WHEN DOES ACCOUNTABILITY WORK?

Texas system had mixed effects on college graduation rates and future earnings

WHEN CONGRESS PASSED the No Child Left Behind Act of 2001 (NCLB), standardized testing in public schools became the law of the land. The ambitious legislation identified test-based accountability as the key to improving schools and, by extension, the long-term prospects of American schoolchildren. Thirteen years later, the debate over the federal mandate still simmers. According to the 2015 *EdNext* poll, about two-thirds of K–12 parents support annual testing requirements, yet a vocal minority want the ability to have their children "opt out" of such tests (see "The 2015 *EdNext* Poll on School Reform," *features,* Winter 2016). Teachers themselves are divided on the issue of high-stakes testing.

NCLB required that states test students in math and reading each year, that average student performance be publicized for every school, and that schools with persistently low test scores face an escalating series of sanctions. We now have ample evidence that these requirements have caused test scores to rise across the country. What we don't know is: Do these improvements on high-stakes tests represent real learning gains? And do they make students better off in the long run? In fact, we know very little about the impact of test-based accountability on students' later success. If academic gains do not translate into a better future, why keep testing?

In this study, we present the first evidence of how accountability pressure on schools influences students' long-term outcomes. We do so by examining how the test-based accountability system introduced in Texas in 1993 affected students' college enrollment and completion rates and their earnings as adults. Though the Texas system predates NCLB, it was implemented under then governor George W. Bush and it served as a blueprint for the federal legislation he

by DAVID J. DEMING, SARAH COHODES, JENNIFER JENNINGS, AND CHRISTOPHER JENCKS

signed as president nearly a decade later. More important, it was implemented long enough ago to allow us to investigate its impact on adult outcomes, since individuals who were in high school in the mid- to late 1990s have now reached adulthood.

Our analysis reveals that pressure on schools to avoid a low performance rating led low-scoring students to score significantly higher on a high-stakes math exam in 10th grade. These students were also more likely to accumulate significantly more math credits and to graduate from high school on time. Later in life, they were more likely to attend and graduate from a four-year college, and they had higher earnings at age 25.

Those positive outcomes are not observed, however, among students in schools facing a different kind of accountability pressure. Higher-performing schools facing pressure to achieve favorable recognition appear to have responded primarily by finding ways to exempt their low-scoring students from counting toward the school's results. Years later, these students were less likely to have completed college and they earned less.

In short, our results indicate that school accountability in Texas led to long-term gains for students who attended schools that were at risk of falling below a minimum performance standard. Efforts to use high-stakes tests to regulate school quality at a higher level, however, did not benefit students and may have led schools to adopt strategies that caused long-term harm.

The Texas school accountability system implemented under then Governor George W. Bush served as a blueprint for the federal legislation he signed as president nearly a decade later.

PHOTOGRAPH/CORBIS; JEFF MITCHELL

The Accountability Movement

A handful of states, such as Texas and North Carolina, began implementing "consequential" school accountability policies in the early 1990s. Under these policies, performance on standardized tests was not only made public but was also tied to rewards and sanctions. The number of states with consequential school-accountability policies rose from 5 in 1994 to 36 in 2000.

Under the accountability system implemented by Texas in 1993, every public school was given one of four ratings: Low-Performing, Acceptable, Recognized, or Exemplary. Schools were rated based on the overall share of students who passed the Texas Assessment of Academic Skills tests in reading, writing, and mathematics; attendance and high-school dropout rates were also considered. Pass rates were calculated separately for four subgroups—white, African American, Hispanic, and economically disadvantaged—if such subgroup made up at least 10 percent of the school's population. Schools were assigned an overall rating based on the pass rate of the lowest-scoring subgroup-test combination (e.g., math for whites), giving some schools strong incentives to focus on particular students and subjects. (Because the state's math test was more difficult than its reading test, low math scores were almost always the main obstacle to improving a school's rating.) School ratings were often published in full-page spreads in local newspapers, and schools that were rated as Low-Performing underwent an evaluation that could lead to serious consequences, including layoffs, reconstitution, and school closure.

