Accountability Texas-Style: The Progress and Learning of Urban Minority Students in a High-Stakes Testing Context

Julian Vasquez Heilig and Linda Darling-Hammond

EDUCATIONAL EVALUATION AND POLICY ANALYSIS 2008; 30; 75
DOI: 10.3102/0162373708317689

Educational Evaluation and Policy Analysis
June 2008, Vol. 30, No. 2, pp. 75–110
DOI: 10.3102/0162373708317689

Accountability Texas-Style: The Progress and Learning of Urban Minority Students in a High-Stakes Testing Context

Julian Vasquez Heilig
University of Texas at Austin

Linda Darling-Hammond
Stanford University

This study examines longitudinal student progress and achievement on the elementary, middle, and high school levels in relation to accountability policy incentives in a large urban district in Texas. Using quantitative analyses supplemented by qualitative interviews, the authors found that high-stakes testing policies that rewarded and punished schools based on average student scores created incentives for schools to "game the system" by excluding students from testing and, ultimately, school. In the elementary grades, low-achieving students were disproportionately excluded from taking the high-stakes Texas Assessment of Academic Skills tests, demonstrating gains not reflected on the low-stakes Stanford Achievement Test–Ninth Edition. Student exclusion at the elementary level occurred through special education and language exemptions and missing scores. Furthermore, gaming strategies reduced educational opportunity for African American and Latino high school students. Sharp increases in 9th-grade student retention and disappearance were associated with increases in 10th-grade test scores and related accountability ratings.

Keywords: *accountability, testing, dropout, pushout, minorities, urban education*

HIGH-STAKES testing and accountability policies are expanding their reach in states and districts nationwide, stimulated in part by the 2002 passage of the No Child Left Behind Act. The prevailing theory of action behind accountability ratings and testing is that schools and students who are held accountable to these measures will automatically increase educational output: Educators will try harder; schools will adopt more effective methods; and students will learn more. Pressure to improve test scores will produce genuine gains in student achievement.

However, the effects of high-stakes testing policies have been debated. Do policies that reward and sanction schools and students based on test scores improve achievement and the quality of education for all or most students? Do they create incentives to "game" the system by teaching to the test or by removing low-achieving students from the testing pool through placement decisions and enrollment actions? Or do they have differential effects on different students in different school contexts? To answer questions like these, it is critical to look not only at aggregate

We wish to gratefully acknowledge the support of the Rockefeller Foundation in conducting this research. The findings and views expressed are ours alone.

trends in average test scores at the school, district, and state levels but at individual student progress through school over time.

In this study, we examine longitudinal student progress and achievement using data on more than 250,000 students over 7 years in a large urban district in Texas that we call Brazos City (pseudonym), where the accountability system adopted in the early 1990s provided the model for the No Child Left Behind Act a decade later. We examine district, school, and individual student trends in test scores on multiple tests, as well as student progress through school and graduation. Our goal is to empirically evaluate whether accountability policies and incentives are associated with changes in student achievement and whether they increase retention, dropout, and disappearance of students from school, with emphasis on outcomes for low-income students and students of color. We evaluate the relationships between these trends and school accountability ratings over time using multivariate statistical methods and analysis of interview data from more than 160 students and staff across seven high schools in the district. These analyses allow us to examine the possibility of gaming actions aimed at boosting school-level accountability ratings that may result in unintended consequences for some subpopulations of students, including efforts to exclude students from testing or from school.

Accountability Texas-Style

Texas was one of the earlier states to develop statewide testing systems during the 1980s, and the state adopted minimum competency tests for school graduation in 1987. In 1993, the Texas legislature mandated the creation of the Texas public school accountability system to rate school districts and evaluate campuses. The Texas accountability system was supported by the Public Education Information Management System (PEIMS) data collection system (a state-mandated curriculum) and the Texas Assessment of Academic Skills (TAAS). Between 1994, the baseline year, and 2002, when the TAAS was replaced by another test, the primary base indicators for determining school accountability ratings involved, first, the proportion of students passing all TAAS subject area tests (with the passing score represented by a 70 on the Texas Learning Index [TLI], a scaled score on the TAAS) and, second, the annual drop-out rate—each disaggregated by student groups: African American, Latino, White, and economically disadvantaged.

A key element of this system involved the use of the 10th-grade TAAS tests in reading, writing, and mathematics as a requirement for graduation from high school as well as the central indicator for establishing high school accountability rankings. Schools were categorized as *exemplary*, *recognized*, *acceptable*, and *low performing*. For a school to be designated *exemplary*, 90% of its students (and 90% of each student subgroup, defined by race/ethnicity and income) had to pass the tests, and the school drop-out rate could not exceed 1%. To be *recognized*, a school needed a 65% pass rate for all subgroups on the test and a drop-out rate of no more than 3.5%. Between 1995 and 1998, the required pass rate increased to 80%. To be *acceptable*, a school needed an initial pass rate of 25% on the test—increasing to 50% for each subgroup by 2000—and a drop-out rate of no more than 6%. In 2001, required drop-out rates were dropped to 3% for a recognized school and 5.5% for an acceptable school. In 2002, these rates were dropped further, to 2.5% and 5.0%, respectively.

High schools' reputations, funding, and their continued existence depended on students' performance on the exit TAAS. Graduation for students also depended on passing the 10th-grade TAAS in reading, writing, and mathematics. Thus, the TAAS was high stakes for students, educators, and schools.

In addition to the increasing expectations of schools, as reflected in these requirements, there was a set of changing rules for students' inclusion in the testing program. Initial exemption for special education students shifted over time, as did that for limited-English-proficient (LEP) students, who could be exempt for a period of time and tested on the Spanish TAAS. Scores for those special education students tested on the TAAS were considered in the accountability system beginning in 1999, and scores for elementary students tested on the Spanish-language TAAS were phased into the accountability system in 1999 and 2000. In 2000, LEP exemptions were limited, and student underreporting and special education compliance could affect a district's rating. By 2001, LEP exemptions were restricted further, and special education exemptions were evaluated. By 2002, a

State-Developed Alternative Assessment for special education students was developed to include them in the accountability system; a social studies test was added to the list; and pass rates were raised once again. After 2002, the tests changed to become more rigorous, and the system evolved. By all accounts, managing this system has been a major preoccupation for Texas schools and districts.

Many perceive Texas-style high-stakes testing and accountability as having become the driving education policy for the nation with its incorporation into the reauthorization of the Elementary and Secondary Education Act in 2002 as the No Child Left Behind Act (McNeil, 2005). The latter requires states and localities to build accountability systems based on assessments that become high stakes, because schools must meet annual test score targets for subgroups of students, thereby making adequate yearly progress, or face federal sanctions and penalties. Thus, studying Brazos City and Texas's first-generation accountability system provides an opportunity to evaluate one aspect of the theory of action underlying No Child Left Behind Act.

Prior Research on High-Stakes Testing

Evidence on the effects of high-stakes testing is mixed. Some studies suggest that students and schools make achievement gains in contexts where tests are used for decision making, whereas other research has found no improvement or even negative consequences. Among these unintended outcomes are school strategies to game the system, by adjusting the testing pool through student placements, admissions, and policies. Also of concern are the ways in which stakes attached to tests can corrupt what the tests measure, making outcomes nongeneralizable to other achievement measures and kinds of learning (for a summary of issues associated with test-based accountability systems, see, e.g., Hamilton, Stecher, & Klein, 2002).

Accountability Systems and Student Achievement

Several studies have used aggregated state-level data to examine whether state high-stakes testing policies appear to increase average student achievement levels. Carnoy and Loeb (2002) used a 5-point index of the strength of accountability systems in all 50 states—with higher ratings assigned to systems using high-stakes testing to reward or sanction schools—to examine whether accountability "strength" was related to student gains on the National Assessment of Educational Progress (NAEP) mathematics test in the 1996–2000 period. They found that students in states with stronger high-stakes accountability systems made significantly higher gains on the eighth-grade national mathematics assessment and that these gains were greater for African American and Latino students, thus narrowing the achievement gap. The study did not find evidence of a relationship between accountability systems and either higher rates of student retention or changes in high school completion rates.

Hanushek and Raymond (2003) also reported positive achievement effects in their analysis of aggregate state-level NAEP mathematics data. They examined the relationship between state-level accountability policies and achievement for cohorts at Grades 4 and 8 and found that accountability schemes appeared to increase state achievement gains. However, they also found that accountability policies did not close the gap in student learning but actually increased it, given that African Americans and Latinos showed lower gains on each test when compared to Whites. In contrast, Lee and Wong (2004) found no evidence that accountability policies resulted in test score gains or changes in the achievement gap, positive or negative.

Another way of examining the question of achievement effects is to evaluate whether gains on high-stakes tests appear to be related to gains in other measures of learning. Amrein and Berliner (2002) examined 18 states with severe consequences attached to their testing programs, to see if high-stakes testing affected student learning on measures other than the high-stakes tests. The authors posited that if student learning is actually increasing under high-stakes testing programs, then transfer learning would be evident on other standardized tests, such as the NAEP, advanced placement tests, and college admissions tests (e.g., ACT, SAT). They found that student learning effects were indeterminate: In most cases, when measured by tests other than the state-mandated high-stakes instruments,

student achievement appeared to remain at the same level it was before the policy was implemented, or it went down when high-stakes testing policies were instituted.

Rosenshine (2003) reanalyzed Amrein and Berliner's NAEP results (2002), arguing that their findings were incomplete because they did not include a comparison group of states without high-stakes testing programs. His study showed that states that attached consequences to testing outperformed a comparison group of states without high-stakes tests on three NAEP tests for the last 4-year period.

Amrein-Beardsley and Berliner (2003) responded to Rosenshine's critique (2003) with their own reanalysis. Using 4 years of NAEP reading and math data, they found that states with high-stakes tests appeared to outperform other states in fourth-grade mathematics but not in fourth-grade reading or eighth-grade mathematics. They also found that states with high-stakes tests exempted more students from participating in the NAEP than did the comparison states without high-stakes tests and that the apparent positive association between high-stakes testing and achievement in fourth-grade math disappeared when test exclusion rates were taken into account. The authors argued that high-stakes testing may provide greater incentives to exclude low-performing students from testing than to increase learning. Increases over time in test exclusion rates for states with strong accountability pressure were confirmed in a study by Nichols, Glass, and Berliner (2006), leading the authors to suggest that "it may be that increasing pressure leads to greater numbers of students dropping out or being held back in school" (p. 50).

Although state-level studies have produced mixed findings regarding the relationship between high-stakes exams and student progress, studies using less aggregated data have found higher rates of retention and dropping out in states and cities that have instituted tougher graduation requirements (Clarke, Haney, & Madaus, 2000; Lilliard & DeCicca, 2001; Orfield & Ashkinaze, 1991; Roderick, Bryk, Jacob, Easton, & Allensworth, 1999; Wheelock, 2003). Using individual-level data from the National Educational Longitudinal Survey, for example, Jacobs (2001) found that graduation tests increased the probability of dropping out among the lowest-ability students. With a similar longitudinal data set, the Chicago Consortium for School Research found that although some students' scores improved in response to a high-stakes testing policy tied to grade promotion, the scores of low-scoring students who were retained declined, relative to those of similar-achieving students who had been promoted, and their drop-out rates substantially increased (Roderick et al., 1999).

Most studies have found that retention in grade negatively affects student achievement and graduation. Summarizing several decades of research, the National Research Council concluded that low-performing students who are held back do less well academically and are far likelier to drop out than are comparable students who are promoted (Heubert & Hauser, 1999). One study, for example, found that retention can increase the odds of dropping out by as much as 250% above those of similar students who were not retained (Rumberger & Larson, 1998).

