columns 1 through 3 show results for postsecondary attainment. We find that 9th graders in cohorts with a risk of being rated Low-Performing were about 1 percentage points more likely to attend college. Column 2 shows that the increase is concentrated entirely among four-year colleges. In column 3 we

[23] We find small but statistically significant increases in reading scores, and no impact on writing scores. We also find small but statistically significant declines in grade retention, transfers to other regular public schools, and transfers to alternative schools for students with behavior problems – see Appendix Table A13 for details. While past work has found that schools respond to accountability pressure by strategically retaining students, schools in Texas that are trying to avoid a Low-Performing rating have little incentive to retain students since the passing standard rises by 5 percentage points each year.

[24] This pattern of increases in special education classification is large enough to be readily apparent in descriptive statistics by school rating. The share of special education students in schools that received a Recognized rating was just over 12 percent, which is actually higher than the share of special education students in schools that received a Low-Performing rating (11 percent).

find a small but highly significant increase (0.4 percentage points, p<0.01) in bachelor's degree receipt. Like the test score results in Table 2, the increase in postsecondary attendance and degree completion is larger for students who previously failed an 8th grade exam. Increases in postsecondary attainment are proportionally much larger for lower-scoring students, who are about 14 percent more likely to attend a four-year college and 12 percent more likely to earn a bachelor's degree. Nonetheless, we also find small increases in attainment for higher scoring students.[25]

We find small and statistically insignificant negative impacts on attainment for schools that face pressure to achieve a Recognized rating. However, the impacts for low-scoring students are very large and negative. Students who previously failed an 8th grade exam are 2.8 percentage points less likely to attend a four-year college and 1.3 percentage points less likely to earn a BA when they are in grade cohorts that are close to achieving a Recognized rating.

Columns 4 through 6 present results for annual earnings in the 9th, 10th and 11th calendar years after the first time a student entered 9th grade. Hereafter we refer to these earnings as occurring in the age 23, 24 and 25 years. Column 7 includes cumulative earnings over the entire three year period, and in Column 8 the outcome is an indicator for being "idle", defined as having zero positive earnings and never being enrolled in school at any point over the age 19-25 period. For 9th graders in cohorts with a risk of being rated Low-Performing, we find small increases in annual earnings at ages 23 through 25 of around $150, or around 1 percent of the baseline mean in each year.[26] The estimates for each year are marginally statistically significant (all three at the 10 percent level, and one at the five percent level), and the estimates for the combined three-year period in Column 6 are statistically significant at the five percent level. Like the results for postsecondary attainment, the impacts are larger for students with lower baseline achievement (about 1.4 percent of the baseline mean for earnings at age 25).

In contrast, we find no overall impact on labor market earnings of accountability pressure to achieve a Recognized rating, with statistically significant and large *declines* in earnings for low-achieving students

[25] Beginning with the graduating class of 1998, Texas high school students who were ranked in the top 10 percent of their class were automatically admitted to all public universities in the state. Despite the impact that the Texas Top 10 percent law may have had on college enrollment in the state, it is unlikely to affect our results, for three reasons. First, we control for class rank of prior test scores directly in our main specifications, and we find no difference when we allow for nonlinear functions of class rank or for a different impact in the top decile. Second, our results are strongest for students who previously failed an 8th grade exam, and very few (less than 7 percent) of these students subsequently end up in the top 10 percent of their class. Third, the law affects all five of the cohorts in our sample (first time 9th graders in 1995 would graduate from high school in 1999), and our identification strategy hinges on across-cohort differences in accountability pressure. It is unlikely that the Texas Top 10 percent law is somehow differentially applicable to students in grade cohorts within a school that are more likely to be rated Low-Performing.

[26] We find very similar results when we estimate impacts on log earnings. Our results are somewhat larger when we exclude students with zero earnings from the calculation, or when we exclude students who are enrolled in a postsecondary institution during that same year.

in these schools. Students who previously failed an 8th grade exam earned about than $700 less at age 25 when they attended schools that were close to achieving a Recognized rating.

