SCHOOL ACCOUNTABILITY, POSTSECONDARY ATTAINMENT, AND EARNINGS

David J. Deming, Sarah Cohodes, Jennifer Jennings, and Christopher Jencks*

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. Schools respond to the risk of being rated Low Performing by increasing student achievement on high-stakes exams. Years later, these students are more likely to have attended college and completed a four-year degree, and they have higher earnings at age 25. However, we find no overall impact of accountability pressure to achieve a higher rating, and large negative impacts on attainment and earnings for the lowest-scoring students.

I. Introduction

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

More than a decade after the passage of NCLB, however, 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 (Koretz & Barron, 1998; Klein et al., 2000; Carnoy & Loeb, 2002; Hanushek & Raymond, 2005; Jacob, 2005; Dee & Jacob, 2011; Reback, Rockoff, & Schwartz, 2014). There are many studies of strategic responses to accountability pressure, ranging from focusing instruction on marginal students, narrow test content and coaching, manipulating the pool of accountable students, boosting the nutritional content of school lunches, and teacher cheating (Haney, 2000; McNeil & Valenzuela, 2001; Jacob & Levitt, 2003; Diamond & Spillane, 2004; Figlio & Winicki, 2005; Booher-Jennings, 2005; Jacob, 2005; Cullen & Reback, 2006; Figlio & Getzler, 2006; Vasquez Heilig & Darling-Hammond, 2008; Reback, 2008; Neal & 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, 1975; 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, & Rockoff, 2014; Jackson, 2012).

While there are many goals of public schooling, test-based accountability is premised on the belief that student achievement gains will lead to long-run improvements in important life outcomes such as educational attainment and earnings. High-stakes testing creates incentives for teachers and schools to adjust their effort toward improving test performance in the short run. Whether these changes make students better off in the long run depends critically on the correlation between the actions that schools take to raise test scores and the resulting changes in earnings and educational attainment at the margin (Holmstrom & Milgrom, 1991; Baker, 1992; Hout & Elliott, 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 were then old enough to examine outcomes in young adulthood. High schools were rated by the share of tenth-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 (Haney, 2000; Cullen & Reback, 2006).

Our research design compares grade cohorts within a school that faced different degrees of accountability pres-

Received for publication February 19, 2014. Revision accepted for publication November 30, 2015. Editor: Asim I. Khwaja.

* Deming: Harvard University and NBER; Cohodes: Harvard University; Jennings: New York University; Jencks: Harvard University.

We thank Dick Murnane, Dan Koretz, David Figlio, Jonah Rockoff, Raj Chetty, John Friedman, and seminar participants at Harvard, Stanford, Columbia, the University of Wisconsin, the NBER Summer Institute, and CESifo for helpful comments. This project was supported by the Spencer Foundation and the William T. Grant Foundation. Very special thanks to Maya Lopuch for invaluable research assistance. We gratefully acknowledge Rodney Andrews, Greg Branch, and the staff of the UT–Dallas Education Research Center for making this research possible. The conclusions of this research do not necessarily reflect the opinions of the Texas Education Agency, the Texas Higher Education Coordinating Board, or the state of Texas.

A supplemental appendix is available online at http://www.mitpressjournals.org/doi/suppl/10.1162/REST_a_00598.

The Review of Economics and Statistics, December 2016, 98(5): 848–862

doi:10.1162/REST_a_00598

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

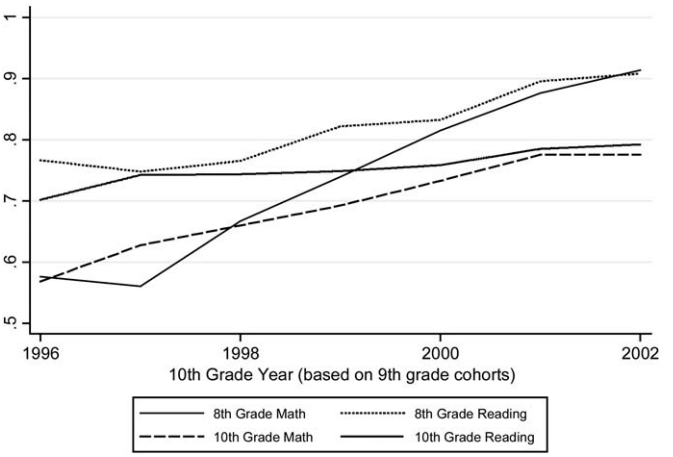
We find that students score significantly higher on the tenth-grade math exam when they are in a grade cohort that is at risk of receiving a Low-Performing rating. These students are more likely to graduate from high school on time and accumulate significantly more math credits, including in subjects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

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

We find that accountability pressure to avoid a Low-Performing rating leads to increases in labor market earnings at age 25 of around 1%. 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 et al., 2011, 2014; Neal & Johnson, 1996; Chetty et al., 2011). Broadly, our results indicate that school accountability led to long-run gains in schools that were at risk of falling below a minimum performance standard. Efforts to regulate school quality at a higher level (through the achievement of a Recognized rating), however, did not benefit students and may have caused long-run harm.

The accountability system Texas adopted 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 increase testing standards over time. Thus, our findings may have broad applicability to the accountability regimes that were rolled out in other states over this period. However, because we compare schools that face different degrees of pressure within the same high-stakes testing regime, our study explicitly differences out any common trend in outcomes caused by school accountability. We estimate the net impact of schools' responses along a variety of margins, including focusing on "bubble" students and subjects, teaching to the test, and manipulating the eligible test-taking pool. Our results are the net impact of schools' responses along a variety of margins and do not imply that school accountability in Texas was optimally designed (Neal, 2013).

FIGURE 1.—SHARE OF STUDENTS PASSING TAAS EXAMS, BY GRADE COHORT



The figure shows time trends in the share of students in Texas who pass the eighth- and tenth-grade exams in math and reading. Students are assigned to cohorts based on the first time they enter ninth grade.