These findings raise a number of issues for further study. First, studies using the state as the unit of analyses can mask variations across schools and districts in organizational responses and student outcomes. In particular, state-level data sets do not allow examination of differences in school and district capacity that may differentially affect student success and school responses. This is a salient issue, given that a growing body of literature shows that school conditions and teacher quality affect student learning. Studies of teachers' effects at the classroom, school, and district levels have found that teacher effectiveness is a strong determinant of differences in student learning (Darling-Hammond, 2000; Jordan, Mendro, & Weerasinghe, 1997; Wright, Horn, & Sanders, 1997).

Second, the issue of students' exclusion from testing and even from school requires further investigation with student- and school-level data sets. These inquiries should evaluate whether exclusion occurs in systematic ways and whether exclusion of different groups of students influences school, district, and state test score trends.

Prior Research About Accountability Effects in Texas

The creation of the Texas accountability system set the stage for later proclamations of a "Texas miracle," featuring sharp increases in

TAAS scores, apparent decreases in the achievement gap, and decreases in recorded rates of dropping out (Klein, Hamilton, McCaffrey, & Stecher, 2000). However, research on these claims has produced divergent findings. Haney's review of Texas data (2000) identified grade retention, testing exclusion for special education, English-proficiency exemptions, student dropout, and other "illusions" as being the underlying reasons for the apparent increases in test scores. Haney found that retention rates in ninth grade and school-leaving rates for high school students had increased substantially since the late 1980s, with fewer than 50% of African American and Latino ninth graders and only about 70% of White ninth graders progressing to graduation 4 years after entering high school. He argued that part of the increase in pass rates on the 10th-grade exit TAAS was attributable to the increases in the rates at which low-achieving students were missing from the testing pool and, hence, the school accountability ratings.

RAND Corporation researchers Klein and colleagues (2000) reported that although the TAAS mathematics scores were soaring, Texas students did not improve significantly more on the NAEP math test than their did their counterparts nationally and the TAAS gains were not reflected in scores on three other tests that the team administered to Texas students. Their analysis also found "stark differences" between the pictures painted by the NAEP and the TAAS tests regarding the narrowing of the gap in scores between White and minority students, arguing that according to the NAEP results, the achievement gap in Texas was not only large but that it even increased slightly. The researchers pointed to test-based instruction focused on the high-stakes test, as well as testing exclusions, as possible reasons for these disparities.

By contrast, Grissmer, Flanagan, Kawata, and Williamson (2000) found that when children from similar families were compared across states, Texas ranked high in achievement on the NAEP. They suggested that the Texas accountability regime might be one among many plausible explanations for the state's NAEP gains but added that the research design could not establish a causal linkage. This study did not examine test exclusion rates.

In a broader study of state achievement in literacy, another RAND team found that the small apparent gap between the scores of White and Latino students on Texas's state reading tests was not replicated on the NAEP, where the score gap between the two groups was substantially larger (McCombs, Kirby, Barney, Darilek, & Magee, 2005). A study by Linton and Kestor (2003) found evidence of ceiling effects on the TAAS test for White students, given that more than 60% of these students scored in the top 10% of possible test scores. This, they argued, may have created the appearance of narrowing the achievement gap without actually doing so.

Finally, in a study using Texas Education Agency (TEA) data, Carnoy, Loeb, and Smith (2001) did not find increased drop-out rates associated with the TAAS 10th-grade exit examination. They argued that statewide TEA data show that grade retention rose when the first minimum competency tests were introduced in the late 1980s, before the start of the TAAS testing in 1990–1991. Their data also suggest that downward trends in 9th- to 12th-grade student progression ratios ended shortly after the 10th-grade TAAS was implemented in the early 1990s, to which it leveled off thereafter, so that the TAAS may not have caused ongoing declines.

Much of the dispute has focused on the relationship between test score trends and changes in the testing pool—including drop-outs at the secondary school level. Although many cities in Texas report low drop-out rates, student data show sharply dwindling cohort sizes between 9th and 12th grades. For example, although the U.S. Department of Education lists Brazos City as having one of the lowest graduation rates in the United States (National Center for Education Statistics, 2003), the city has reported annual drop-out rates to the TEA below 2% (TEA, 2003). TEA auditors who checked the dropout coding at Brazos City high schools uncovered school use of PEIMS leaver codes that artificially reduced reported drop-out rates at most of them. When the auditors reviewed the records of nearly 5,500 students who left those schools, they found that almost 3,000 students should have been coded as dropouts but were not. This is one example of the broader phenomenon of gaming that has been documented

in some studies of districts and states with high-stakes testing policies.

Accountability Systems and Gaming

Evidence of the effects of high-stakes testing and accountability policies on school responses suggests that high-stakes testing systems that reward or sanction schools on the basis of average student scores may create incentives for schools to boost scores by manipulating the population of students taking the test. In addition to retaining students in grade so that their relative standing will look better on grade-equivalent scores, schools have been found to label large numbers of low-scoring students for special education placements so that their scores are not factored into school accountability ratings (Allington & McGill-Franzen, 1992; Figlio & Getzer, 2002); they have excluded low-scoring students from admission to open-enrollment schools; and they have encouraged poorly performing students to leave school, transfer to GED programs, or drop out (Darling-Hammond, 1991; Haney, 2000; Smith, 1986).

Some studies have found evidence of gaming actions in the grade level just before the one for which school-level scores produce school accountability rankings. For example, Allington and McGill-Franzen (1992) examined trends in the incidence of retention, remediation, and identification of students as handicapped in New York State elementary schools during a period when a high-stakes testing and accountability plan was implemented (1978–1989). They theorized that schools, in response to the fact that special education students' achievement is not included in the school accountability system, might respond by increasing special education assignments as well as by retaining students in grade. They reported a statistically significant increase in the proportion of children retained in grade or identified as handicapped in third grade, just before the fourth-grade high-stakes tests. In considering this relationship, they argued that the removal of low-achieving students from the stream and their delay of entry inflated the reported assessment results, thereby demonstrating no real increase in school effectiveness. Figlio and Getzer (2002) also found that Florida schools tended to reclassify low-income and low-performing students as disabled at significantly higher rates following the introduction of a new test-based accountability policy and that these behaviors were concentrated among the low-income schools most likely to be on the margin of failing the state's accountability system.

Similarly, when Massachusetts began requiring a 10th-grade high school exit exam for graduation in 2002, with scores tied to school accountability rankings, graduation rates decreased sharply for African American and Latino students whereas grade retention and dropout/disappearance rates escalated, especially in the 9th grade. Schools with the highest grade retention and drop-out rates experienced some of the steepest increases in test scores. For example, high schools receiving state awards for gains in 10th-grade pass rates on the Massachusetts test showed substantial increases in prior-year 9th-grade retention rates and in the percentage of students who disappeared between 9th and 10th grades (Wheelock, 2003).

At the high school level, gaming may include not only student placements in program categories such as special education but the denial of admissions and the encouragement to leave. Smith (1986) explained the widespread engineering of student populations, which he found in his study of New York City's implementation of test-based accountability as a basis for school-level sanctions:

> Student selection provides the greatest leverage in the short-term accountability game. . . . The easiest way to improve one's chances of winning is (1) to add some highly likely students and (2) to drop some unlikely students, while simply hanging on to those in the middle. (pp. 30–31)

More recent evidence regarding New York City's new exit exam requirements, imposed in 1999, suggests that many of the city's high schools are trying to improve their test scores by pushing out students who are unlikely to pass the tests. By 2000–2001, more than 55,000 high school students were discharged without graduating, a number far larger than the 34,000 seniors who actually graduated from high school (Advocates for Children, 2002), and the number of school-age students in GED programs run by the city schools increased by more than 50%,

from 25,500 to more than 37,000 (*The New York Times,* May 15, 2001, p. A1). A study of England's high-stakes accountability system, which tied school rankings to student scores, also found that it led to a large increase in student exclusion rates (Rustique-Forrester, 2005).

A study by Schiller and Muller (2000) suggests that the nature of the incentive structure may affect school responses. The authors found that more frequent testing increased the odds of graduating when tests carried consequences for students and that teachers used scores to identify at-risk students, presumably for greater attention. They also found, however, that test-based consequences for schools increased the odds of students' dropping out: When schools stood to be sanctioned for low scores, teachers' identifications of at-risk students were associated with more of those students leaving school. This finding is consistent with studies that suggest that when schools are rewarded or punished for students' average scores, there are substantial incentives for low-scoring students to be pushed out of the testing pool in one way or another.

The extent to which high-stakes tests may lead to gaming actions and student exclusion rather than efforts to improve teaching may also depend on school capacity—including whether a school has a stable cadre of skilled teachers who can develop strategies that will meet the needs of struggling students. In many states, schools serving the highest-need students are those with the highest turnover, the greatest numbers of untrained and inexperienced teachers, the fewest monetary and curricular resources, and the least knowledgeable administrators and senior staff (Darling-Hammond & Sykes, 2003). In these contexts, designations that a school is failing may be less likely to result in improvements than in actions to improve average scores by removing the lowest-performing students. Indeed, Rustique-Forrester (2005) found that British schools with lower rates of exclusion had stronger, more expert staffs with more engagement in decision making and greater investments in professional development whereas those with high rates of exclusion had large numbers of inexperienced, untrained, and substitute teachers and few resources devoted to improving staff skills to better meet students' needs.

Similarly, Diamond and Spillane (2004) found that high-performing schools, when under high-stakes accountability policies focused on school scores, increased academic press (i.e., the normative emphasis placed on efforts to improve academic achievement), worked to discover and adopt more effective instructional strategies, and created interventions for students who were lower performing. This was found in contrast to low-performing schools that were on probation—schools with more needy students and less school capacity—which drilled students on test format and narrowed, rather than expanded, their instructional strategies. These latter schools also focused on their higher-performing students, in hopes of getting them to raise their scores, and gave up on their lower-performing students.

DeBray, Parson, and Woodworth (2001) also documented the compliance-without-capacity responses of low-performing schools to accountability pressures in Vermont and New York. Whereas higher-performing schools used the policies to create greater internal accountability around the construction of shared goals, curriculum changes, professional development, and teacher evaluation, the low-performing schools lacked the capacity to mobilize themselves for productive change. In these schools, superficial compliance that focused on the tests was not accompanied by schoolwide initiatives to improve curriculum and instruction.

Mintrop's study (2003) of 11 low-performing schools that were placed on probation in Maryland and Kentucky revealed that most of these schools—with high levels of teacher turnover and little teacher expertise—did not know how to improve. In these schools, teachers did not know how to better teach the students, and they often blamed them for the low performance; furthermore, rather than institute teacher learning processes, administrators responded with control strategies that rigidified teaching. The few schools that were able to improve were those that had more skilled teachers and a principal who created a collegial learning process that could tap their expertise.

In sum, although testing policies may mobilize schools to improve teaching for some students, it appears that these policies—especially where schools and teachers have little knowledge and

TABLE 1
Student Demographics for Texas, Brazos City, and Large Urban School Districts in Texas (2001–2002; in percentages)

	Texas	Brazos City	Large urban district average
African American	14	31	28
Latino	42	56	53
White	41	10	18
Asian / Pacific Islander	3	3	2
Native American	0	0	0
Economically disadvantaged	51	79	67
Limited-English proficient	15	28	27

capacity to improve instruction—can cause the most difficult-to-educate students to be held back, placed in special education, and encouraged to leave. Such actions may make schools look as though they are succeeding on aggregated measures without actually improving their quality.