In Column 8, we find no impact of accountability pressure on the probability of being "idle" between the ages of 19 and 25 for any students. The standard errors are sufficiently small that we can rule out differences of more than one percentage point in most cases.[27] Since our data only cover postsecondary attendance and employment in the state of Texas, students could potentially have college attendance and labor market earnings in other states. 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. However, it is unclear which way the bias would go. If the students who are induced by accountability pressure to attend four-year colleges and obtain BAs are more likely to move out of state after college, or if their earnings are low because they just graduated, the results would be biased downward. On the other hand, if accountability pressure makes students more likely to attend college out-of-state and stay there, our results may be biased upward. In Appendix Tables A14 and A15, we find that our results are robust to imputing missing earnings values and to allowing the impacts to vary for schools that send large shares of students out-of-state.[28]

Ideally we would measure labor market earnings beyond age 30, when nearly all schooling is complete and earnings are more stable (e.g. Haider and Solon 2006). Measuring earnings at age 25 is less than ideal, particularly since accountability pressure also appears to affect postsecondary attendance and persistence. To try to understand how the attainment and earnings results fit together, we estimate equation (1) for three mutually exclusive outcomes at ages 19 through 25 – (1) enrolled in any college; (2) earnings, excluding students with zero earnings or who are enrolled in college (regardless of their earnings); and (3) "idle" – defined as zero earnings and not enrolled in college. Appendix Table A16 shows three important findings from this exercise. First, the impact of accountability pressure to avoid a Low-Performing rating on labor market earnings for students who are not enrolled in school is positive and statistically significant in every year from age 20 to age 25. Second, the impact on postsecondary enrollment becomes very small around age 24 (around 0.2 percentage

[27] We also find no impacts on other measures of labor force participation, such as an indicator for zero earnings at ages 19-25, or indicators for idle / zero earnings in individual years between ages 19 and 25.

[28] In Appendix Table A14 we impute mean earnings by prior achievement and school type for students who are not currently enrolled in school and have zero reported earnings. We test the sensitivity of the results to imputation by adding and subtracting one standard deviation. We find that the results for schools on the Low-Performing margin are robust to adding or subtracting a standard deviation of earnings for missing values. However, the results for schools on the Recognized margin are much more sensitive to imputation. In Appendix Table A15 we show that our results are actually slightly larger when we exclude schools that send more than 10 percent of graduates to out-of-state colleges (the TSP data include a match of the graduating classes of 2008 and 2009 to the National Student Clearinghouse, which includes out-of-state enrollment). Finally, in results not reported, we also find no difference in impacts for schools that are close to the Texas border.

points). Moreover, overall rates of college attendance drop sharply around age 23, with fewer than 10 percent of all students still enrolled in college at age 25. Among low-scoring students, only 35 percent ever enroll in any college and only 10 percent ever enroll in a four-year college. Taken together, the evidence suggests that accountability pressure to avoid a Low-Performing rating has an impact on earnings independent of its impact on postsecondary attainment. Finally, we note a statistically significant increase in the probability of being "idle" beginning at age 23 for students in schools that face pressure to achieve a Recognized rating. While not shown, these increases are concentrated among low-scoring students.

V.3 Variation by school risk and evidence on changes in inputs

Table 4 shows results for the main outcomes of the paper for estimates of equation (3), where we allow the impacts to vary by terciles of the school's predicted accountability rating. In general, we find slightly larger impacts for schools that are more likely to receive a Low-Performing rating. But the differences are relatively small and we are never able to reject equality across categories. We also find that the overall negative impacts in schools that were close to achieving a Recognized rating are relatively constant across categories of predicted risk.

A strict interpretation of the predicted ratings might imply that some schools would not respond to pressure, because they estimate that they are almost certain to receive a Low-Performing rating. However, administrators and teachers in these schools were likely to be quite concerned about job security. Staff could potentially be fired for lack of effort, or conversely could be retained if some progress toward goals was being made. Reback, Rockoff and Schwartz (2011) find that teachers in schools that were well below the AYP margin were still significantly more likely to report that they were "concerned about job security due to student performance", relative to schools that were projected to make AYP or be close to the margin.