II. Background

Beginning in the early 1990s, a handful of states, including 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 & Loeb, 2002; Hanushek & Raymond, 2005; Dee & Jacob, 2011; Figlio & 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 & Raymond, 2005; Figlio & Ladd, 2007; Figlio & Loeb, 2011). Under then Governor and future President George W. Bush, test-based accountability in Texas served as a template for the federal NCLB Act of 2002.

Figure 1 shows pass rates on the eighth- and tenth-grade reading and mathematics exams for successive cohorts of first-time ninth graders in Texas. Pass rates on the eighth-grade math exam rose from about 58% in the 1994 cohort to 91% in the 2000 cohort, only six years later. Similarly,

pass rates on the tenth-grade exam, a high-stakes exit exam for students, rose from 57% to 78%, 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. Research has found that scores on high-stakes tests improve, often dramatically, whereas performance on a low-stakes test with a different format but similar content improves only slightly or not at all, a phenomenon known as "score inflation" (Koretz & Barron, 1998; Klein et al., 2000; Jacob, 2005). Studies of the implementation of accountability in Texas and other settings have found that schools raised test scores by retaining low-performing students in ninth grade, classifying them as eligible for special education or otherwise exempt from taking the exam, and encouraging them to drop out (Haney, 2000; McNeil & Valenzuela, 2001; Jacob, 2005; Cullen & Reback, 2006; Figlio & Getzler, 2006; Vasquez Heilig & Darling-Hammond, 2008; McNeil, Coppola, Radigan, & Vasquez Heilig, 2008; Jennings & Beveridge, 2009).

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

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

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

A few studies have examined the impact of accountability in Texas on high school dropout, with inconclusive findings (Haney, 2000; Carnoy, Loeb, & Smith, 2001, McNeil et al., 2008; Vasquez Heilig & 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 American Community Survey (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, earn more as adults, or benefit over the long run in other important ways.

III. 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 prekindergarten through twelfth grade with a unique student identifier. From 1994 to 2003, state exams were referred to as the Texas Assessment of Academic Skills (TAAS). Students were tested in reading and math in grades 3 through 8 and again in grade 10, with writing exams also administered in grades 4, 8, and 10. Raw test scores were scaled using the Texas Learning Index (TLI), which was intended to facilitate comparisons across test administrations. For each year and grade, students are considered to have passed the exam if they reach a TLI score of 70 or greater. 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 (Klein et al., 2000).

Our analysis sample consists of five cohorts of first-time ninth-grade students from spring 1995 to spring 1999. The TSP data begin in the 1993–1994 school year, and we need eighth-grade test scores for our analysis. The first cohort with eighth-grade scores began in the 1994–1995 school year. Our last cohort began high school in 1998–1999 and took the tenth-grade exam in 1999–2000. We use these five cohorts because Texas's 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 ninth 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 & Valenzuela, 2001; Cullen & Reback, 2006; Jennings & Beveridge, 2009). Our solution is to assign students to the first high school they attend and to measure outcomes based on initial assignment. For example, if a student attends school A in ninth grade, transfers to school B in tenth grade and then graduates, she is counted as graduating from school A. This is similar in spirit to an intent-to-treat design.

High school students were required to pass each of the tenth-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 eighth-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 (Haney, 2000). Although students were allowed to retake the tenth-grade exit exams up to eight times, we use the first score only in our analysis. We also create an indicator variable equal to 1 if a student first took the test at the usual time for his or her ninth-grade cohort. This helps us test for the possibility that schools might increase pass rates by strategically retaining, exempting, or reclassifying students. Since the TSP data cover the entire state, we can measure graduation from any public school in the state of Texas, even if a student transfers several times, but we cannot track students who left the state.

The TSP links PK–12 records to postsecondary attendance and graduation data from the Texas Higher Education Coordinating Board (THECB). The THECB data contain records of enrollment, course taking, and matriculation for all students who attended public institutions in Texas. While the TSP data do not contain information about out-of-state college enrollment, less than 9% of graduating seniors in Texas who attend college do so out of state, and they are mostly high-scoring students.[1] 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.[2]

The TSP has also linked PK–12 records to quarterly earnings data for 1990 to 2010 from the Texas Workforce Commission (TWC). The TWC data cover 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. Our main outcomes of interest here are annual earnings in the age group 23 to 25 years (the full calendar years that begin nine to elven years after the student's first year in ninth grade). Since the earnings data are available through 2010, we can measure earnings in the age 25 year for the 1995 through 1999 ninth-grade cohorts. We also construct indicator variables for having any positive earnings in the age 19 to 25 years and over the seven years combined. Zero positive earnings could indicate a true lack of labor force participation, having unemployment insurance–ineligible earnings or employment in another state.

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

IV. Policy Context

Figure 2 summarizes the key Texas accountability indicators and standards from 1995 to 2002. Schools were grouped into one of four possible performance categories: Low-Performing, Acceptable, Recognized, and Exemplary. Schools and districts were assigned performance grades based on the overall share of students who passed TAAS exams in reading, writing, and mathematics, as well as attendance and high school dropout. Indicators were also calculated separately for four key subgroups—white, African American, Hispanic, and economically disadvantaged (based on the federal free lunch eligibility standard for pov-

[1] Authors' calculation based on a match of two graduating classes (2008 and 2009) in the TSP data to the National Student Clearinghouse (NSC), a nationwide database of college attendance.

[2] Our youngest cohort of student (ninth graders in spring 2001) had seven years after their expected high school graduation date to attend college and complete a B.A. While a small number of students in the earlier cohorts received a B.A. after year 7, almost none attended a four-year college for the first time after seven years.