Method

We use a mixed-methods approach to understand students' school experiences and progression through school, combining descriptive and multivariate analyses of a longitudinal student-level administrative data set secured from Brazos City Independent School District, with interviews with students and staff to learn about their direct experience of the policy.

As one of the large urban school districts in Texas, the district of Brazos City is fairly typical. As in other Texas cities, it mostly serves low-income students who are Latino and African American, and in 2001–2002, just over one quarter of its students were identified as LEP. In sum, these cities have a much greater share of students of color, LEP students, and low-income students when compared to the state as a whole (see Table 1).

Overview of Quantitative Data Set

The student-level data set includes 2,500 variables for 270,000 students over a 7-year period (1995–2002), providing information about student background characteristics (race/ethnicity, income, language proficiency), school placements (grade level), and achievement scores for each year, as linked to teacher and school characteristics: percentages of students by race/ethnicity, income, language status, special education status; the percentage "at risk," as defined by a multifaceted state index; the school accountability rating; and the percentages of certified teachers, new teachers, and rates of teacher turnover.

Unique student identifiers provide the ability to follow students throughout their tenure in the district. The data allowed for the examination of elementary and middle school data from 1995 to 2002. Three cohorts could be followed through high school: those that began high school between 1996 and 1998, representing the graduating classes of 2000, 2001, and 2002. The data set included PEIMs codes designating students as dropouts, withdrawn from school, and graduated. We constructed additional variables for students who were retained in grade and for those who "disappeared" from the data set but who were not coded as withdrawn or dropout.

Qualitative Data Collection

To examine how school staff and students experienced the accountability policies, a research team undertook companion qualitative investigations in a set of Brazos City high schools. The research was initiated with informal focus group discussions with school staff from the Brazos City area about the history of the accountability system and their experiences of the policies. These meetings were conducted to inform the research team on key issues and questions to include in the instrumentation. These meetings were followed by interviews

with staff and students in seven traditional district high schools that agreed to participate in the study. Because most of Brazos City's high schools are "majority minority," we selected three high schools with a majority of Latino students and four with a majority of African American students, all of which had high rates of ninth-grade retention and dropout/disappearance. Nontraditional schools (e.g., charters, schools for previous dropouts) were excluded from the sample. Our sample of schools represents about a third of the traditional high schools in Brazos City School District (BCSD).

We used a key informant strategy and sought out school staff with significant levels of institutional memory in each high school. Schools were asked to randomly choose administrators, staff, and teachers who had more than 5 years of experience in BCSD. A total of 24 math and English teachers were interviewed across the seven high schools. Fourteen BCSD high school administrators and staff (principals, counselors, testing coordinators, etc.) were included in the sample. Additional interviews were conducted with 122 current and former BCSD high school students. Schools were asked to randomly choose 18-year-old students from senior English classes. English classes were chosen because fewer seniors take mathematics; as such, sampling 12th-grade math classes would have biased the sample toward higher-achieving students.

The mixed-methods approach provides an opportunity to triangulate focus group interviews and individual field interviews from almost 200 individuals, representing administrators, students, and teachers, with the trends examined in the quantitative research. In combination, these sources of data allow us to gain a macro-level perspective regarding trends in student achievement, progression, and graduation in Brazos City, along with a micro-level view of the dynamics of student, teacher, and school responses to the evolving incentives offered by the accountability system.

Analyses

Our analyses were designed to address many of the questions raised in the literature about the effects of accountability systems on student performance and school continuation, as well as the effects on school ratings of strategies for serving or excluding students. We sought to understand whether test scores improved and for whom did they improve and whether scores were affected by excluding students from the testing pool or from school altogether, as some previous studies have hypothesized. We also sought to understand whether schools improved their ratings in the accountability system by undertaking gaming strategies affecting student participation.

Individual-level student achievement trends. Our first set of analyses uses the longitudinal student-level data set to track student test score trends on the TAAS, the Texas high-stakes test, and the Stanford Achievement Test–Ninth Edition (SAT-9), a low-stakes test offered in Brazos City. We examine the relationships between students' scores on the two tests, as well as numbers and characteristics of students who were excluded from each test. If achievement gains are real and generalizable, one should expect scores on the high-stakes TAAS to strongly predict performance on the low-stakes SAT-9. To take account of student characteristics and the test exclusions that we found on the TAAS, we use an ordinary least squares regression model to examine the predictors of students' scores on the SAT-9 test as a function of their TAAS scores (or their exclusion from TAAS), their personal background characteristics, and their teachers' qualifications. For this case of k independent variables, the ordinary least squares multiple regression equation model is

$$Y_i = \alpha + \beta_1 X_1 + \beta_2 X_{i2} \ldots + \beta_k X_{ik} + \varepsilon_i.$$

The βs equal the regression coefficients for the independent variables: student demographic characteristics, teacher quality indicators, proportion of at-risk students at the school level, and whether the student had a valid TAAS score. The dependent variable Y represents SAT-9 math and reading scores for student i. The constant α is where the regression line intercepts the y-axis, and ε is the error term reflected in the residuals.

Test score trends and student exclusions. Our second set of analyses focus on the Grade 10

TAAS exit exam, which provides the basis for high school rankings and which some previous analyses have suggested may be associated with high rates of grade retention and eventual dropout. Using the individual level data, we examine trends in test participation and pass rates on each of the 10th-grade tests (reading, writing, and mathematics) over time, and we examine students' grade-to-grade progression through school, tracking students' trajectories over time in relation to race/ethnicity, language status, and income, as well as test-taking history. This allows us to empirically examine whether test exclusions, grade retentions, and dropout or disappearance were widespread, which students were affected, and how they interacted with test score trends. We use individual-level data to track the trajectory of one cohort of entering ninth graders through high school, following them for 6 years (from 1996–1997 to 2001–2002) to track their graduation status, including their test-taking and test-passage experience on the TAAS exit exam. We record not only the district's codes for dropouts, transfers, and withdrawals but also the disappearance rates of students of different types from the database and the district.

School-level achievement and gaming behaviors. Our third set of analyses seeks to ascertain whether high schools may deliberately engage in gaming the system by retaining low-scoring students in grade or by keeping or pushing them out of school to raise scores on the tests used for school rankings. As noted earlier, we conducted in-depth interviews with 160 students, teachers, and administrators in a cross-section of Brazos City high schools to understand how they perceive the influences of the testing and accountability system and what strategies they use to boost scores. We then test whether these strategies have the effects on school rankings that practitioners believe they have.

One of the unique aspects of this data set is that it allows us to use the district's individual-level data to calculate student progress on the school level more accurately than what school self-reported data may actually reflect. This is important given the growing discussion in the literature about whether the data that are publicly reported by schools, districts, and TEA adequately represent student progress through Texas high schools (Greene, 2002; Haney, 2000; Orfield, Losen, Wald, & Swanson, 2004). We use individual-level data to calculate school-level student progress variables from grade to grade and through graduation, in a set of regressions that allow us to investigate the impact of student 9th-grade retention, dropout, and disappearance on 10th-grade high-stakes test scores and school accountability ratings tied to these scores.

Our school-level variables are constructed from BCSD-provided data on students who were officially coded as dropping out or withdrawing from high school—plus a disappearance variable that we created to reflect the proportion of students who were not officially coded by schools or the district as withdrawing or dropping out but who did not continue in a BCSD high school. A variable was also created to measure the proportion of students who were retained in the ninth grade in each school. The school-level variables are expressed as proportions of students in BCSD high schools experiencing each event.

We use a set of regressions to consider the statistical relationships between year-to-year changes in student progress (grade retention, dropout/disappearance, withdrawal) and changes in school test scores and accountability ratings, controlling for changes in the school's teaching capacity and changes in the school's student demographics. First, we use fixed-effects generalized least squares regression models to test the relationship between school-level changes in average exit exam scores and changes in student progression trends, demographics, and teacher capacity. We analyze achievement trends for the population of 24 traditional high schools, arranged in a panel format with school and years as the units of analysis. The model is

$$Y_{it} = \beta_0 + \Sigma\beta_k X_{kit} + \varepsilon_{it},$$

where

$$\varepsilon_{it} = u_i + v_i + w_{it}.$$

As such, β denotes generalized least squares regression coefficients; k indexes measure

independent variables; *i* indexes, high schools; *t* indexes, school years; ε is the error term; *u* is the school component of error; *v* is the error across years; *w* is the random component of error; and β_0 is the intercept. The dependent variable, *Y*, is measured as year-to-year changes in average 10th-grade exit TAAS mathematics and reading TLI scores for each school from 1997 to 2002.

The TLI is a scaled score derived for the TAAS that describes how far a student's performance is above or below the passing standard on each test. The TEA used the TLI as a metric to permit comparisons between TAAS administrations and across grades for use in the accountability system (TEA, 2000). However, the TLI does not represent a vertical scale that extends across the grades; instead, scores are scaled for each grade level each year, using a scale that has a maximum value of 100. If a student receives the same numerical score at consecutive grade levels, he or she is said to have achieved a year of growth.

To predict changes in school-level exit TAAS–TLI scores, we estimate both a random-effects and a fixed-effects model. (A school-fixed effects model is often used to remove bias created by the inability to include controls for unmeasured school characteristics, for example, unchanging aspects of school culture, school staff capacity, parental involvement, and other characteristics that have additive effects.) In this case, effects are fixed for schools and years. We compare the results of the two models and conduct a Hausman test to consider whether the coefficients estimated by the efficient random-effect estimator are the same as the ones estimated by the consistent fixed-effects estimator (Stock & Watson, 2003).

The equations use controls for changes in school-level demographic variables and measures of teaching capacity, including year-to-year changes in student characteristics (percentage of White students, LEP students, special education students, at-risk students) and teacher characteristics (percentage of teachers certified, teachers with less than 3 years of experience, annual teacher turnover). The dependent variable in the fixed- and random-effect regressions considers change in school-level average exit TAAS–TLI for each high school (see appendix for descriptive statistics for variables used in the analysis). Each year-to-year change represents a separate observation in the random, fixed, and multinomial regression models. Year-to-year change variables for student progress, school capacity, and student demographics, as well as exit TAAS–TLI scores, were calculated as

$$\Delta V_t = V_t - V_{t-1}.$$

After considering the relationship between changes in student progress, demographics, and teacher quality and high-stakes exit test scores in the general linear model regressions, we then examine how these sets of independent variables are related to changes in TEA accountability rankings, by conducting a comparable yet independent analysis for the same years (1997–2002). This regression analysis uses multinomial logistic regression, which estimates the probability of a specific event's occurring and so allows consideration of more than two categorical outcomes of the dependent variable. The dependent variable in the multinomial regression is the year-to-year change in a high school's accountability rating. Using blocks of predictor variables (changes in student progress, demographics, and teacher quality), regression coefficients were obtained for three contrasting situations: a decrease in TEA school rating (used as the reference group), no change in rating, or an increase in the school rating. TEA ratings were determined by the state as a function of increases in TLI scores, coupled with officially reported drop-out rates below threshold levels for each rating. The model is

$$\log(\pi_j/\pi_J) = \alpha_j + \Sigma^K_{k=1}\beta_{jk}X_k.$$

Independent variables are denoted by X_k. These influence the probability π_j that category *j* of the response variable will be chosen. In this analysis, 1 is used as the reference category. This analysis tests the relationship between year-to-year school-level changes in TEA accountability ratings and changes in student progression, controlling for changes in student demographics and teacher capacity.