The accountability system adopted by Texas bore many similarities to the accountability requirements of NCLB, enacted nine years later. NCLB mandated reading and math testing in grades 3 through 8 and at least once in high school, and it required states to rate schools on the basis of test performance overall and for key subgroups. It also called for sanctions on schools that failed to meet statewide targets for student proficiency rates. Finally, the system required states to report subgroup test results and to increase their proficiency rate targets over time.

Too Good to Be True?

Scores on high-stakes tests rose rapidly in states that were early adopters of school accountability, and Texas was no exception. Pass rates on the state's 10th-grade exam, which was also a high-stakes exit exam for students, rose from 57 percent to 78 percent between 1994 and 2000, with smaller yet still sizable gains in reading (see Figure 1).

The interpretation of this so-called Texas miracle, however, is complicated by studies of schools' strategic responses to high-stakes testing. Research on how high-stakes accountability affects test performance has found that scores on high-stakes tests tend to improve with accountability, often dramatically, whereas performance on low-stakes tests with a different format but similar content improves only slightly or not at all. Furthermore, studies in Texas and elsewhere have found that some schools raised their published test scores by retaining low-performing students in 9th grade, by classifying them as

School accountability in Texas led to long-term gains for students who attended schools at risk of falling below the minimum performance standard.

eligible for special education (or otherwise exempting them from the exam), and even by encouraging them to drop out.

Clearly, accountability systems that rely on short-term, quantifiable measures to drive improved performance can lead to unintended consequences. Performance incentives may cause schools and teachers to redirect their efforts toward the least costly ways of raising test scores, at the expense of actions that do not boost scores but may be important for students' long-term welfare.

Our study overcomes the limits of short-term analysis by asking: when schools face accountability pressure, do their efforts to raise test scores generate improvements in higher education attainment, earnings, and other long-term outcomes?

Our Study

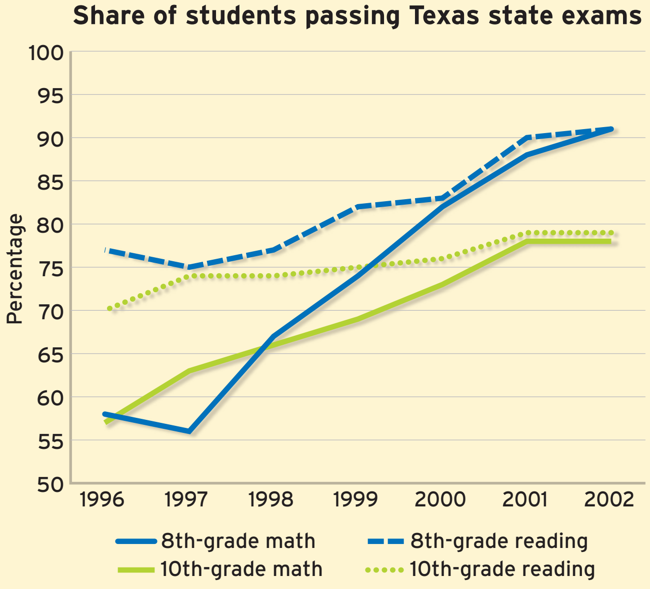
An ideal experiment to address this question would randomly assign schools to test-based accountability and then observe changes in both test scores and long-term outcomes, comparing the results to those of a control group of schools. Such an experiment is not possible in this case because of the rapid rollout of high-stakes testing in Texas and (later) nationwide. And unfortunately, data limitations preclude us from looking at prior cohorts of students who were not part of the high-stakes testing regime.

Instead, our research design compares successive grade cohorts within the same school—cohorts that faced different degrees of accountability pressure owing to changes in how the state defined school performance categories over time. Beginning in 1995, each Texas school received its overall rating based on its lowest subgroup-test pass rate. That year, at least 25 percent of all tested students in a high school were required to pass the 10th-grade exit exam in each subject in order for the school to receive an Acceptable rating. This standard rose by 5 percentage points every year, up to 50 percent in 2000. The standard for a Recognized rating also rose, from a 70 percent pass rate 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. We use these changes in performance standards to estimate the "risk" that each school will receive a particular rating, and we compare cohorts who attended a school when it was on the brink of receiving a Low-Performing or Recognized rating to cohorts in the same school in years that it was all but certain to be rated Acceptable—and therefore plausibly "safe" from accountability pressure.