TABLE 1.—DESCRIPTIVE STATISTICS

	Overall (1)	Black (2)	Latino (3)	Free Lunch (4)	Passed Eighth-Grade Exams (5)	Failed an Eighth-Grade Exam (6)
Eighth-grade covariates						
White/other	0.52			0.20	0.64	0.33
Black	0.14			0.19	0.09	0.21
Latino	0.34			0.61	0.27	0.46
Free lunch	0.38	0.54	0.68		0.29	0.55
Passed 8th math (TLI >= 70)	0.67	0.48	0.56	0.53		
Passed 8th reading	0.79	0.66	0.69	0.66		
High school outcomes						
10th grade math score	78.2	72.6	75.6	74.6	83.2	66.3
Passed 10th math on time	0.76	0.59	0.67	0.64	0.90	0.40
Ever passed 10th math	0.81	0.74	0.76	0.72	0.92	0.62
Passed 10th reading on time	0.88	0.75	0.77	0.75	0.95	0.51
Special ed. in 10th, not 8th	0.01	0.01	0.01	0.01	0.00	0.02
Total math credits	1.93	1.78	1.73	1.65	2.29	1.33
Graduated from high school	0.74	0.69	0.69	0.65	0.82	0.59
Later outcomes						
Attended any college	0.54	0.46	0.45	0.39	0.65	0.35
Attended four-year college	0.28	0.24	0.19	0.15	0.39	0.10
B.A. degree	0.13	0.09	0.09	0.07	0.18	0.05
Age 25 earnings (in 1000s)	17.7	13.6	16.1	14.6	19.8	14.0
No earnings/college, all years	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

The sample consists of five cohorts of first-time rising ninth graders in public high schools in Texas from 1995 to 1999. Postsecondary attendance data include all public institutions and, from 2003 onward, all nonprofit institutions in the state. Earnings data are drawn from quarterly unemployment insurance records from the state. Column 6 shows students who received a passing score on both the eighth-grade math and reading exams. Column 7 shows descriptive statistics for students who failed either exam. Students who are first-time ninth graders in year T and who pass a tenth-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 precalculus, for a total maximum value of four.

erty)—but only if the group constituted at least 10% 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 twenty total performance indicators (five measures by four subgroups). The TAAS passing standard for a school to receive an Acceptable rating rose by 5 percentage points every year, from 25% in 1995 to 50% in 2000. The standard for a Recognized rating also rose, from 70% in 1995 and 1996 to 75% in 1997, and 80% from 1998 onward. In contrast, the dropout and attendance rate standards remained constant over the period we study.

The details of the rating system meant that math scores were almost always the main obstacle to improving a school's rating. The lowest subgroup-indicator was a math score in over 90% 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.[3]

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

Schools rated as Low-Performing were also forced to undergo an evaluation process that carried potentially serious consequences, such as allowing students to transfer out, firing school leadership, and reconstituting or closing the school (TEA, 1994; Cullen & Reback, 2006). Although the most punitive sanctions were rarely used, surveys of principals and teachers indicate that threat of dismissal or transfer for failing to achieve a particular rating was more common (Toenjes & Garst, 2000; Evers & Walberg, 2002; Lemons, Luschel, & Siskin, 2003; Mintrop & Trujillo, 2005). Schools receiving high ratings were eligible for cash bonuses of up to $5,000 per school, and higher-rated schools did indeed receive additional funding as a perfor-

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

FIGURE 2.—ACCOUNTABILITY INDICATORS AND STANDARDS, 1995–2002

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

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

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

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

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

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

mance incentive (Izumi & Evers, 2002; Craig, Imberman, & Perdue, 2013).

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

The Texas accountability system was in many ways the template for NCLB. 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.

V. Measuring Accountability Pressure

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

Since the focus of our study is on long-run outcomes, we first examine descriptive evidence of trends in four-year college attendance and earnings at age 25 for the five cohorts of first-time ninth-grade students in Texas included in our study. Appendix figures A2 and A3 show that college attainment and earnings rose modestly for successive cohorts following the introduction of school accountability.[4]

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

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

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

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

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

We aggregate the individual predictions up to the school-subgroup-test level to estimate the risk that schools will receive a particular rating.[5] The prediction proceeds in three steps. First, we use the predicted values from the student-level regressions in equation (1) to form mean pass rates and standard errors at the school-subgroup-test level: $\overline{PassRate^t_{js}}$.

Second, we integrate over the mean pass rates and standard errors to get predicted accountability ratings for each subgroup, school, and test. For example, if the predicted pass rate for white students in school A on the math exam is 35% with a standard error of 2.5%, our model would predict the probability of receiving an Acceptable rating as 50% in 1997 (since the threshold was at exactly 35%) but only about 5% in 1998 (since the threshold increased to 40%, which is 2 standard deviations above the mean).

Third, since Texas's accountability rating system specifies an overall school rating that is based on the lowest subgroup-test pass rate, the probability that a school receives a rating of Acceptable or higher (and likewise for other ratings) is equal to the probability that every eligible subgroup rates Acceptable or higher on each test.[6] Thus, we simply multiply the probabilities for each subgroup and test together to get the probability that school j in cohort c receives a particular rating.

There are two sources of variation in perceived accountability pressure within schools over time: (a) changes over time in the ratings thresholds shown in figure 2 and (b) changes in the demographics and prior test scores of a school's incoming grade cohort, which may have altered the school's incentives to focus on particular subgroups.