Together, these analyses help us to understand student achievement trends in Brazos City

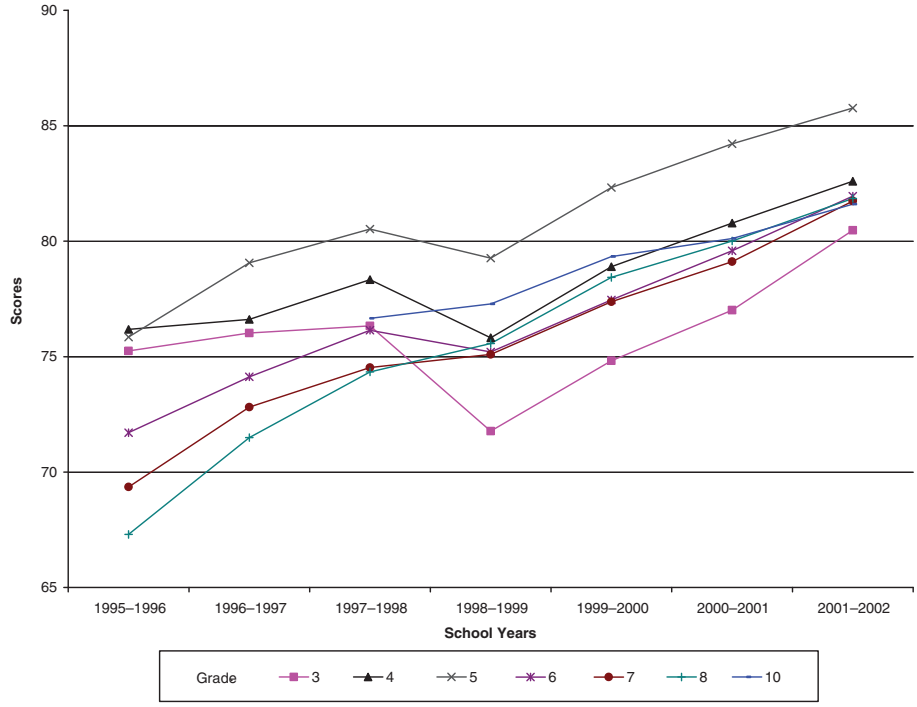

FIGURE 1. *Mean Texas Assessment of Academic Skills–Texas Learning Index scores in mathematics, by grade level.*
Note. Grade 10 was phased into the data set in 1997.

in conjunction with evidence about student exclusion from the testing pool through testing exemptions, grade retention, and exclusion from school. The analyses explore the influences of accountability testing on students' classifications and progress through school for different subgroups of students, as well as the influences on school accountability ratings of engaging in these classification and retention practices.

Student Achievement and Attainment in Brazos City

Achievement and Exclusions on the TAAS

Improvements in TAAS achievement scores for White students and students of color have been widely cited as evidence of major improvements in education in Texas. The TAAS was administered statewide in Grades 3–8 and Grade 10 from 1994 to 2002. Cut scores were set for the 10th-grade TAAS, which served as an exit exam from high school, and TEA accountability ratings for schools were based on the percentage of students passing the TAAS at that grade level.

For cross-year comparisons, we used the TAAS–TLI. We found sharp increases in student performance on this indicator in reading and mathematics, with the exception of a drop in 1998–1999, when there was a large decrease in TAAS testing exemptions for LEP students and a small decrease in special education exemptions, occasioned by public criticism about the numbers of students not tested (see Figure 1). After the drop in scores that occurred when exemptions decreased, math and reading TLIs stabilized for all grades in 1999–2000 and improved steadily until 2001–2002. Grade 3 scores dropped most in 1998–1999, when exemptions were reduced and then remained lower than those of other grade levels. This coincides with a doubling of retention rates at

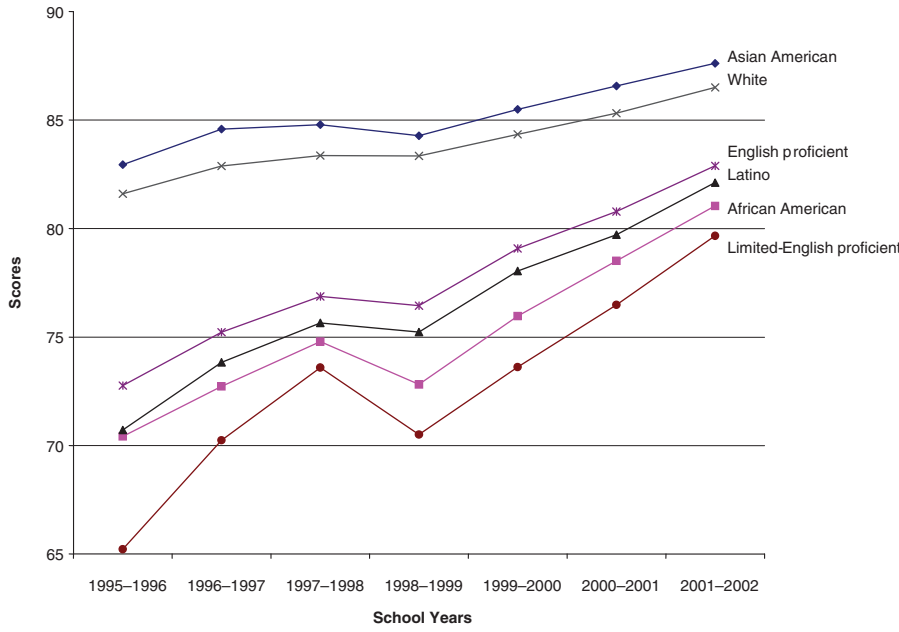

FIGURE 2. *Mean Texas Assessment of Academic Skills–Texas Learning Index scores for Grades 3–8 and Grade 10 in mathematics, by demographic group.*

Grade 3 between 1999 and 2000 (from 5% to 10%) and a continuation of higher-than-average retention rates at this grade in the subsequent years. Holding back low-scoring third-grade students may have depressed scores at that grade level.

Trends in TAAS scores appear to suggest a substantial closing of the achievement gap between racial and ethnic groups, as shown in Figure 2, and between English-proficient and LEP students. African Americans, who represent a disproportionate share of special education students in Brazos City, and LEP students showed the largest drops in scores in 1998–1999, when exemptions from testing were reduced; they also showed the greatest apparent recovery thereafter. However, as we discuss below, the second wave of score increases occurred while there was a corresponding increase in the number of elementary students who were missing test scores altogether.

Two important factors influence the interpretation of these trends. First, the TEA chose not to include achievement on the Spanish TAAS as part of the reported TLI for elementary students (Grades 3–6), which excluded approximately 7,000 Latino students per year from calculations of TLI trends. The proportion of students taking the Spanish TAAS increased over the period that we examined, from 7% in 1995–1996 to 10% in 2000–2001. In addition, the number of students with missing scores increased as exemptions decreased, growing from 2% of scores in 1995–1996 to 10% of all scores in 2000–2001 and 9% in 2001–2002 (see Table 2). Thus, whereas LEP and admission, review, and dismissal exemptions from TAAS testing decreased from 16.5% of all students in 1995–1996 to 2.5% by 2001–2002, most of the students no longer exempted appear to have shifted to the Spanish TAAS test or to missing scores.

Despite an increase in the share of students tested over time, the proportion of Brazos City students included in English TLI scores reached only 78% in 2002, up from 72% in 1996. Thus, more than 20% of the district's students were not included in the state-reported TLI scores in each year. This figure suggests that although some of the gains in TAAS scores may have reflected improvements in learning, much of the increase was likely associated with changes in

TABLE 2
English and Spanish Texas Assessment of Academic Skills (TAAS) Mathematics Score Codes: Grades 3–8 (1995–2002; in percentages)

Students	1995–1996	1996–1997	1997–1998	1998–1999	1999–2000	2000–2001	2001–2002
English TAAS	71.9	71.6	72.4	77.4	77.1	76.2	77.5
Spanish TAAS	7.3	6.3	8.6	9.8	10.3	10.4	10.1
Missing	2.0	4.4	1.7	1.0	1.5	10.0	8.9
ARD exempt	8.5	9.1	9.1	8.4	7.8	0.0	0.0
LEP exempt	8.0	6.6	6.3	1.9	2.0	2.3	2.4
Absent	1.9	1.5	1.5	1.2	0.9	0.8	0.7
Other	0.5	0.5	0.4	0.2	0.3	0.3	0.3

Note. ARD = admission, review, and dismissal; LEP = limited-English proficient.

TABLE 3
Texas Assessment of Academic Skills (TAAS) and Harcourt Mathematics Testing by Race/Ethnicity: Grades 3–8 (1997–2002; in percentages)

Students		Not Harcourt math tested	Harcourt math tested
Asian American	No math TAAS	3	9
	TAAS math tested	1	87
African American	No math TAAS	3	12
	TAAS math tested	2	83
Latino	No math TAAS	2	12
	TAAS math tested	1	85
White	No math TAAS	3	6
	TAAS math tested	2	89

Note. Harcourt = Stanford Achievement Test–Ninth Edition or Aprenda.

the testing pool. This possibility was reinforced by a conversation with a BCSD board member who noted,

> The LEP students were not tested because it would affect the scores. So the outcry was "Your scores are inflated because you don't test everybody." You can't have it both ways. You have to understand what the results are and what went into it, if you test everyone, like we did . . . like [the superintendent] recommended. Okay, you think our scores are inflated, next year we're going to test everybody. And so the scores went down, of course, because you throw everyone into the mix. . . . That's one of the negative consequences: When you test everyone, the scores are going to plummet.

We found much lower rates of student exclusion on the low-stakes SAT-9 tests, instituted by BCSD in 1996 but not considered in the state or district accountability system. In 1997–1998, approximately 93% of BCSD elementary and middle school students were tested on the SAT-9 and Aprenda 2 (Spanish version) mathematics tests, compared to 81% on the English and Spanish TAAS. By 2001–2002, 96% of elementary and middle school students took the SAT-9 and Aprenda 2, compared to 88% who took the TAAS in English or Spanish. In contrast to the TAAS, there was little difference in test-taking patterns by race/ethnicity throughout the period. Given our analysis of the individual-level data, we found that throughout the period of 1997 to 2002, 96% of Latino and Asian students had scores reported on the SAT-9 tests, as did 95% of African American and White students (see Table 3).

On average, about 12% of African American and Latino elementary and middle school students were excluded from the TAAS but not from Harcourt testing (SAT-9 or Aprenda) from 1997 to 2002, which was about double the

TABLE 4
Stanford Achievement Test–Ninth Edition Mean Scores, by English Texas Assessment of Academic Skills (TAAS) Participation Status: Grades 3–8 (1997–2002)

	Stanford Achievement Test–Ninth Edition			
	Reading score (*n*)		Math score (*n*)	
Students	TAAS	No TAAS	TAAS	No TAAS
Overall	459.70 (341,169)	210.54 (65,307)	504.9 (346,663)	273.75 (59,813)
African American	439.95 (125,008)	196.78 (25,737)	467.12 (127,584)	233.91 (23,161)
Latino	415.29 (163,522)	201.21 (33,097)	479.22 (166,004)	282.49 (30,615)
Native American	556.33 (253)	354.90 (28)	588.76 (258)	427.38 (23)
Asian American	598.29 (11,144)	272.27 (1,614)	705.97 (11,221)	455.31 (1,537)
White	654.97 (41,242)	327.76 (4,831)	664.86 (41,596)	361.09 (4,477)

proportion of Whites excluded. Furthermore, Latino students taking the TAAS in Spanish—accounting for 10% of all test takers by 2002—were excluded from the TLI. As a result, much greater proportions of Latino and African American students' test scores were excluded from TEA accountability ratings than what was true for White and Asian American students.

Further examination demonstrated that those excluded from the English TAAS—the basis of the state and district accountability rankings—scored significantly lower on the SAT-9 than did those who took the TAAS, across all racial/ethnic groups ($p < .001$; see Table 4).