Most research on school accountability has studied how schools respond to receiving a poor rating, but our approach focuses instead on the much larger group of schools that face pressure to avoid a Low-Performing rating in the first place. Because the ratings thresholds rose over time, the set of schools experiencing the most pressure also changed. Consider, for example, students in a school that was plausibly safe from

The "Texas Miracle"? (Figure 1)

The percentage of students passing Texas state assessments increased rapidly after the introduction of the state accountability system in 1993.

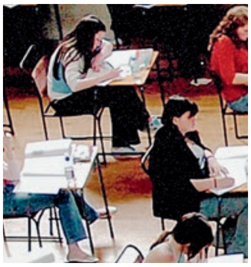


NOTE: Students are assigned to cohorts based on the first time they enter 9th grade.

SOURCE: Authors' calculations based on data from the Texas Assessment of Academic Skills accountability pressure in 1995 but was at risk of a Low-Performing rating in 1996. Students in the 1996 cohort are likely quite similar to students in the class before them, except for the fact that they were subject to greater accountability pressure. (Our analysis does include controls for various ways in which those cohorts may have differed initially, such as by incoming test scores and demographic makeup.) By comparing grade cohorts who faced different degrees of accountability pressure, we can ascertain how much their level of risk affects not only 10th-grade exam scores but also how much schooling they completed and their earnings later in life.

Findings

We find that students, on average, experience better outcomes when they are in a grade cohort that puts its school at risk of receiving a Low-Performing rating. They score higher on the 10th-grade math exam, are more likely to graduate from high school on time, and accumulate more math credits, including in subjects beyond a 10th-grade level.



Students in a grade cohort that puts its school at risk of receiving a Low-Performing rating fare better than students whose schools are not facing as much accountability pressure. On average, they earn about 1 percent more at age 25.

Later in life, these students are 0.6 percentage points more likely to attend a four-year college and 0.37 percentage points more likely to graduate. They also earn about 1 percent more at age 25 than those who were in cohorts whose schools were not facing as much accountability pressure. The earnings increase is comparable to the impact of having a teacher at the 87th percentile, in terms of her "value added" to student achievement, versus a teacher at the value-added median (see "Great Teaching," *research,* Summer 2012).

Since the Texas state test was a test of basic skills, and the accountability metric is based on pass rates, schools had strong incentives to focus on helping lower-scoring students. While schools surely varied in how they identified struggling students, one reliable predictor that students might fail the 10th-grade exam was whether they failed an 8th-grade exam.

In fact, when we take into account 8th-grade failure rates, we find that *all* of the aforementioned gains are concentrated among students who previously failed an exam. These students are about 4.7 percentage points more likely to pass the 10th-grade math exam, and they score about 0.2 standard deviations higher on the exam overall (see Figure 2). More importantly, they are significantly more likely to attend a four-year college (1.9 percentage points) and earn a bachelor's degree (1.3 percentage points). These impacts, while small in absolute terms, represent about 19 and 30 percent of the mean for students who previously failed an 8th-grade exam. We also find that they earn about $300 more annually at age 25.

In contrast, we find *negative* long-term impacts for low-scoring students in grade cohorts attending a school in a year when it faced pressure to achieve a Recognized rating. Students from these cohorts who previously failed an exam are about 1.8 percentage points less likely to attend a four-year college and 0.7 percentage points less likely to earn a bachelor's degree, and they earn an average of $748 less at age 25. This negative impact on earnings is larger, in absolute terms, than the positive earnings impact in schools at risk of being rated Low-Performing. However, there are fewer low-scoring students in high-scoring schools, so the overall effects on low-scoring students roughly cancel one another other out. Again we find no impact of accountability pressure on higher-achieving students.

What worked well. Higher test scores in high school do not necessarily translate into greater postsecondary attainment and increased earnings in adulthood, yet our study demonstrates that, for many students, accountability pressure does seem to positively influence these long-range outcomes. Additional knowledge of mathematics is one plausible explanation for these favorable impacts on postsecondary attainment and earnings. Accountability pressure could have caused students to learn more math through: 1) additional class time and resources devoted to math instruction and 2) changes in students' later course-taking patterns, sparked by improved on-time passage of the exit exam.