The lowest-performing schools were often targeted with significant external resources such as improvement efforts from the district office, focused remediation outside of school hours, and external funding (e.g. Scheurich, Skrla and Johnson 2000, Evers and Walberg 2002, Lemons, Luschei and Siskin 2004). Craig, Imberman and Perdue (2013) find that school districts in Texas allocate additional funds for instruction to schools that receive lower accountability ratings. While the TSP data provide very limited information on staffing, we are able to measure the total number of teacher full-time equivalents (FTEs) by school and year. We also know broad categories of classrooms to which teachers are assigned. Appendix Figure A3 presents estimates of the impact of accountability pressure on the allocation of

regular classroom and remedial classroom teacher FTEs, using the setup in equations (1) and (3). We find some evidence that schools respond to the risk of being rated Low-Performing by increasing staffing, particular in remedial classrooms. This is consistent with studies of accountability pressure in Texas and elsewhere that find increases in instructional time and resources devoted to low-performing students (e.g. Evers and Walberg 2002, Lemons, Luschei and Siskin 2004, Hamilton et al 2007, Rouse, Hannaway, Goldhaber and Figlio 2013). However, we are limited in this exercise by our inability to match individual teachers to subjects or grades, and by the possibility of spillover to adjacent grade cohorts.

One less than ideal feature of our empirical strategy is that we are sometimes comparing 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 A17 and A18 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.

V.4 Is the impact of accountability pressure greater for targeted subgroups?

As we discussed in Section III, the Texas accountability system created strong incentives for schools to focus on low-scoring students in targeted subgroups. We investigate the impact of accountability pressure on targeted subgroups by estimating:

$$Y_{isc} = \propto + \delta I[pr(LP)_{sc} > 0] * Target_{isc} + \theta I[pr(R)_{sc} > 0] * Target_{isc} + \beta X_{isc} + \emptyset_{sc} + \varepsilon_{isc} \quad (4)$$

This setup modifies equation (1) in three ways. First, we augment the X vector with eight race (white or other vs. black or Latino) by poverty by prior test score (failed either 8th grade exam or passed both) indicators. Second, we include school-by-cohort fixed effects, which difference out the level impact of accountability pressure and restrict our identifying variation to changes in the *distribution* of outcomes across cohorts. Third, we interact the accountability pressure treatment indicators with indicators for the most "targeted" group – poor minority student with low baseline test scores. This specification asks whether the *difference* in outcomes between targeted subgroups and all other students in the same grade cohort is greater in a year when the school faces accountability pressure. We

also restrict the analysis sample to schools that have enough poor and/or minority students to meet the minimum size requirements described in Section III.

The results for the key outcomes in the paper are in Table 5. In Columns 1 and 2, we see that gains in 10th grade math for poor, low-scoring minorities are significantly greater *relative to all other students* in grade cohorts that are at risk of receiving a Low-Performing rating. We also find statistically significant relative gains in four-year college attendance, BA receipt and earnings at age 25 for poor, low-scoring minority students. These results confirm that the overall gains shown in Tables 2 and 3 from accountability pressure to avoid a Low-Performing rating are concentrated among poor minority students with low levels of baseline achievement. Because school ratings are assigned based on the lowest-scoring subgroup and indicator, schools under accountability pressure had strong incentives to target these particular students. In contrast, we find no relative change in the outcomes of targeted subgroups for schools that face pressure to achieve a Recognized rating.

VI. Discussion and Conclusion

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 (2012) 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 (2001) 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.

One possible explanation for the long-run impacts is that passing the math exit exam at a higher rate led to mechanical increases in high school graduation, which increased earnings through the "sheepskin" effect of having a high school diploma. Martorell and Clark (2010) use data on students who barely passed exit exams in two states (including Texas) to estimate the signaling value of a high school diploma, and find that the diploma has little or no impact on earnings for students with similar 10th grade test scores. Moreover, low-scoring students in schools that were close to achieving a Recognized rating were actually *more* likely to graduate from high school despite experiencing significant declines in

earnings (Table 2, Panel B, Column 6). Thus, we conclude that "sheepskin" effects of acquiring a high school diploma are unlikely to explain our results.

A second possible mechanism is that accountability pressure affected students' knowledge of mathematics, which in turn affected labor market earnings. Accountability pressure could have caused students to learn more math in two ways – 1) changes in class time and instructional resources devoted to math instruction, and 2) changes in students' later course-taking patterns, due either to increases in passing the exit exam "on time" or to exemption from test-taking requirements. Using cross-state variation in the timing of high school graduation requirements, Goodman (2012) finds that an additional year of math coursework increases annual earnings by between 4 and 8 percent, with larger impacts for students who attend schools with high shares of nonwhites, and who have relatively low levels of skill. 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 (2013) find increases in high school graduation and college attendance for students who are assigned to a "double dose" Algebra I class in 9th grade. While the impacts we find on math course credit accumulation in Table 2 are broadly consistent with this story, the evidence for increased math knowledge as a causal mechanism is only speculative.