These differences in exclusions are likely a major reason why the increases in scores seen on the TAAS were not found on the SAT-9. Although some increase was seen in SAT-9 reading and mathematics scores between 1997 and 1998, average scores were relatively flat from 1998 to 2001 and then decreased in 2002. There was little reduction of the achievement gap by race/ethnicity (Figure 3), and the gap between English speakers and LEP students increased noticeably in mathematics (Figure 4) and to an even greater extent in reading (Figure 5).

Predictors of Student Achievement

The substantial exclusion of low-achieving students on the TAAS probably explains why the correlations between student SAT-9 normal curve equivalent scores and the TAAS–TLI scores (in Grades 3–8 plus Grade 10), though substantial, are more modest than what might be expected ($r = .64$ in reading and .59 in math, $p < .001$). To evaluate the extent to which test exclusion and other policy variables (such as the provision of qualified and experienced teachers) might be associated with student performance independent of student and school demographic characteristics, we estimated the determinants of SAT-9 scores in Grades 3–5, given that they were the grade levels at which students had single teachers who could be attached to their test score records.

A generalized least squares regression analysis uses individual-level data to examine students' test scores on the SAT-9 in relation to their scores (and exclusions) on the TAAS as well as their demographic characteristics. As shown in Model A (see Table 5) and as expected, students' SAT-9 scores are strongly predicted not only by their TAAS–TLI scores but also by their race/ethnicity, language status, income status, and school-level proportions of at-risk students. Through Model B, we examined the effects of a dummy variable, representing inclusion or exclusion in the English TAAS, which proved to exert a strong influence on SAT-9 achievement.

Finally, although exerting a smaller influence, achievement was significantly higher for students with teachers who were certified and had more than 3 years of experience—an additional important policy variable in a district with large proportions of uncertified teachers who are disproportionately allocated to poor and minority students. Together, these findings suggest that low scorers were significantly less

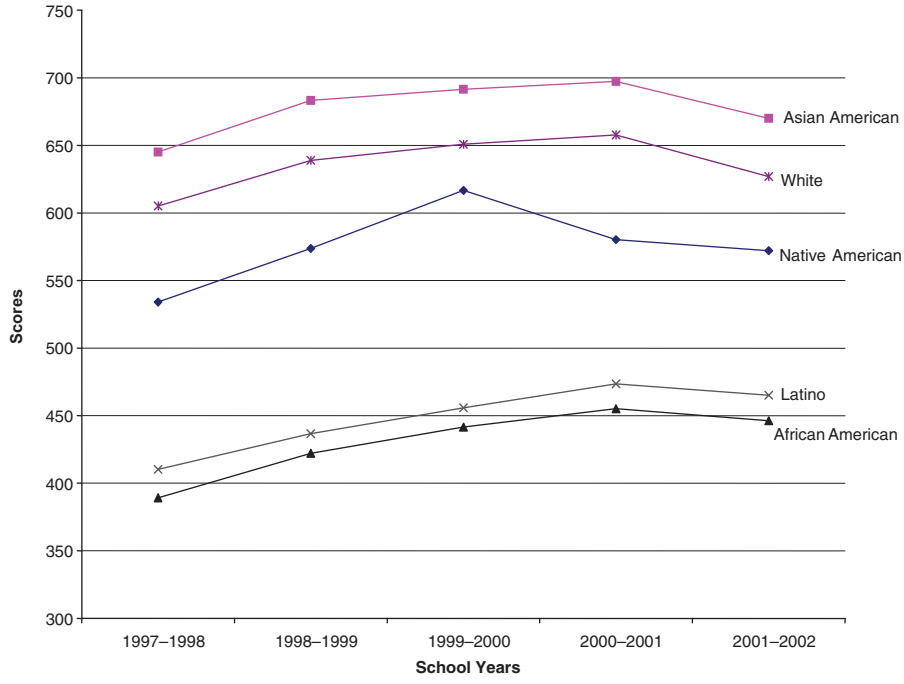

FIGURE 3. *Mean Stanford Achievement Test–Ninth Edition math normal curve equivalent scores for Grades 3–10, by race/ethnicity.*

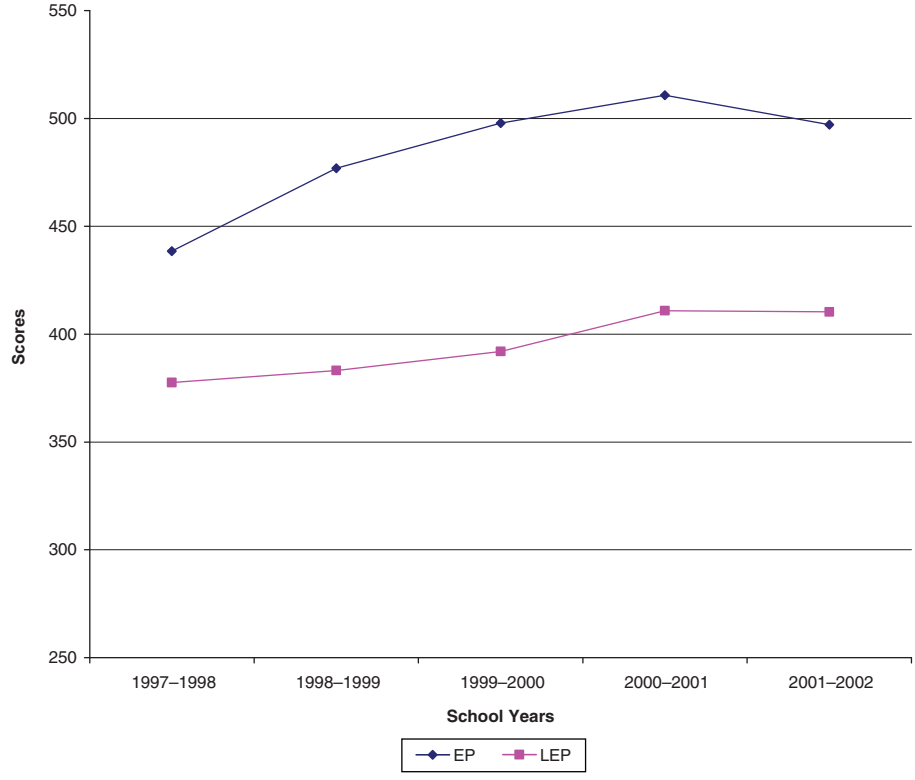

FIGURE 4. *Mean Stanford Achievement Test–Ninth Edition math normal curve equivalent scores for Grades 3–10, by limited-English-proficient status.*

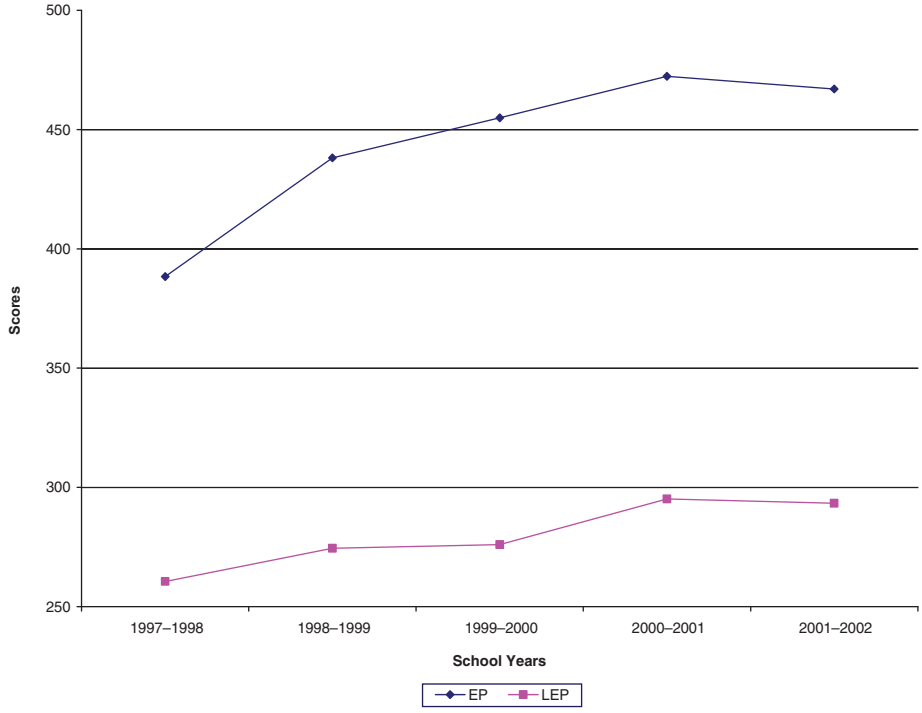

FIGURE 5. *Mean Stanford Achievement Test–Ninth Edition reading normal curve equivalent scores for Grades 3–10, by limited-English-proficient status.*

TABLE 5

Predictors of Stanford Achievement Test–Ninth Edition Scores, Grades 3–5: Random-Effects Generalized Least Squares Regression

	Stanford Achievement Test–Ninth Edition			
	Reading		Math	
Predictor	Model A	Model B	Model A	Model B
Constant	97.833*** (2.751)	487.377*** (1.982)	–10.546*** (2.843)	498.527*** (2.112)
Limited-English proficient	–41.990*** (1.151)	–56.635*** (1.215)	–12.634*** (1.170)	–19.084*** (1.296)
Economically disadvantaged	–47.051*** (0.963)	–62.252*** (1.082)	–34.813*** (0.978)	–52.603*** (1.153)
Native American	–47.755*** (13.748)	–38.186* (16.665)	–28.777** (13.647)	–19.589 (17.482)
Asian American	–8.065*** (2.490)	–17.520*** (3.049)	52.439*** (2.457)	62.756*** (3.176)
African American	–109.519*** (1.422)	–150.155*** (1.724)	–83.471*** (1.408)	137.602*** (1.800)
Latino	–91.527*** (1.509)	–127.648*** (1.807)	–55.757*** (1.492)	–87.796*** (1.889)
Teacher fully certified	4.399*** (0.690)	8.275*** (0.705)	6.786*** (0.725)	10.843*** (0.764)
Teacher experienced	13.814*** (0.726)	12.471*** (0.747)	11.042*** (0.762)	9.877*** (0.809)
School at-risk index	–0.741*** (0.022)	–0.474*** (0.024)	–0.676*** (0.022)	–0.160*** (0.025)
Texas Learning Index score	6.361*** (0.026)	—	8.227*** (0.029)	—
Student has valid TAAS score	—	159.102*** (1.062)	—	159.032*** (1.170)
R^2	.506	.342	.495	.272
Number of observations	163,709	193,339	166,950	193,339

Note. TAAS = Texas Assessment of Academic Skills.

*$p < .05$. **$p < .01$. ***$p < .001$.

likely to be included in high-stakes accountability formulas and more likely to have underprepared and inexperienced teachers and to be in schools with a greater proportion of at-risk students.

High School Exit TAAS Testing Trends

We found that the number of students untested by the TAAS high-stakes tests grew even larger in the high school years but that the reasons changed. In 2002, 79% of sophomores in the BCSD were reported to be passing the TAAS exit exams, up from 72% in 2001, a gain widely celebrated in the local press. We found, however, that these 10th-grade pass rates did not signify that most high school students in BCSD took and passed the exit exam or successfully graduated from high school. Indeed, only a minority did so.

Figure 6 shows the cumulative testing and passing rates by subject for the group of students who began high school as ninth graders in Brazos City in 1997 and would have graduated on time in 2001. This cohort contains more than 13,000 students who were in the eighth grade in BCSD in 1996–1997 and then the ninth grade in 1997–1998. This method leaves out retained eighth graders and previously retained ninth graders from the cohort.