Indeed, we find an average increase of about 0.06 math course credits per student in schools that face pressure to avoid a Low-Performing rating. We also find that the impacts on both math credits and long-range outcomes grow with cohort size and with the number of students who previously failed an 8th-grade exam, suggesting that students particularly benefited from accountability pressure when it prompted schoolwide reform efforts.

Prior research has demonstrated that additional mathematics coursework in high school is associated with higher earnings later in life, and that even one additional year of math coursework increases annual earnings by between 4 and 8 percentage points. In our study, controlling for the amount of math coursework reduces the effects of accountability pressure on bachelor's degree receipt and earnings at age 25 to nearly zero, and lowers the impact on four-year college attendance by about 50 percent. This suggests that additional math coursework may be a key mechanism for the long-term impacts of accountability pressure under the Texas policy.

Additionally, we find some evidence that schools respond to the risk of being rated Low-Performing by adding staff, particularly in remedial classrooms. This response is consistent with studies of accountability pressure in Texas and elsewhere that find increases in instructional time and resources devoted to low-scoring students, and provides another possible explanation for the positive effects of accountability pressure for certain students.

Dangers of a poorly designed system. Despite finding evidence of significant improvements in long-range outcomes for some students, those same improvements were not enjoyed by others. Why might an accountability system generate seemingly contradictory results?

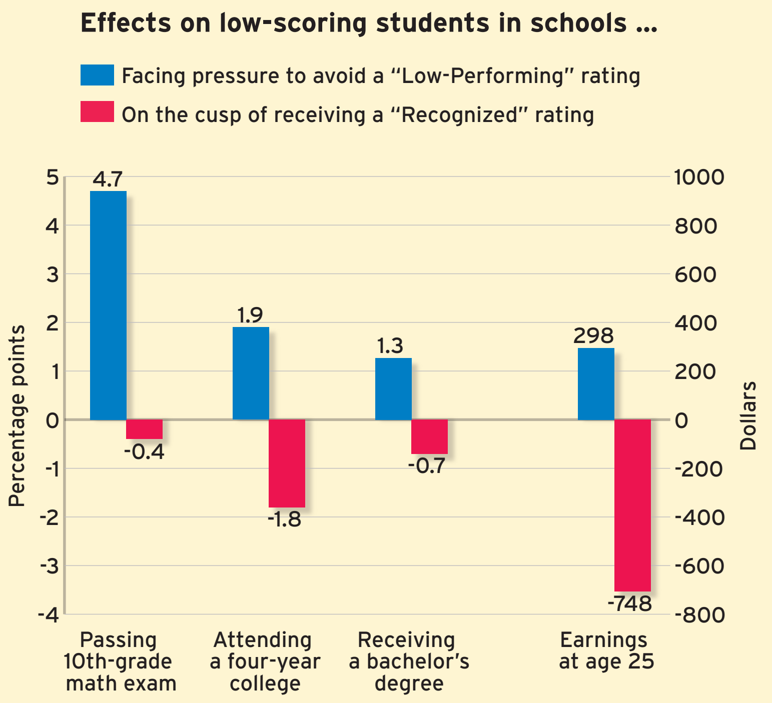
As mentioned earlier, high-stakes testing poses the risk that it may cause teachers and schools to adjust their effort toward the least costly (in terms of dollars or effort) way of boosting test scores, possibly at the expense of other constructive actions. Thus, one can try to understand the difference in impacts between the two kinds of accountability by asking: in each situation, what was the least costly method of achieving a higher rating?

In our data, the student populations of schools at risk of a Low-Performing rating were, on average, 23 percent African American and 32 percent Hispanic, and 44 percent of students were poor. The mean cohort size was 212, and the mean pass rate on the 8th-grade math exam was 56 percent. Since the overall cohort and each tested subgroup were on average quite large, these schools could only escape a Low-Performing rating through broad improvement in test performance. In contrast, school populations closer to the high end of the performance spectrum were only about 5 percent African American, 10 percent Hispanic, and 16 percent poor, with a mean cohort size of only 114 and a mean pass rate of 84 percent on the 8th-grade math exam. Thus, many of the schools that were aspiring to a Recognized rating could achieve it by affecting the scores of only a small number of students.