It is also possible that test score gains from accountability pressure could reflect gains in student motivation, perseverance or other "non-cognitive" skills. Heckman, Stixrud and Urzua (2007) argue that unlike cognitive skills, non-cognitive skills remain malleable through adolescence. Jackson (2012) finds evidence that teachers have important influences on non-cognitive skills that lead to increased attainment and earnings later in life. We have no direct evidence on the impact of accountability pressure on non-cognitive skills.

In Appendix Tables A19 through A21 we examine heterogeneous treatment effects by gender, limited English proficiency (LEP), and urbanicity.[29] For schools that face pressure to avoid a Low-Performing rating, we find very similar test score impacts by gender, but larger impacts on postsecondary attainment and earnings for males. We also find somewhat larger long-run impacts for LEP students, though we cannot reject equality across categories. Finally, we find that the negative long-run impacts in schools that face pressure to achieve a Recognized rating are concentrated in urban areas. While we do not report Appendix results by race and poverty, the targeted subgroup results in

[29] For urbanicity, we interact the main treatment indicators from equation (1) with an indicator variable that is equal to one if a student attends school in one of the six large urban districts in Texas (Houston, Dallas, Fort Worth, Austin, San Antonio and El Paso).

Table 5 show that the impacts of accountability pressure to avoid a Low-Performing rating are larger for poor and minority students with low baseline test scores.

Overall, we find that the long-run impact of accountability pressure in Texas high schools was very different for schools along different margins. Overall math achievement increased in schools that were at risk of receiving a Low-Performing rating, with larger gains for students with low baseline achievement. Later in life, these students were more likely to attend and graduate from a four-year college, and they had higher earnings at age 25. These long-run impacts are corroborated by gains in high school graduation and increased accumulation of math credits, including credits in advanced subjects such as Algebra II and Pre-Calculus, which the 10^{th} grade exit exam does not cover.

On the other hand, schools that were close to achieving a Recognized rating responded not by improving math achievement, but by classifying more low-scoring students as eligible for special education and thus exempt from the "accountable" pool of test-takers. Low-scoring students in these schools also accumulated fewer high school math credits. Such changes in high school experiences may have played some role in the large decline in the postsecondary attainment and earnings of these students later in life. Why do schools on the margin of being rated Low-Performing not also respond with strategic exemptions? One possibility is that they enroll too many low-achieving students to meet rising passing standards purely through gaming behavior.

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.

However, our results are specific to a particular time period, state and grade level. Moreover, our approach is not designed to measure the impacts of long-run, structural changes that schools make in response to accountability pressure. Our findings provide an estimate of the net impact of schools' responses to test pressure along a variety of margins, including strategic student exemption, focusing on tested subjects, teaching to the test, and manipulating the eligible test-taking pool. Our findings do not imply that these strategic responses do not occur, nor do they imply that school accountability in Texas was optimally designed (Neal, 2013). Improving the design of the next generation of school accountability policies is an important problem for future work.

References

Abdulkadiroglu, A. et al., 2011. Accountability and flexibility in public schools: Evidence from Boston's charters and pilots. *Quarterly Journal of Economics*, 126(2), pp.699–748.

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

Angrist, J.D., Pathak, P.A. & Walters, C.R., 2011. *Explaining charter school effectiveness*, Cambridge, MA: National Bureau of Economic Research.

Baker, G.P., 2002. Distortion and risk in optimal incentive contracts. *Journal of human resources*, pp.728–751.

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

Belfield, C.R. et al., 2006. The High/Scope Perry Preschool Program Cost–Benefit Analysis Using Data from the Age-40 Followup. *Journal of Human Resources*, 41(1), pp.162–190.

Betts, J.R. & Shkolnik, J.L., 1999. The behavioral effects of variations in class size: The case of math teachers. *Educational Evaluation and Policy Analysis*, 21(2), pp.193–213.

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)*.