Only 40% of the BCSD 1997–1998 ninth-grade cohort ever passed the writing section of the exit TAAS. The reading and math sections show slightly less student success at 39% and 38%, respectively. Approximately 20% of students took and failed the math, reading, and writing sections of the test. Most surprising, of the original ninth graders in BCSD in 1997–1998, about 40% did not take each section of the exit exam in the district during any academic year between 1997 and 2001. Although there were differentials by race/ethnicity, the pass rates were not high for any ethnic group. For example, only 62% of Asian American students and 54% of White students in the 1997 9th-grade cohort ever passed the 10th-grade reading test in BCSD. Only 38% of African American students and 36% of Latino students passed this portion of the test.

One reason for the low proportions of students taking the exit TAAS is that many did not reach the 10th grade when it was first offered. Among 9th graders who entered high school in 1997, 26% were retained in 9th grade. Ninth-grade retention increased to 31% of all students by 2001. Most of these students never took the exit test. For example, of ninth-grade-retained students in the 1997–1998 cohort, 64% never took the reading portion of the exit test, and only 12% ever passed it. During their high school careers, only 209 of 3,489 retained students (about 8% of the total) ever became eligible to graduate by passing all three subjects on the spring exit TAAS. What happened to these students and others who did not complete high school in BCSD?

Where Did Students Go? Analyses of Student Enrollment Trends

Carnoy et al. (2001) propose two scenarios about the possible relationship between the TAAS and student enrollment outcomes:

> In the first, an emphasis on increasing TAAS scores increases the overall quality of schooling, leading to gains in student learning on multiple levels and decreases in the dropout rate. In an alternative scenario, however, increased emphasis on TAAS comes at the expense of other learning or leads to efforts to screen students before they take the TAAS. This may lead to increases in the dropout rate, either as low performing students are forced out of schools in order to increase school average TAAS scores or as students choose to leave. (p. 18)

To examine whether the alternative scenario occurred in BCSD during the formative years of the Texas test-based accountability system, we examined several measures of student progress in school: grade retention, dropout, withdrawal, disappearance, and graduation. First, however, we examined student mobility because some of the students who never took the tests undoubtedly transferred to other schools outside the district. To estimate the potential upper bound on mobility owing to transfers, we use BCSD's individual-level data set to analyze potential mobility for students who attended middle school (Grades 6–8) in BCSD but were no longer in the district for ninth grade. We expected that the greatest mobility would be found in the transition from middle school to high school, given that this is the point at which many students opt for private schools and suburban public high schools.

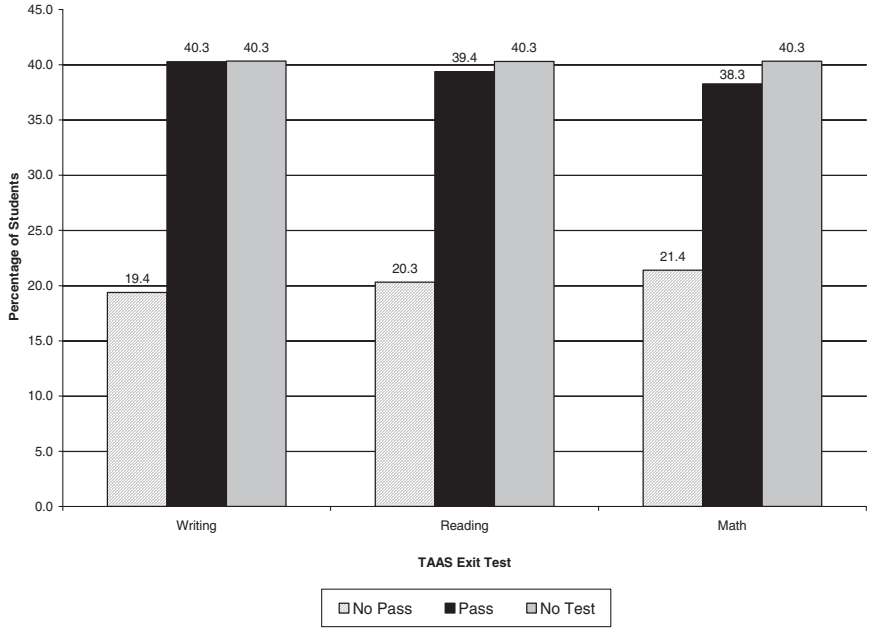

FIGURE 6. *Ninth-grade cohort (1997–1998): Cumulative testing results on the exit Texas Assessment of Academic Skills, by subject.*

Among eighth graders who attended BSCD for middle school in 1998–1999, 97% remained in BCSD the following year. About 1% were officially coded as having dropped out. Given the nature of the accountability system, there was no incentive for schools to overreport dropout data. If there is a bias, it would be to underreport dropouts. This suggests a potential upper-bound mobility rate of a little more than 2% for students advancing to high school from middle school. Assuming an average rate of 2% mobility for each of the 4 years of high school, no more than 8% of BCSD high school students might be considered to have transferred to other schools.

High School Grade Progression

We looked at how students progressed through high school and how many made it through to graduation by following three entering ninth-grade cohorts over a 4-year period. As Figure 7 shows, about half of each class did not progress from 9th to 10th grade in their 2nd year of high school, and an additional 10% did not progress from 10th to 11th grade in their 3rd year. For each cohort, African Americans and Latinos showed the steepest loss between the 9th and 10th grades, given that 50% to 55% of the freshman class did not progress on time, as compared to 30% to 35% of Whites and Asian Americans. All student groups and cohorts showed a smaller (10%) failure to progress from the 10th grade to the 11th grade on time, and showed fairly level progression trends from the 11th to 12th grades. What is interesting is that the 1997–1998 African American and Latino 9th-grade cohort gained students between the 11th and 12th grades. This resulted from another gaming practice: that of "skipping" students past the 10th grade to avoid the TAAS test at the time when it would be factored into the school accountability ratings. Four years after entering 9th grade, about 40% of African American and Latino students in each cohort were enrolled in the 12th grade, as compared to just more than 50% for Whites and Asians.

This analysis highlights the volatile size of the 9th-grade class from year to year: The number of 9th graders increased by nearly 30% between 1996–1997 and 1997–1998 and dropped again by about 10% in the subsequent

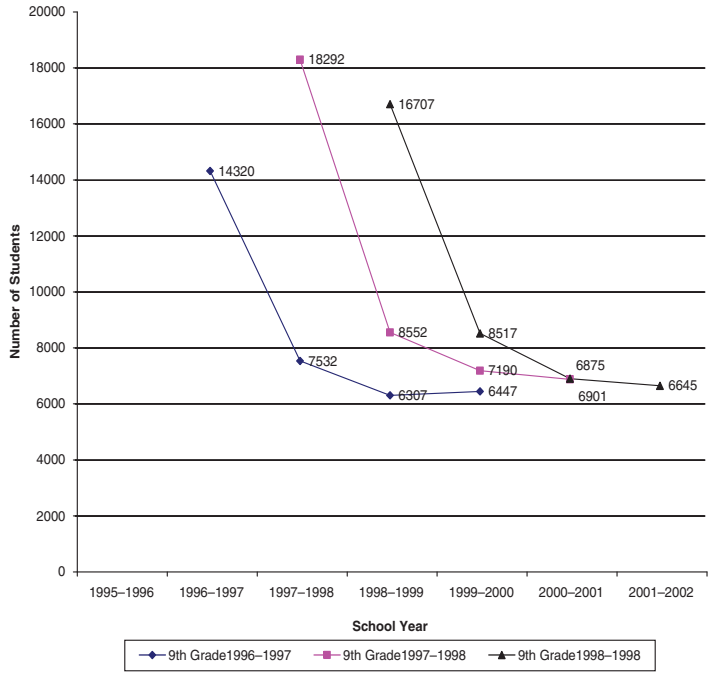

FIGURE 7. *Brazos City School District high school cohort progression (entering ninth graders, 1996–1998).*

year. These changes were not a function of district enrollment changes. Instead, they reflect fluctuations in ninth-grade retention rates, given that the BCSD district accountability system—which was piggybacked on the state system—was introduced in 1995. Some schools began to use substantial grade retention in the 9th grade shortly thereafter, and as we learned in our interviews, other schools followed suit when they learned about the effects of the practice on 10th-grade scores, a point that we explore below.

To understand in detail how students progressed through their high school careers, we followed a first-time 9th-grade cohort (1996–1997) for 7 years. Before entering the 9th grade, about 97% of this entering cohort was in the 8th grade in Brazos City (see Table 6). After the 9th-grade year, 53% of the cohort advanced to the 10th grade, whereas 30% remained in the 9th grade—including one third of African Americans and Latinos and one eighth of Whites and Asians. What is interesting to note is that about 4% of the students in the 9th grade skipped to the 12th grade in the 2nd year, almost all of them African American and Latino. The remaining 13% of students disappeared. By the 3rd year, 44% had progressed to the 11th grade; 32% had disappeared; 20% were still in 9th or 10th grade; and 6% skipped to the 12th grade. In the 4th year, 45% of students had made it to the 12th grade; 40% had left; and 15% still remained in school in grades below the 12th. By the 5th and 6th years, only a few remained in school. Thus, there was not a large increase in graduation rates after the traditional 4 years of high school.

Dropout and Graduation Rates

Clearly, not all the students who failed to graduate dropped out, but as we have suggested, figuring out how many students actually graduated from BCSD and how many transferred to other districts or dropped out is not an easy matter. Earlier, we reviewed differing reports of drop-out rates at the state level in Texas and noted TEA audits revealing that schools severely underreported dropout data in response to incentives in the accountability system. In the TEA accountability baseline year of 1994, drop-out rate standards were first established for schools rated *exemplary* (no more than 1.0%)

TABLE 6
Progression of Ninth-Grade Cohort (1996–1997) Through High School (in percentages)

Grade	Students	1995–1996	1996–1997	1997–1998	1998–1999	1999–2000	2000–2001	2001–2002
8	White/Asian American	98						
	African American/Latino	97						
	All students	97						
9	White/Asian American		100	13	2	2		
	African American/Latino		100	33	6	4	1	
	All students		100	30	6	4		
10	White/Asian American			66	7	2		
	African American/Latino			50	15	4	1	
	All students			53	14	4	1	0.2
11	White/Asian American				55	3	1	0.0
	African American/Latino				42	8	3	1.0
	All students				44	7	1	0.5
12	White/Asian American			1	2	55	2	
	African American/Latino			4	6	43	6	1
	All students			4	6	45	2	1

and *recognized* (no more than 3.5%). In 1995–1996, drop-out rate standards were established for schools rated *acceptable* (no more than 6% for each demographic group).

In our data set, school-reported drop-out rates by grade level hovered below 2%, except those for the ninth grade, which peaked around 3.5% in 1999–2000. Rates for all demographic groups dipped sharply in 1995–1996 and remained just below accountability system threshold levels, positioning schools to meet expectations for a key indicator in the TEA accountability system.

BCSD reported in 2002 that the district graduation rate had soared 21 percentage points over 5 years, from 54% in 1997 to 75.3% in 2002. However, our analyses show that most students failed to pass the exit exam and were not graduation eligible. Arriving at an estimate close to ours, a management team member at King High School estimated from her experience, "I would think that the graduation rate is closer to 40%–45%, not 85%." She continued, "Ultimately what's happening is that we're letting kids down. We're using some kind of system to disguise where they are. If you've got 600 [freshmen] in your school and 300 graduate, they're somewhere . . . "

In a study of dropout data, Orfield and colleagues (2004) noted that the overstatement of graduation rates in Texas may occur partly because Texas's individualized tracking system (PEIMS) includes many ways to exclude students from enrollment data used to calculate graduation rates. For example, a missing student may be taken off the books by a school if he or she is presumed to be in school elsewhere or to have graduated, when that student may in fact have dropped out. In practice, what this means is that if a student does not have an official PEIMS code for the TEA cohort graduation calculation, he or she is dropped from the denominator (TEA, 2002).