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

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

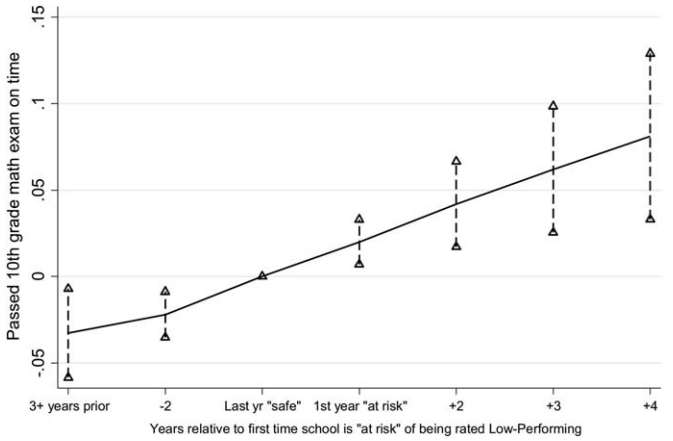
One limitation of computing a single prediction across cohorts is that it discards potentially useful variation, such as whether a particular subgroup is large enough to count toward the rating. Moreover, there is much less yearly variation along the Acceptable/Recognized rating threshold. Thus we also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

Our approach is similar in spirit to Reback et al. (2014), who compare students across schools that faced differential accountability pressure because of variation in state standards. We follow their approach in constructing subgroup

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

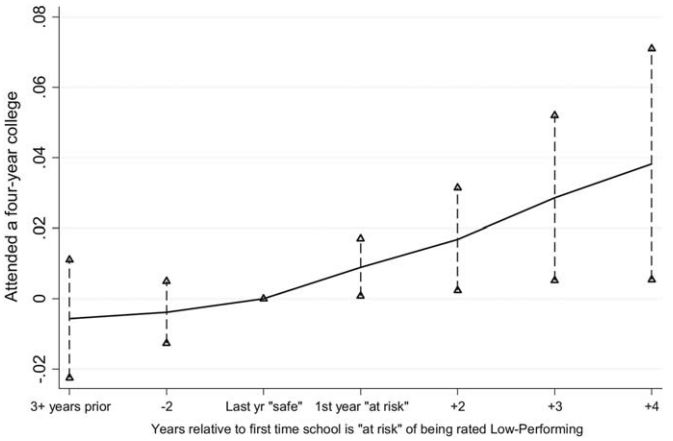
[6] Formally, $Pr(Rating \geq Acceptable)_{jc} = \prod_{s=1}^{S} \prod_{t=1}^{T} Pr(Rating \geq Acceptable)_{jstc}$. Consider the following example for a particular high school. Based on the predicted pass rates on the tenth-grade mathematics exam in math, reading, and writing for each of the four rated subgroups, we calculate that white students have a 96.3% chance of receiving an A rating and a 3.7% chance of receiving an R rating. Black students have an 18.8% chance of receiving an LP rating and an 81.2% chance of receiving an A rating. Latinos have a 4.7% chance of receiving an LP rating and a 95.3% chance for an A rating. Economically disadvantaged students have an 11.3% chance of receiving an LP rating and an 88.7% chance for an A rating. Since only whites have any chance of getting an R and the rating is based on the lowest-rated subgroup and test, the probability of getting an R is 0. The probability of an A rating is equal to the probability that all subgroups rate A or higher, which is $(0.963 + 0.037) \times (0.812) \times (0.953) \times (0.887) = 0.766$. 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 three tests to arrive at an overall probability, although in almost all cases, math is the only relevant test since math scores are so much lower than reading and writing.

FIGURE 3.—EVENT STUDY OF THE IMPACT OF ACCOUNTABILITY PRESSURE PASSED TENTH-GRADE MATH EXAM ON TIME



The figure presents point estimates and 95% confidence intervals from equation (2), where the outcome is whether a student passed the tenth-grade math exam on time (defined as one year after the first time a student enters ninth grade). We estimate the risk of a high school being rated Low-Performing based on the demographics and eighth-grade test scores of the grade cohort, combined with policy variation over time in the passing standards shown in figure 2 (see the text for details). We then define grade cohorts according to the first year each school was at risk of being rated Low-Performing, with the prior year as the baseline category. The regression includes school fixed effects.

FIGURE 4.—EVENT STUDY OF THE IMPACT OF ACCOUNTABILITY PRESSURE FOUR-YEAR COLLEGE ATTENDANCE



The figure presents point estimates and 95% confidence intervals from equation (2), where the outcome is whether a student attended a four-year college in Texas within eleven years of the first time he or she entered ninth grade. We estimate the risk of a high school being rated Low-Performing based on the demographics and eighth-grade test scores of the grade cohort, combined with policy variation over time in the passing standards shown in figure 2 (see the text for details). We then define grade cohorts according to the first year each school was at risk of being rated Low-Performing, with the prior year as the baseline category. The regression includes school fixed effects.

and subject-specific pass rate predictions based on measures of prior achievement. Several papers have studied the impact of actually receiving a low school rating, in a regression discontinuity (RD) framework (Figlio & Lucas, 2004; Chiang, 2009; Rockoff & Turner, 2010; Rouse, Hannaway, Goldhaber, & Figlio, 2013). Our approach focuses on the much larger group of schools that feel pressure to avoid a Low-Performing rating.

VI. Results

A. Event Study Using Policy Variation

For an initial graphical examination of accountability pressure, we align each school's predicted pass rates with the ratings threshold in an event study framework. Many schools, particularly in the early years, have a predicted pass rate that is far above the Low-Performing threshold; formally, their risk of being rated Low-Performing (according to the estimation procedure above) approaches 0. Depending on each school's average eighth-grade test scores and demographic characteristics, the model predicts that they will have some positive probability of being rated Low-Performing beginning in a particular year. Because the policy threshold for a Low-Performing rating only rises over time (see figure 2) and the prediction does not vary by cohort, once a school is "at risk," it remains so in subsequent cohorts. We organize schools according to the first year they have a positive probability of being at risk and estimate[7]