Champion, S., 2011. *Increased Accountability, Teachers' Effort, and Moonlighting*, Stanford University Graduate School of Business.

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., 2011. *The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood*, Cambridge, MA: National Bureau of Economic Research.

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., 2013. *Intensive Math Instruction and Educational Attainment: Long-Run Impacts of Double-Dose Algebra*, Harvard Kennedy School.

Cullen, J.B., Jacob, B.A. & Levitt, S.D., 2005. The impact of school choice on student outcomes: an analysis of the Chicago Public Schools. *Journal of Public Economics*, 89(5), pp.729–760.

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.

Currie, J. & Thomas, D., 1999. *Early Test Scores, Socioeconomic Status and Future Outcomes*, National Bureau of Economic Research, Inc.

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. et al., 2011. School choice, school quality and academic achievement. *NBER Working Paper*, 17438.

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.

Duflo, E., Dupas, P. & Kremer, M., 2011. Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya. *American Economic Review*, 101(5), pp.1739–1774.

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

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. & Page, M.E., 2002. School choice and the distributional effects of ability tracking: does separation increase inequality? *Journal of Urban Economics*, 51(3), pp.497–514.

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.

Fryer, R.G., 2011. *Injecting Successful Charter School Strategies into Traditional Public Schools: Early Results from an Experiment in Houston*, National Bureau of Economic Research.

Gamoran, A. & Mare, R.D., 1989. Secondary school tracking and educational inequality: Compensation, reinforcement, or neutrality? *American Journal of Sociology*, pp.1146–1183.

Goodman, J., 2012. *The Labor of Division: Returns to Compulsory Math Coursework*, Harvard University, John F. Kennedy School of Government.

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.

Hamilton, L.S., Berends, M. & Stecher, B.M., 2005. *Teachers' Responses to Standards-Based Accountability*, Santa Monica, CA: RAND Corporation.

Haney, W., 2000. The Myth of the Texas Miracle in Education. *Education Policy Analysis Arcives*, 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.

Heckman, J.J., Stixrud, J. & Urzua, S., 2006. The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior. *Journal of Labor Economics*, 24(3), pp.411–482.

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.

Imberman, S.A. & Lovenheim, M., 2013. *Does the Market Value Value-Added? Evidence from Housing Prices after a Public Release of School and Teacher Value-Added*, CESifo Group Munich.

Jackson, C.K., 2012. *Non-Cognitive Ability, Test Scores, and Teacher Quality: Evidence from 9th Grade Teachers in North Carolina*, Cambridge, MA: National Bureau of Economic Research.

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. et al., 1991. The Effect of High-stakes Testing on Achievement: Preliminary Findings about Generalization Across Tests. In Annual Meetings of the American Educational Research Association. Chicago, IL.

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.

Ladd, H.F. & Lauen, D.L., 2010. Status versus growth: The distributional effects of school accountability policies. *Journal of Policy Analysis and Management*, 29(3), pp.426–450.

Lauen, D.L. & Gaddis, S.M., 2012. Shining a light or fumbling in the dark? The effects of NCLB's subgroup-specific accountability on student achievement. *Educational Evaluation and Policy Analysis*, 34(2), pp.185–208.

Leventis, A., 1997. *Cardiac surgeons under the knife*, Princeton, NJ: Princeton University.

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

Linn, R.L., 2000. Assessments and accountability. *Educational researcher*, 29(2), pp.4–16.

Martorell, P. & Clark, D., 2010. *The signaling value of a high school diploma*, RAND Corporation.

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.

Neal, D., 2010. Aiming for Efficiency Rather than Proficiency. *The Journal of Economic Perspectives*, 24(3), pp.119–131.

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., 2012. *Stop using one assessment system to pursue two objectives*, Chicago, IL: University of Chicago.

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.

Prendergast, C., 1999. The provision of incentives in firms. *Journal of economic literature*, 37(1), pp.7–63.

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., 2011. *Under pressure: Job security, resource allocation, and productivity in schools under NCLB*, Cambridge, MA: National Bureau of Economic Research.

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. et al., 2007. *Feeling the Florida heat? How low-performing schools respond to voucher and accountability pressure*, National Bureau of Economic Research.