Using a cohort method to track graduation rates for students entering ninth grade in 1997, we calculated a graduation eligibility variable to consider students who were eligible for graduation based on having reached senior year and having passed all sections of the TAAS exit exam. As Table 7 shows, of the 13,651 students in the 1997–1998 ninth-grade cohort, 4,111 (30%) appear to be graduation eligible within 5 years. Using the district-provided graduation status code, 4,458 (33%) students are coded as having graduated by 2002 (5-year span).

Asian Americans had the highest completion rate, with almost 50% coded as having graduated. Whites and African Americans had the next-highest rates at 43% and 39%, respectively. Less than a quarter of Latino students in the cohort were coded as having graduated from BCSD by their 5th year. Economically disadvantaged students showed a graduation rate of

TABLE 7
Graduation Rate of Ninth-Grade Cohort Entering High School in 1997 (in percentages)

Student characteristics	Graduation eligible within 5 years	Graduated within 5 years[a]
Overall	30.1	32.7
White	44.9	43.3
Latino	26.1	24.8
African American	29.3	39.4
Asian American	53.1	49.4
Economically disadvantaged	26.3	28.3
Limited-English proficient	14.1	20.0
Did not pass eighth-grade TAAS in reading	7.3	19.3
Did not pass eighth-grade TAAS in math	9.7	22.3

Note. TAAS = Texas Assessment of Academic Skills.
a. As coded by district.

approximately 28%, whereas only 20% of LEP students were coded as graduated.

Of note, African American students appear to graduate at rates 10 percentage points higher than their rates of graduation eligibility. There are two possible explanations for this difference. One possibility is that large numbers of African American students took and passed the make-up summer exit TAAS to gain eligibility for graduation. However, a check of the district's 2001 summer school report shows that even if African Americans composed half of all students passing the make-up exit TAAS, it would increase the overall African American eligibility for graduation by only 2 percentage points, from 29% to 31%.

A more plausible explanation is that African American students received special education (admission, review, and dismissal) exemptions from the exit TAAS testing and could therefore graduate even though they were not identified as being eligible by our measure, which included successful completion of the exam. Unfortunately, the data set does not include a special education identifier for individual students. However, there is a significant and positive correlation between the proportion of African American and special education students ($r = .45$) in traditional BCSD high schools. In national data sets, African Americans are much more likely than other students to be identified for special education. Use of admission, review, and dismissal exemptions for some or all sections of the exit TAAS could have allowed a significant subset of African Americans to graduate without passing all sections of the exit TAAS.

In addition to the 33% of students who were coded as graduates by the 5th year after entering 9th grade, 15% were still seniors and 6% were enrolled in Grades 9–11. About 25% officially withdrew from school; 6% were coded as dropouts; and 18% disappeared from the data set. An evaluation of the status of these students with respect to the exit exam reveals that aside from seniors, of whom two thirds had passed the exam, most students in other categories were not in position to graduate and would have had a strong probability of having dropped out of school (see Table 8). Among those who disappeared from the data set, only 132 of 2,419 had passed all three exit tests. Similarly, among those who withdrew from school, only 362 out of 3,348 had passed the exit tests. Among those still enrolled in Grades 9–11, only a minority (121 out of 809) had passed the tests.

If we assume, on the basis of our earlier analysis, that up to 8% of BCSD high school students may have transferred to other schools and if we paint the cohort's progress in the rosiest possible hues, about 33% graduated within 5 years; 19% were still in school (about a third of them in a position to graduate); an estimated 8% had transferred to public and private schools out of district; and about 40% dropped out, "withdrew" without having passed the exit exam, or disappeared without having passed the test.

TABLE 8
Ninth-Grade Cohort (1997–1998) by Student Progress and Exit Test Status (percentages in parentheses)

	Exit test status		
Students	Not passed	Passed	Total
Enrolled in Grades 9–11	688 (5)	121 (1)	809
Coded as dropout	782 (6)	65 (< 1)	847
Disappeared from data set	2,287 (17)	132 (1)	2,419
Senior (includes graduates)	2,115 (15)	4,113 (30)	6,228
Coded as withdrawn	2,986 (22)	362 (3)	3,348
Total	8,858 (65)	4,793 (35)	13,651

School Gaming and Accountability

It would appear that the large numbers of Brazos City students who failed to take the high school tests and who left school without being coded as dropouts could have enabled schools to appear to meet accountability standards without actually doing so. In this section, we evaluate whether this occurred and, if so, how. We do this by examining what educators and students told us about their experiences and by modeling changes in school accountability ratings as a function of specific practices, such as grade retention and student pushout.

Accountability Pressure in BCSD High Schools

If institutional theorists are correct, schools may react to accountability pressures with ceremonial conformity and various types of gaming responses, some of which may be educationally destructive (Meyer & Rowan, 1977; Oliver, 1991). In hierarchical systems, policy pressure in the form of rewards and sanctions is first applied to district superintendents, then to their subordinates (including principals), then to teachers, and on to students. In Brazos City, incentives included bonuses for principals who raised scores and the probability of termination otherwise, given that contracts were at will. A former BCSD board member described how these rewards and sanctions affected school administrators.

> In the BCSD evaluation system, you [feel pressure] because you have to measure progress. . . . The people who are manipulating it . . . feel pressure because there's a stipend involved. . . . If it's the goal [of accountability] that's set to raise the scores, then you're going to do whatever it takes to raise the scores.

The flip side of monetary incentives involved the threat of firing. An administrator at Edgeview High School explained, "All of us are on at-will contracts. So . . . we can be let go at the end of the year. . . . It's a lot of pressure . . . not even subtle pressure . . . just hard pressure put on you to get those scores up."

We found substantial consensus among the BCSD staff whom we interviewed about the influences of these pressures and the accountability system itself. Several principals identified a widespread culture of gaming the system in BCSD. One said,

> When you talk the company talk, you forget what honesty is. And my fear is that . . . we've forgotten what honesty is. I think that what has happened is that we've gotten all caught up in [accountability] that we don't know what honesty is anymore. I was down at the administration building just today, this morning. I was passing by some secretaries who were there, and they said . . . "Why are they making such a big to-do over these schools that cheat? Everybody cheats." They think everybody is doing it, and it makes it right. . . . That's the feeling that permeates everything that I hear and see.

Gaming Strategies

Administrators and teachers described a culture of gaming and a set of strategies that had been devised to boost ratings. A former Crockett High School staff member observed,

> [You] have to understand the culture of Texas. Texas is a standardized testing machine. We believe in

> evaluation and assessments, quantitative numbers. That's the way the state is run, and that's the culture. So if a principal knows that his ninth graders are not prepared for that examination . . . he is going to put certain mechanisms in place that are going to involve students [being] held back, suspended, so on and so forth.

A management team member at Edgeview described how schools have approached gaming high-stakes testing and accountability year to year.

> I think each year we get a new set of regs, and we try and figure out how is the best way to use it to our advantage. . . . The game changes . . . it's . . . like any—like a game that has a set of instructions. And everybody gets the same set of instructions, and everybody follows the same set of instructions. . . . If you're really savvy and if you're really into everything as a principal, you may see a problem . . . you may give your campus an advantage that another campus doesn't have.

Retention and the waiver policy. One strategy described by administrators and students was the waiver promotion policy, which required freshman and sophomore students to pass all their core courses to move onto the next grade. A management team member at King explained how the waiver allowed schools to avoid testing students on the 10th-grade exam:

> If you're not in a 10th-grade homeroom, then you don't get officially listed as a test taker. . . . So what you ended up with in 10th grade then were all the people who could pass all their core classes. The scores jumped because you put up a barrier, and everybody else was still in ninth grade. . . . With that wall that you could create with the waiver, those kids never entered the accountability picture.

A member of Edgeview's management team detailed how this policy functioned both to increase scores on the 10th grade test and to secure rewards:

> The waiver was set up so that if you did not pass all four core subjects at the ninth-grade level, no matter how many credits you had, you could not push forward. . . . Well, I am taking all these 10th-grade classes, except I have to wait a whole semester to take the [failed section of] Algebra I B. When I finally get that credit, I have enough credits to be a junior. I now have 12 credits rather than my 7 credits. So I skip over taking that 10th-grade test. It's the 10th-grade test that had been used to judge the school. So I have a large group of people who are skipping over the accountability grade. . . . Here in our school, that waiver was used basically to boost the test scores. It had a lot to do with who got the [incentive] money. They wrote a waiver, so they'll circumvent the rule, so they'll know they'll have a higher percentage of students passing the test. . . . There was like a pool of money each school would get.

Focus groups at all the high schools we visited included seniors who discussed peers who had been retained in the ninth grade one or more times. Most focus groups included students who were retained in the ninth grade for 3 or 4 years. Students consistently reported that many of their peers who were retained multiple times eventually gave up and dropped out of school. This comment by a Latino student was typical:

> I have a friend that was in ninth grade for 2 years, and she was 19 or 20 years old. She did not pass algebra, and the school told her that if she didn't improve her grades, they were going to drop her since she was older. So she . . . dropped out of school.

Skipping over 10th grade. Our analysis surfaced the interesting practice of grade skipping, in which students stayed in 9th grade for 2 years or more and then suddenly reappeared in the 12th grade. This practice could have two benefits for schools: First, by skipping 10th grade, students would not take the TAAS exit test in the year that it counted for school accountability ratings. Second, by showing up in 12th grade, they would contribute to a more favorable statistic where school progression is examined as the proportion of 9th graders who appear in 12th grade 4 years later. The practice was described to us as being widespread but mysterious to the students. A Clearbend student talked about her brother's experience:

> This year he was supposed to graduate, but last year he was in the ninth about this time. But he's in the 11th right now, so if he did community service, he'll probably be in 12th. . . . [He was in ninth grade] like about 3 years . . . [then] they made him 12th.

Another Clearbend student spoke about grade skipping and how it had affected a friend:

> This person I was telling you about . . . she was in the same year when we got here but all of a sudden she was in the 12th grade, and then she didn't know

> why. . . . But even though she was a senior, she wouldn't be able to graduate because of [not having test results for the] TAAS.

A disturbing aspect of this grade-skipping practice is that many students never tested on the exit TAAS whether they were held back in 9th grade, progressed to 10th grade, or jumped over their sophomore year. This allowed the schools that they attended to effectively sidestep the high school accountability system. And if students did not test on the exit exam, then they were not eligible to graduate unless they received an exemption, available to special education students.

Keeping or pushing out low-scoring students. We encountered a number of administrators who described schools' refusing to enroll low-scoring students, and we spoke to students who reported trying to enroll at schools but getting turned away. Staff were clear that the ground rules provided no incentive for high schools to keep students who, they believed, would negatively affect their accountability ratings. A member of Edgeview's management team described how students are affected by these pressures:

> I encountered a student just a week ago, and he is 16 years old; this is his first year in the ninth grade. His chances of graduating are slim. . . . Most of the ninth-grade kids are like this—he is going to give up. . . . If he does not make it to the 10th grade, he is going to be 17 years old, and he is going to be a dropout. . . . No school is going to want to take him. They are not going to want him. He is going to screw up their test scores. There are no incentives [to keep him in school] unless you have a principal that is willing to work with these kids. These kids move from school to school and then drop out.