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

The X vector includes the same covariates as equation (1). However, in this specification, we have added school fixed effects (η_s) to account for persistent differences across schools in unobserved factors such as parental education, wealth, or school effectiveness. Intuitively, we ask whether the school-specific trend in outcomes varies systematically around the first year that a school was at risk of being rated Low-Performing. Because we have only five cohorts, the panel is unbalanced for any individual school. However, by controlling for cohort fixed effects (γ_c), we can obtain estimates for up to four years before and after the first year a school was at risk of being rated Low-Performing. Since our main independent variables are nonlinear functions of generated regressors, we adjust the standard errors by block bootstrapping at the school level here and for the remainder of the paper.[8]

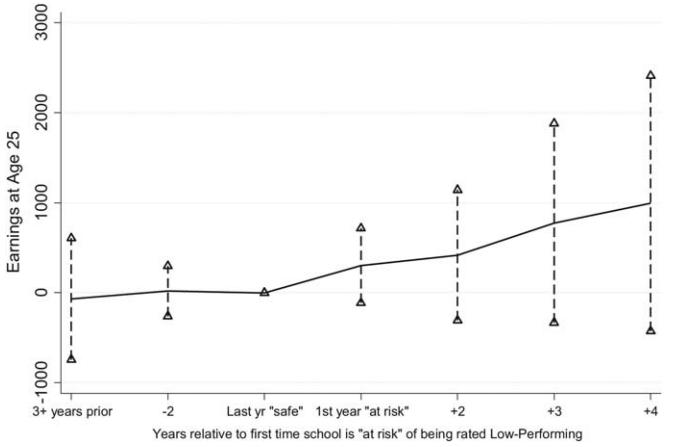
Figures 3 through 5 present results from equation (2) for the three key outcomes in the paper: tenth-grade math pass rates, four-year college attendance, and earnings in the eleventh calendar year after the student's ninth-grade cohort, which we refer to from here on as the "age 25" year. Estimates for each cohort include 95% confidence intervals, with the last year a school is "safe" as the baseline.

Figure 3 shows that students in the same school and with similar prior characteristics are about 2 percentage points more likely to pass the math exam on time (defined as the

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

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

FIGURE 5.—EVENT STUDY OF THE IMPACT OF ACCOUNTABILITY PRESSURE EARNINGS AT AGE 25



The figure presents point estimates and 95% confidence intervals from equation (2), where the outcome is earnings in the age 25 year, defined as the eleventh year after the first time a student enters ninth grade. Students with 0 reported earnings are included in the calculation. We estimate the risk of a high school being rated Low-Performing based on the demographics and eighth-grade test scores of the grade cohort, combined with policy variation over time in the passing standards shown in figure 2 (see the text for details). We then define grade cohorts according to the first year each school was at risk of being rated Low-Performing, with the prior year as the baseline category. The regression includes school fixed effects.

year after the first time a student enters ninth grade) if their grade cohort is the first to be at risk. This difference is statistically significant at the 95% level. However, we also find evidence of pretrends in math pass rates—the difference between two years and one year prior to being at risk is also statistically significant.

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

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

Our main interest is in long-run outcomes, which are less easily manipulated. Figure 4 presents results from equation (2) for four-year college attendance. Students in the first grade cohort at risk are about 0.9 percentage points more likely to attend a four-year college within eight years of the first time they enter ninth grade, and the difference is statistically significant at the 5% level. Moreover, we see no significant evidence of pretrends. We also see that the impact on four-year college attendance continues to rise for subsequent cohorts (who are also "at risk").

The pattern is very similar for earnings. Figure 5 shows that students in the first cohort at risk earn about $300 more at age 25 (this estimate is significant at the 10% level), and the impact rises gradually over time with no evidence of pretrends. Thus, it appears that the pressure to avoid a Low-Performing rating led to gains in postsecondary attainment and earnings for students in Texas. Note that point estimates are always less precise for years further away from the last year a school is "safe." This is because of the unbalanced nature of the panel: with only five cohorts, estimates at either end are identified using fewer years of data.[9]

B. Regression Results Using Policy Variation

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

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

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

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

Table 2 shows that students in grade cohorts that were at risk of being rated Low-Performing were about 0.8 percentage points more likely to pass the tenth-grade math exam on time (column 1) and scored about 0.3 scale score points (about 0.05 SDs) higher overall (column 2). We also find statistically significant increases in the probability of four-year college attendance (0.6 percentage points, column 3) and receipt of a bachelor's degree by age 25 (0.37 percentage

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

TABLE 2.—IMPACT OF ACCOUNTABILITY PRESSURE: ONLY POLICY VARIATION IN THE PREDICTION MODEL

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

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

points, column 4). The impact on earnings is positive but not statistically significant. In contrast, we find no significant impacts of accountability pressure to achieve a Recognized rating.

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

We find that all of the gains from accountability pressure to avoid a Low-Performing rating are concentrated among students who previously failed an exam. These students are about 4.7 percentage points more likely to pass the math exam (column 1), and they score about 1.3 scale score points (0.2 SDs) higher on the exam overall. More important, they are significantly more likely to attend a four-year college (1.9 percentage points, column 3) and earn a bachelor's degree (1.27 percentage points, column 4). These impacts, while small in absolute terms, represent about 19% and 30% of the mean for students who previously failed an eighth-grade exam. We also find that they earn about $298 more at age 25, and that impact is statistically significant at the 5% level.

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

C. Regression Results Using All Cohort Variation

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

To make use of cohort variation in prior characteristics, we estimate equation (1) again, but with separate predictions for each school and cohort. This allows for much more flexibility in schools' perceptions of accountability pressure over time; for example, a school may be at risk initially because of a particular subgroup, then switch to safe because the group becomes too small in subsequent cohorts.[10]

[10] We follow the minimum size requirements outlined by accountability policy and exclude subgroups that are less than 10% of the ninth-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% within five years. Appendix table A4 presents a transition matrix that shows the relationship between schools' predicted ratings in year T and year $T + 1$.