One example of how a small school might "game the system" is by strategically classifying students in order to influence who "counts" toward the school's rating. Indeed, we find strong evidence that some schools trying to attain a Recognized rating did so by exempting students from the high-stakes test. These schools classified low-performing students as eligible for special education services to keep them from lowering the school's

Long-Term Gains and Losses from Accountability Pressure (Figure 2)

Low-scoring students in Texas schools at risk of receiving a "Low-Performing" rating were more likely to attend and graduate from college and earned more at age 25. However, pressure on higher-performing schools to achieve a "Recognized" rating led to negative long-term outcomes.



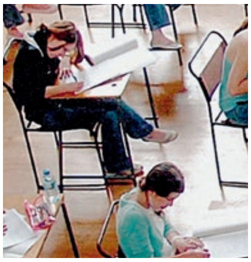
NOTE: Under the high-stakes accountability system implemented in Texas in 1993, Texas schools were grouped into one of four possible performance categories: Low-Performing, Acceptable, Recognized and Exemplary.

Source: Authors' calculations

rating (special education students could take the 10th-grade state test, but their scores did not count toward the rating).

In schools that had a chance to achieve a Recognized rating, low-scoring students who were not designated as eligible for special education in 8th grade were 2.4 percentage points more likely to be newly designated as such in 10th grade, an increase of more than 100 percent relative to the 2 percent designation rate in other schools. The designation of low-scoring students as eligible for special education was more common in schools where a small number of students had failed the 8th-grade exam, making it easier for educators to target specific students. We also find a small but still noteworthy *decrease* of 0.5 percentage points in special education classification for high-scoring students in these schools.

As a result of this strategic classification, marginal students in certain schools were placed in less-demanding courses and acquired fewer skills, accounting for the negative impact of accountability pressure on long-term outcomes for those students. In essence, those students did not receive the attention they needed in order to improve their learning.



Some schools trying to attain a Recognized rating did so by exempting students from the high-stakes test, classifying low-performing students as eligible for special education services to keep them from lowering the school's rating.

Summing Up

Why do some students benefit from accountability pressure while others suffer? Our results suggest that Texas schools responded to accountability pressure by choosing the path of least resistance, which produced divergent outcomes. The typical school at risk of receiving a Low-Performing rating was large and had a majority nonwhite population, with many students who had previously failed an 8th-grade exam. These schools had limited opportunity to strategically classify students as eligible for special education services. Instead, they had to focus their efforts on truly helping a large number of students improve. As a result, students in these schools were more likely to pass the 10th-grade math exam on time, acquire more math credits in high school, and 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.

In other schools, the accountability system produced strong incentives to exempt students from exams and other requirements. In these schools, accountability pressure more than doubled the chances that a low-scoring student would be newly deemed eligible for special education. This designation exempted students from the normal high-school graduation requirements, which then led them to accumulate fewer math credits. In the long run, low-scoring students in these schools had significantly lower postsecondary attainment and earnings.

In some respects, though not all, the accountability policy in Texas served as the template for No Child Left Behind, and thus our findings may have applicability to the accountability regimes that were rolled out later in other states. In Texas, and under NCLB nationwide, holding schools accountable for the performance of every student subgroup has proven to be a mixed blessing. On the one hand, this approach shines light on inequality within schools in an attempt to ensure that "no child is left behind." On the other hand, when schools can achieve substantial "improvements" by focusing on a relatively small group of students, they face a strong incentive to game the system. In Texas, this situation led some schools to strategically classify students as eligible for special education, which may have done them long-run harm.

What policy lessons can we draw from this study as Congress works out a new iteration of the Elementary and Secondary Education Act to replace NCLB? First, policy complexity can carry a heavy cost. As many other studies have shown, high-stakes testing creates strong incentives to game the system, and the potential for strategic responses grows as the rules become more complicated. The second lesson is that, at least in Texas, school accountability measures only worked for schools that were at risk of receiving a failing grade. Therefore, the federal government might consider approaching school accountability the way the Food and Drug Administration regulates consumer products. Instead of rating and ranking schools, the feds could develop a system that ensures a minimum standard of quality.

David J. Deming is associate professor of education and economics at the Harvard Graduate School of Education. Sarah Cohodes is assistant professor of education and public policy at Teachers College, Columbia University. Jennifer Jennings is assistant professor of sociology at New York University. Christopher Jencks is the Malcolm Wiener Professor of Social Policy at the Harvard Kennedy School.