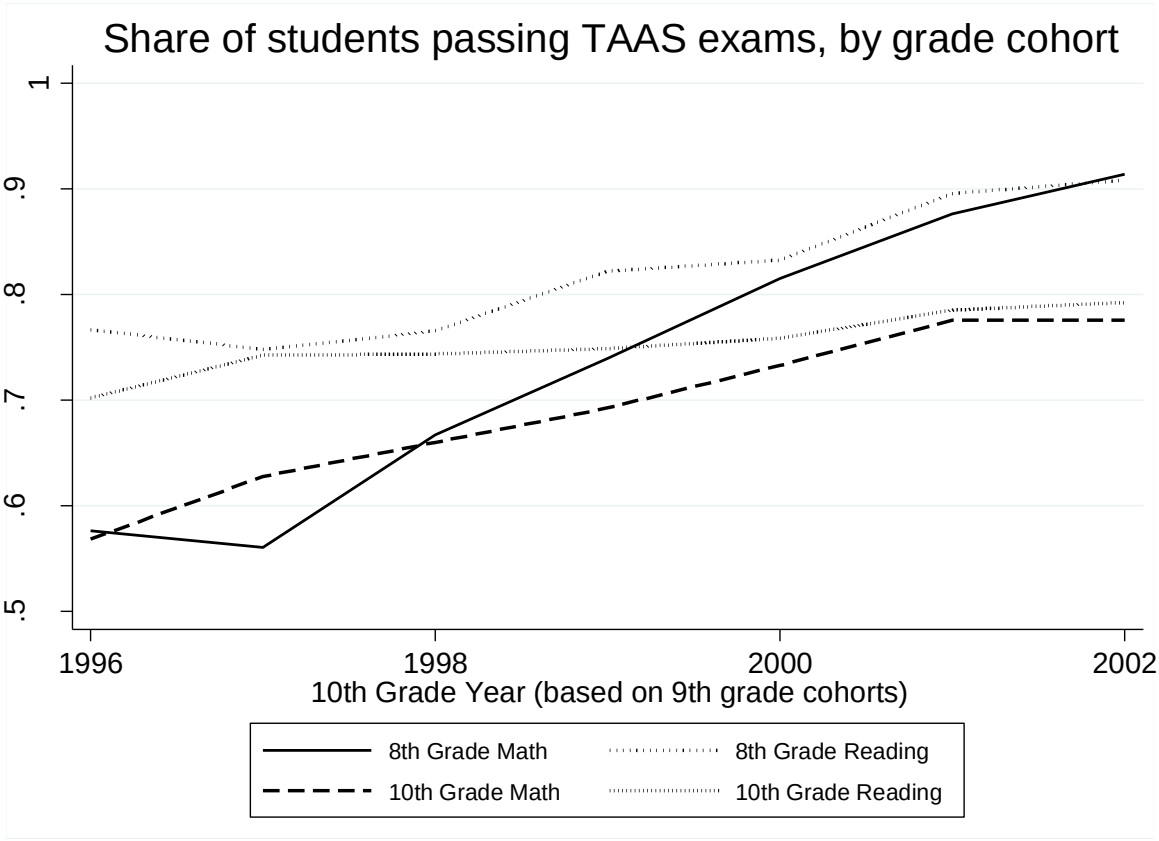
Skrla, L., Scheurich, J.J. & Johnson, J.F., 2000. Equity-driven achievement-focused school districts: A report on systemic school success in four Texas school districts serving diverse student populations. *Austin, TX: Charles A. Dana Center*.

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.

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.

Table 1 - Descriptive Statistics

	Overall	By Student Race / SES			By 8th Grade Math Score	
		Black	Latino	FRPL	Passed both 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 (TLI>=70)	0.79	0.66	0.69	0.66		
High school outcomes						
10th grade scale score - math	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, but not in 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
Annual Earnings, Age 25	$17,683	$13,605	$16,092	$14,580	$19,818	$14,019
Idle, Ages 19 to 25	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 on High School Outcomes