TABLE 3.—IMPACT OF ACCOUNTABILITY PRESSURE: ALL VARIATION IN THE PREDICTION MODEL

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

Each column is a single regression of the indicated outcome on the set of variables from equation (3), which includes controls for math and reading scores, demographics, and year and school fixed effects. Standard errors are block bootstrapped at the school level. Each coefficient gives the impact of being in a grade cohort that has a positive estimated risk of being rated Low-Performing or Recognized. (See the text for details on the construction of the ratings prediction.) A 1 standard deviation change in the math score is equal to about 7 scale score points. College attendance outcomes are measured within an eight-year time window beginning with the student's first-time ninth 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 eleventh year after the first time a student enters ninth grade (which we refer to as the age 25 year), including students with 0 reported earnings. Significant at **5%, ***1% or less.

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

Some schools would be at risk of being rated Low-Performing or Recognized because of a particular subgroup, but are actually safe because that subgroup is too small to count toward the rating. Thus the minimum subgroup size requirement provides us with a useful placebo test. In appendix table A5, we show that estimated impacts for placebo subgroups are near 0 and statistically significantly smaller than subgroups that are truly at risk.[11]

D. Robustness Checks

One potential concern is that the relationship between tenth-grade scores and eighth-grade characteristics is contaminated by endogenous responses to perceived risk. Concretely, if the prediction model in equation (1) used an identical set of covariates as equation (2), our estimates would be identified purely from functional form. However, the timing of perceived risk is a function of policy variation that is not in the prediction model. As a check on the endogeneity of the prediction model, in appendix table A6 we simply allow impacts to vary by the eighth-grade pass rate of the lowest-scoring subgroup in a school rather than estimating risk directly.[12]

Another concern is that the timing of a school's predicted rating is correlated with other contemporaneous shocks that might also affect long-run outcomes. We test for the possibility of contemporaneous shocks in appendix table A6 by regressing a school's predicted risk of being rated Low-Performing on time-varying high school inputs such as principal turnover, teacher pay, and teacher experience.[13]

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

We also measure possible attrition directly by constructing an indicator variable that is equal to 1 if a student has 0

[11] We select 8% as the placebo because schools face some uncertainty around the threshold, which is based on tenth-grade cohorts rather than first-time ninth graders.

[12] The results in table A6 are obtained by calculating the share of students in an incoming high school cohort who passed the eighth-grade exam for all test-subgroup combinations (e.g., Latinos in reading, blacks in math) We then take the difference between the minimum eighth-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 tenth grade, and divide schools into bins based on their relative distance from the yearly threshold. In this approach, there is no mean reversion or correlated estimation error because we do not estimate anything.

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

earnings and never attends any college between the ages of 19 and 25. This provides an upper bound on students who left the state and did not return (incarcerated or deceased students would have a value of 0, for example). In table 1, we see that the mean of this variable is 13% for the full sample.[14] When we estimate the impact of accountability pressure on this indicator for possible attrition, the estimate is −0.001 with a standard error of 0.002 for Low-Performing and 0.004 (0.003) for Recognized. Thus, there is no evidence of differential attrition, and our standard errors allow us to rule out all but very small impacts.

Our empirical strategy sometimes compares students who are only one or two grades apart in the same school. If accountability pressure causes schools to shift resources toward some students at the expense of others (Reback, 2008), comparisons across cohorts may be problematic. In appendix tables A11 and A12, we therefore restrict our analysis to nonconsecutive cohorts (i.e., 1995, 1997, and 1999) and nonoverlapping 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.

VII. What Explains the Pattern of Results?

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

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

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

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

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

In table 4 we provide some evidence on possible mechanisms by estimating results for additional outcomes in high school. The outcome in column 1 is an indicator for whether a student is receiving special education services in the tenth-grade year, but did not receive special education services in eighth grade. Panel B of column 1 shows strong evidence of strategic special education classification in schools that had a chance to achieve a Recognized rating. Low-scoring students in these schools are 2.4 percentage points more likely to be newly designated as eligible for special education, an increase of over 100% relative to the baseline mean of 2%. We also find a smaller (0.5 percentage points) but still highly significant decrease in special education classification for high-scoring students in these schools.

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

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

TABLE 4.—IMPACT OF ACCOUNTABILITY PRESSURE ON HIGH SCHOOL OUTCOMES

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

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

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

Increased knowledge of mathematics is a plausible mechanism for long-run impacts on postsecondary attainment and earnings. Using cross-state variation in the timing of high school graduation requirements, Levine and Zimmerman (1995) and Rose and Betts (2004) also find that additional mathematics course work in high school is associated with increases in labor market earnings. Cortes, Goodman, and Nomi (2015) find increases in high school graduation and college attendance for students who are assigned to a "double-dose" algebra I class in ninth grade.

In appendix table A14, we show that controlling for math course work reduces the estimates of accountability pressure on bachelor's degree receipt and earnings at age 25 to nearly 0, and lowers the impact on four-year college attendance by about 50%. This suggests that increases in math course work are a key mediator for explaining the long-run impacts of accountability pressure. In appendix table A15, which contains results for a number of additional high school outcomes, we show that these increases in math credits extend beyond the requirements of the tenth-grade math exit exam to upper-level course work such as algebra II and precalculus.

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

VIII. Discussion and Conclusion

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

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

We find that accountability pressure to avoid a Low-Performing rating leads to increases in labor market earnings at age 25 of around 1%. By comparison, Chetty et al. (2014) find that having a teacher in grades 3 through 8 with 1 standard deviation higher "value-added" also increases earnings at age 25 by about 1%. 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% 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 have 2% higher earnings at ages 26 to 29. Similarly, Chetty et al. (2011) find cross-sectional relationships between test scores at age 5 to 7 and adult earnings that are similar in size to our results for high school students.

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

REFERENCES

Allen, R., and S. Burgess, *How Should We Treat Under-Performing Schools? A Regression Discontinuity Analysis of School Inspections in England* (London: University of London, 2012).

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

Booher-Jennings, J., "Below the Bubble: "'Educational Triage' and the Texas Accountability System," *American Educational Research Journal* 42 (2005), 231–268.

Booker, K., T. R. Sass, B. Gill, and R. Zimmer, "The Effects of Charter High Schools on Educational Attainment," *Journal of Labor Economics* 29 (2011), 377–415.

Campbell, D. T., *Assessing the Impact of Planned Social Change* (Hanover, NH: Dartmouth College, Public Affairs Center, 1976).

Carnoy, M., and S. Loeb, "Does External Accountability Affect Student Outcomes? A Cross-State Analysis," *Educational Evaluation and Policy Analysis* 24 (2002), 305–331.

Carnoy, M., S. Loeb, and T. L. Smith, "Do Higher State Test Scores in Texas Make for Better High School Outcomes?" unpublished paper (2001).

Chetty, R., J. N. Friedman, N. Hilger, E. Saez, D. W. Schanzenbach, and D. Yagan, "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project STAR," *Quarterly Journal of Economics* 126 (2011), 1593–1660.

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

Chiang, H., "How Accountability Pressure on Failing Schools Affects Student Achievement," *Journal of Public Economics* 93 (2009), 1045–1057.

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

Craig, S. G., S. A. Imberman, and A. Perdue, "Does It Pay to Get an A? School Resource Allocations in Response to Accountability Ratings," *Journal of Urban Economics* 73 (2013), 30–42.

Cullen, J. B., and R. Reback, "Tinkering toward Accolades: School Gaming under a Performance Accountability System," in T. Gronberg and D. Jansen, eds., *Advances in Applied Microeconomics* (Amsterdam: Elsevier, 2006).

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

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

——— "Better Schools, Less Crime?" *Quarterly Journal of Economics* 126 (2011), 2063–2115.

Deming, D., J. S. Hastings, T. J. Kane, and D. O. Staiger, "School Choice, School Quality and Academic Achievement," *American Economic Review* 104 (2014), 991–1013.

Diamond, J. B., "Where the Rubber Meets the Road: Rethinking the Connection between High-Stakes Testing Policy and Classroom Instruction," *Sociology of Education* 80 (2007), 285–313.

Diamond, J., and J. Spillane, "High-Stakes Accountability in Urban Elementary Schools: Challenging or Reproducing Inequality?" *Teachers College Record* 106 (2004), 1145–1176.

Donovan, C., D. N. Figlio, and M. Rush, *Cramming: The Effects of School Accountability on College-Bound Students* (Cambridge, MA: National Bureau of Economic Research, 2006).

Evers, W. M., and H. J. Walberg, *School Accountability* (Stanford, CA: Hoover Press, 2002).

Figlio, D. N., and L. S. Getzler, "Accountability, Ability and Disability: Gaming the System?" (pp. 35–49), in T. Gronberg and D. Jansen, eds., *Advances in Applied Microeconomics* (Amsterdam: Elsevier, 2006).

Figlio, D. N., and L. W. Kenny, "Public Sector Performance Measurement and Stakeholder Support," *Journal of Public Economics* 93 (2009), 1069–1077.

Figlio, D. N., and H. F. Ladd, "School Accountability and Student Achievement" (pp. 166–182), in Helen Ladd and Edward B. Fiske, eds., *Handbook of Research in Education Finance and Policy* (New York: Routledge, 2008).

Figlio, D., and S. Loeb, "School Accountability" (pp. 383–421), in Eric A. Hanushek, Stephen, J. Machin, and Ludger Woessman, eds., *Handbook of the Economics of Education* (Amsterdam: North-Holland, 2011).

Figlio, D. N., and M. E. Lucas, "What's in a Grade? School Report Cards and the Housing Market," *American Economic Review* 94 (2004), 591–604.

Figlio, D. N., and C. E. Rouse, "Do Accountability and Voucher Threats Improve Low-Performing Schools?" *Journal of Public Economics* 90 (2006), 239–255.

Figlio D. N., and J. Winicki, "Food for Thought: The Effects of School Accountability Plans on School Nutrition," *Journal of Public Economics* 89 (2005), 381–394.

Gould, E. D., V. Lavy, and M. D. Paserman, "Immigrating to Opportunity: Estimating the Effect of School Quality Using a Natural