	10th Grade Math			Non-Test High School Outcomes		
	Passed Test "On Time"	Ever Passed	Scale Score	Special Ed. In 10th	Graduated High School	Total Math Credits
Panel A	(1)	(2)	(3)	(4)	(5)	(6)
Risk of Low Performing Rating	0.007**	0.004	0.265**	-0.001	0.009**	0.060**
	[0.003]	[0.003]	[0.080]	[0.001]	[0.002]	[0.015]
Risk of Recognized Rating	-0.001	-0.008	-0.238	0.002	-0.009*	0.011
	[0.003]	[0.005]	[0.127]	[0.001]	[0.004]	[0.016]
Panel B						
Risk of Low Performing Rating						
Failed an 8th grade exam	0.015**	0.008*	0.435**	-0.003**	0.010**	0.073**
	[0.006]	[0.004]	[0.125]	[0.001]	[0.003]	[0.016]
Passed 8th grade exams	0.004	0.002	0.181*	0.000	0.009**	0.051**
	[0.002]	[0.004]	[0.075]	[0.000]	[0.002]	[0.017]
Risk of Recognized Rating						
Failed an 8th grade exam	-0.008	-0.019**	-0.395*	0.024**	0.013	-0.106**
	[0.009]	[0.007]	[0.173]	[0.004]	[0.007]	[0.023]
Passed 8th grade exams	-0.007	-0.005	-0.215	-0.005**	-0.016**	0.044*
	[0.003]	[0.006]	[0.121]	[0.001]	[0.004]	[0.018]
Sample Size	697,728	887,713	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, 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. 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". The outcome is Column 4 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.

Table 3: Impact of Accountability Pressure on Long-Run Outcomes

	Postsecondary Outcomes			Annual Earnings				Idle
	Attend Any College	Attend 4 Yr College	BA Degree	Age 23	Age 24	Age 25	Age 23-25	Age 19-25
Panel A	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Risk of Low Performing Rating	0.011**	0.012**	0.0043**	140*	148	172	459*	-0.001
	[0.002]	[0.002]	[0.0011]	[63]	[77]	[97]	[221]	[0.002]
Risk of Recognized Rating	-0.006	-0.005	-0.0041	48	-50	-121	-123	0.004
	[0.004]	[0.004]	[0.0037]	[129]	[198]	[198]	[508]	[0.003]
Panel B								
Risk of Low Performing Rating								
Failed an 8th grade exam	0.009*	0.014**	0.0060**	148	176*	194*	518*	-0.001
	[0.004]	[0.002]	[0.0016]	[78]	[75]	[89]	[203]	[0.002]
Passed 8th grade exams	0.013**	0.010**	0.0032*	133	127	153	412	-0.002
	[0.003]	[0.003]	[0.0015]	[75]	[78]	[99]	[275]	[0.002]
Risk of Recognized Rating								
Failed an 8th grade exam	0.002	-0.028**	-0.0129**	-135	-379	-707**	-1,221**	0.003
	[0.007]	[0.006]	[0.0045]	[157]	[211]	[212]	[451]	[0.004]
Passed 8th grade exams	-0.009*	0.002	-0.0018	101	43	49	193	0.004
	[0.004]	[0.005]	[0.0039]	[129]	[181]	[155]	[455]	[0.003]
Sample Size	887,713	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 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. 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 outcomes in Columns 4 through 6 are annual earnings in the 9th through 11th years after the first time a student enters 9th grade (which we refer to as the age 23 to 25 years), including students with zero reported earnings. Column 7 gives total earnings over the age 23-25 period. Column 8 is an indicator variable that is equal to one if the student has zero reported earnings and no postsecondary enrollment over the age 19-25 period. * = sig. at 5% level; ** = sig. at 1% level or less.

Table 4: Impact of Accountability Pressure, by Terciles of Predicted Rating

	10th Grade Math		Four Year College		Earnings
Risk of Low-Performing Rating	Passed Test	Score	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 boostrapped 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". 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 5: Impact of Differential Accountability Pressure for Targeted Subgroups

	10th Grade Math		Four Year College		Earnings
	Passed Test	Score	Attend	BA	Age 25
Risk of Low-Performing Rating	(1)	(2)	(3)	(4)	(5)
Targeted Subgroup, Failed 8th	0.011*	0.279*	0.012*	0.008**	579**
Grade Exam	[0.005]	[0.134]	[0.005]	[0.002]	[141]
Risk of Recognized Rating					
Targeted Subgroup, Failed 8th	0.009	-0.370	-0.012	-0.006	-193
Grade Exam	[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 boostrapped 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. TThe 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". 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.