Experiment on Ethiopians in Israel," *Quarterly Journal of Economics* 119 (2004), 489–526.
Greene, J., M. Winters, and G. Forster, "Testing High-Stakes Tests: Can We Believe the Results of Accountability Tests?" *Teachers College Record* 106 (2004), 1124–1144.
Hamilton, L. S., B. M. Stecher, J. A. Marsh, J. S. McCombs, and A. Robyn, *Standards-Based Accountability under No Child Left Behind: Experiences of Teachers and Administrators in Three States* (Santa Monica, CA: RAND Corporation, 2007).
Haney, W., "The Myth of the Texas Miracle in Education," *Education Policy Analysis Archives* 8 (2000).
Hanushek, E. A., and M. E. Raymond, "Does School Accountability Lead to Improved Student Performance?" *Journal of Policy Analysis and Management* 24 (2005), 297–327.
Holmstrom, B., and Paul Milgrom, "Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design," *Journal of Law, Economics and Organization* 7 (1991), 24–52.
Hout, M., and S. W. Elliott, *Incentives and Test-Based Accountability in Education* (Washington, DC: National Academies Press, 2011).
Izumi, L. T., and W. M. Evers, "State Accountability Systems," in W. M. Evers and H. J. Wallbert, eds., *School Accountability: An Assessment by the Koret Task Force on K-12 Education* (Stanford, CA: Hoover Institution Press, 2002).
Jackson, C. K., "Non-Cognitive Ability, Test Scores, and Teacher Quality: Evidence from 9th Grade Teachers in North Carolina," NBER working paper 18624 (2012).
Jacob, B. A., "Accountability, Incentives and Behavior: The Impact of High-Stakes Testing in the Chicago Public Schools," *Journal of Public Economics* 89 (2005), 761–796.
Jacob, B. A., and S. D. Levitt, "Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating," *Quarterly Journal of Economics* 118 (2003), 843–877.
Jennings, J. L., and A. A. Beveridge, "How Does Test Exemption Affect Schools' and Students' Academic Performance?" *Educational Evaluation and Policy Analysis* 31 (2009), 153–175.
Kerr, S., "On the Folly of Rewarding A, While Hoping for B," *Academy of Management Journal* 18 (1975), 769–783.
Klein, S. P., L. Hamilton, D. F. McCaffrey, and B. Stecher, *What Do Test Scores in Texas Tell Us?* (Santa Monica, CA: Rand, 2000).
Koretz, D. M., and S. I. Barron, *The Validity of Gains in Scores on the Kentucky Instructional Results Information System (KIRIS)* (Santa Monica, CA: RAND, 1998).
Ladd, H. F., "The Dallas School Accountability and Incentive Program: An Evaluation of Its Impacts on Student Outcomes," *Economics of Education Review* 18 (1999), 1–16.
Ladd, H. F., and D. L. Lauen. "Status versus Growth: The Distributional Effects of School Accountability Policies." *Journal of Policy Analysis and Management* 29 (2010), 426–450.
Lemons R., T. Luschel, and L. Siskin, "Leadership and the Demands for Standards-Based Accountability" (pp. 99–128), in M. Carnoy, R. Elmore, and L. S. Siskin, eds., *The New Accountability: High Schools and High-Stakes Testing* (New York: Routledge-Falmer, 2003).
Levine, P. B., and D. J. Zimmerman, "The Benefit of Additional High-School Math and Science Classes for Young Men and Women," *Journal of Business and Economic Statistics* 13 (1995), 137–149.
Levy, F., and R. J. Murnane, *The New Division of Labor: How Computers Are Creating the Next Job Market* (Princeton, NJ: Princeton University Press, 2012).
McNeil, L. M., E. Coppola, J. Radigan, and J. Vasquez Heilig, "Avoidable Losses: High-Stakes Accountability and the Dropout Crisis," *Education Policy Analysis Archives* 16:3 (2008), 1.
McNeil L., and A. Valenzuela, "The Harmful Impact of the TAAS System of Testing in Texas: Beneath the Accountability Rhetoric" (pp. 127–150), in Gary Orfield and Minday Kornhaber, eds., *Raising Standards or Raising Barriers? Inequality and High Stakes Testing in Public Education* (New York: Century Foundation, 2001).
Mintrop, H., and T. Trujillo, "Corrective Action in Low Performing Schools: Lessons for NCLB Implementation from First-Generation Accountability Systems," *Education Policy Analysis Archives* 13 (2005), 48.
Murphy, K. M., and R. H. Topel, "Estimation and Inference in Two-Step Econometric Models," *Journal of Business and Economic Statistics* 20 (2002), 88–97.
Neal, D. A., "*The Consequences of Using One Assessment System to Pursue Two Objectives*," NBER working paper 19214 (2013).
Neal, D. A., and W. R. Johnson, "The Role of Premarket Factors in Black-White Wage Differences," *Journal of Political Economy* 104 (1996), 869–895.
Neal, D., and D. W. Schanzenbach, "Left Behind by Design: Proficiency Counts and Test-Based Accountability," this REVIEW 92 (2010), 263–283.
Reback, R., "Teaching to the Rating: School Accountability and the Distribution of Student Achievement," *Journal of Public Economics* 92 (2008), 1394–1415.
Reback, R., J. Rockoff, and H. L. Schwartz, "Under Pressure: Job Security, Resource Allocation, and Productivity in Schools under NCLB," *American Economic Journal: Economic Policy* 6 (2014), 207–241.
Rockoff, J., and L. J. Turner, "Short-Run Impacts of Accountability on School Quality," *American Economic Journal: Economic Policy* 2 (2010), 119–147.
Rose, H., and J. R. Betts, "The Effect of High School Courses on Earnings," this REVIEW 86 (2004), 497–513.
Rouse, C. E., J. Hannaway, D. Goldhaber, and D. Figlio, "Feeling the Florida Heat? How Low-Performing Schools Respond to Voucher and Accountability Pressure," *American Economic Journal: Economic Policy* 5 (2013), 251–281.
Skrla, L., J. J. Scheurich, and J. F. Johnson, *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, 2000).
Spillane, J. P., L. M. Parise, and J. Z. Sherer, "Organizational Routines as Coupling Mechanisms Policy, School Administration, and the Technical Core," *American Educational Research Journal* 48 (2011), 586–619.
Stecher B. M., S. L. Barron, T. Chun, and K. Ross, *The Effects of the Washington State Education Reform on Schools and Classrooms* (Santa Monica, CA: RAND Corporation, 2000).
Stotsky, S., "Analysis of the Texas Reading Tests, Grades 4, 8 and 10, 1995–1998," Independent report prepared for the Tax Research Foundation, Houston, Texas (November 1999).
Toenjes, L. A., and J. E. Garst, *Identifying High Performing Texas Schools and School Districts and Their Methods of Success* (Austin: Texas Education Agency, 2000).
Vasquez Heilig, J. V., and L. Darling-Hammond, "Accountability Texas-Style: The Progress and Learning of Urban Minority Students in a High-Stakes Testing Context," *Educational Evaluation and Policy Analysis* 30 (2008), 75–110.
Wong K., Looking Beyond Test Score Gains: State Accountability's Effect on Educational Attainment and Labor Market Outcomes (Irvine: University of California, Irvine, 2008).