An administrator at Clearbend—a high-minority, low-income school—described the disincentives for high-ranking schools to take or keep students who might lower their accountability ratings:

> I don't think that schools that are blue-ribbon schools in the state of Texas or that are exemplary schools will take students like we have at Clearbend. . . . I've heard stories of schools in our district that turn our kids away. They find a way, and that's wrong. That's morally wrong, but they get away with it . . . and it starts at the top.

The administrator described high schools' financial incentives to keep students in school through October and then push them out: "Many schools unload their troublemakers right after the [enrollment] snapshot. They keep them so they can get the dollars." Students at several high schools explained that their schools were so full at the start of the year that there were no empty desks in classrooms. However, by the time that testing occurred in the spring, there were many desks available because many students were no longer attending the school. The means for pushing students out ranged from enforcing zero-tolerance discipline policies, especially on low-achieving students, to expelling students for attendance problems, and to counseling them out by encouraging them to enroll in GED programs or by transferring them to other nontraditional settings. Students were often explicitly discouraged from taking the exit exam. One administrator explained:

> I think that the kids are being forced out of school. I had a kid who came here from Fine Oaks High School and said, "Miss, if I come here, could I ever take the [exit exam]?" And I said, "What do you mean? If you come here, you must take the [exam]. And he said, "Well, every time I think I'm going to take the test, they either say, "You don't have to come to school tomorrow, or you don't have to [take the test]. . . . We're told different things." That's when kids drop out . . . when you never give them a chance. . . . I think that what has happened at Fine Oaks is what happens at many schools. I think we've done a lot to force kids out of school.

Most administrators whom we interviewed believed that practices were commonplace that manipulated the student population to game the accountability system. As one administrator put it, BCSD was billed as a Texas miracle:

> And I think it's a nonmiracle. It's not a miracle to manipulate things. A miracle is saving kids actually in reality; that's what miracles are. To go out and get these kids who were dropped out or to get kids who are not achieving and find ways—that's a miracle. . . . It's not to manipulate things so that it appears [to be something it's not.] It's a façade.

The Relationship Among Student Retention, Dropout, and School Ratings

The qualitative interviews describe how high schools had responded to the press of

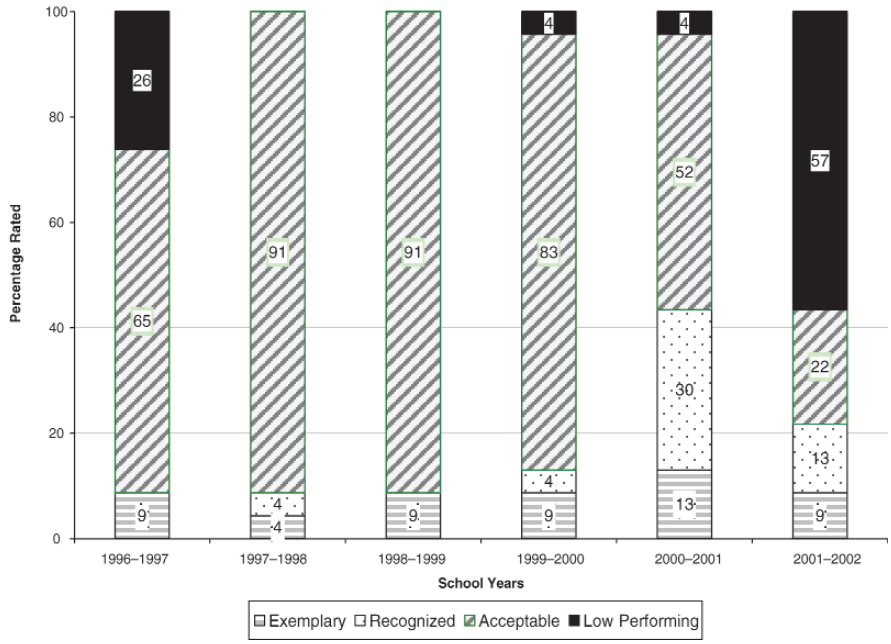

FIGURE 8. *Texas Education Agency accountability ratings for Brazos City School District High Schools (1996–2002).*

accountability by manipulating student populations to raise test scores. Brazos City practitioners perceived that a variety of gaming activities (especially purposeful school-level escalation of grade retention and student exclusion) boosted school test scores and accountability ratings. To triangulate these findings, quantitative analyses were conducted to consider whether BCSD high schools were actually able to raise their test scores and accountability ratings by increasing retention and dropout or disappearance of low-scoring students.

During the formative years of Texas accountability, the primary base indicators for determining a high school's accountability ratings were the exit TAAS scores and the annual dropout rate. Based on these measures, the four typical TEA accountability rating levels were *exemplary*, *recognized*, *acceptable*, and *low performing*. As Figure 8 suggests, between 1996 and 2001, schools figured out how to achieve high ratings. Although 26% were rated *low performing* in 1996–1997, by the next year, no schools were rated as such. From 1997–1998 to 2000–2001, the proportion of schools falling in the top two categories (*recognized* and *exemplary*) increased from 8% to 43%, even while test score standards were also increasing.

In the last year of the period, a large shift occurred as TEA again increased the TAAS passing-rate thresholds required for each category and phased in lower dropout targets. Following these readjustments and with a decrease in ninth-grade retention rates (from 31% to 27% between 2001 and 2002), most BCSD high schools were rated *low performing* once again.

We wanted to evaluate whether BCSD high schools that responded to the press of accountability by escalating dropout, disappearance, and 9th-grade retention were indeed able to increase 10th-grade test scores and, ultimately, TEA accountability ratings. We conducted generalized least squares regression analyses to evaluate whether exiting students raised test scores independent of the change in accountability rating thresholds.

Tables 9 and 10 show the results of analyses examining predictors of changes in reading and mathematics scores, using random effects and fixed effects for year and school. (Fixed effects are often used to account for unmeasured

variables—for example, in the school environment—that may influence the outcomes in question, in this case, test scores.) We use the Hausman test to check the differences in outcomes of the two models, and we find no significant difference, suggesting that the use of fixed effects is not necessary in this case. In each case, we add the student progress measures (changes in school average retention, dropout, disappearance, and withdrawal rates) as a separate block, having controlled for changes in student characteristics and school capacity (teacher certification, experience, and turnover).

We find that adding the vector of variables that represent student progress through school to the equations sharply increases the proportion of explained variance in reading scores, from 19% to 29%, and in math scores, from 5% to 15%. Some changes in school demographic variables influence changes in 10th-grade scores. For example, an increase in the proportion of students identified as "at risk" significantly depresses reading achievement on the TAAS at the 10th-grade level. Furthermore, increases in the proportion of students identified for special education marginally increase reading scores in the fixed-effects model (perhaps because special education identification also increases exemptions of students from the test).

After controlling for these changes, we find that the most powerful predictor of changes in 10th-grade scores in reading and math in all models is an increase in 9th-grade retention. Disappearance from school is marginally significant in the reading models employing random and fixed effects. No other variables are significant for either test.

In terms of effect size, a 7-percentage-point increase in 9th-grade retention predicts a 1-point increase on the math TAAS–TLI score. An 8-percentage-point increase in 9th-grade retention predicts a 1-point increase in a high school's average reading TAAS–TLI index score. A 1-point increase on the reading TLI could also be achieved by an increase of 11 percentage points in student disappearance between 9th grade and 10th grade. By increasing their average school TLI, schools could ascend to a higher TEA accountability rating category.

We examine how school strategies translate into changes in TEA accountability rankings by conducting a separate analysis for the same years (1997–2002) using multinomial logistic regression, which estimates the probability of an event's occurring and allows consideration of more than two categorical dependent variables. Using the same predictor variables, we obtained regression coefficients for three contrasting situations: a decrease in TEA school rating (used as the reference group), no change in rating, or an increase in the school rating. TEA ratings were a function of increases in TLI scores coupled with officially reported drop-out rates below threshold levels for each rating.

The results are compatible with the accountability system incentives. Table 11 shows that, once again, ninth-grade retention rates strongly predict better TEA ratings. The odds of a 1% increase in ninth-grade retention in a school that increased its TEA rating was about 24% greater than in a high school with a TEA rating decrease, both before and after changes in student characteristics and school capacity are controlled. Additionally, as called for in the accountability system, officially reported ninth-grade drop-out rates show a negative coefficient in schools whose TEA ratings rose, as compared to those where ratings declined. As we have seen, these dropout codes bore little relationship to actual school leaving for students.

Not all the responses of schools suggest gaming. It appears that schools could also improve their ratings by increasing their teaching capacity. An increase in school ratings was associated with a positive and significant increase in the percentage of fully certified teachers. Such improvements in teacher qualifications were 20% more likely in high schools that had an increase in TEA ratings than in schools where ratings decreased. Also, schools that had stable ratings were more likely than others to have experienced an increase in students' being classified as "at risk" in models that already controlled for race, language, and special education status, whereas schools that increased their ratings were more likely to have lost special education students than were those whose ratings decreased.

The multinomial regression findings hold despite the fact that TEA test score thresholds were increased and official dropout levels lowered in the final year. Essentially, high schools

(text continues on p. 106)

TABLE 9

Changes in 10th-Grade Scores—Exit Texas Assessment of Academic Skills, Reading: Generalized Least Squares Regression With Random Effects and School/Year Fixed Effects

Predictor	Random effects: Model A	Random effects: Model B	Fixed effects: Model A	Fixed effects: Model B
Constant	1.180** (.452)	.939 (.445)	1.202** (.533)	816 (.529)
Δ School capacity (%)				
Fully certified	–.071 (.045)	–.035 (.045)	–.073 (.050)	–.036 (.050)
Novice teacher	–.040 (.065)	–.033 .(063)	–.039 (.076)	–.032 (.073)
Teacher turnover	–.050 (.046)	–.027 (.045)	–.059 (.051)	–.038 (.049)
Δ School demographic (%)				
White	.127 (.166)	.205 (.166)	.230 (.219)	.301 (.219)
Limited-English proficient	.025 (.078)	.052 (.078)	.081 (.088)	.121 (.088)
Special education	.213 (.155)	.262† (.155)	.239 (.196)	.398** (.203)
At risk	–.083** (.030)	–.099** (.031)	–.087** (.034)	–.101** (.035)
Δ Ninth-grade student progress (%)				
Disappearance	—	.064† (.039)	—	.082† (.046)
Retained	—	.119** (.044)	—	.133** (.051)
Withdrawal	—	.059 (.050)	—	.078 (.059)
Dropout	—	–.214 (.181)	—	–.170 (.201)
R^2	.189	.288	.184	.278
n	94	94	94	94

Note. Numbers in parentheses are standard errors.
†$p < .10$. *$p < .05$. **$p < .01$. ***$p < .001$.

TABLE 10

Changes in 10th-Grade Scores—Exit Texas Assessment of Academic Skills, Math: Generalized Least Squares Regression With Random Effects and School/Year Fixed Effects

Predictor	Random effects: Model A	Random effects: Model B	Fixed effects: Model A	Fixed effects: Model B
Constant	.923 (.504)	.764 (.499)	.983 (.608)	.743 (.622)
Δ School capacity (%)				
Fully certified	–.035 (.049)	.001 (.050)	–.027 (.057)	.003 (.059)
Novice teacher	–.068 (.073)	–.060 (.070)	–.071 (.087)	–.063 (.086)
Teacher turnover	.014 (.051)	.037 (.050)	.014 (.058)	.032 (.058)
Δ School demographic (%)				
White	.256 (.185)	.281 (.186)	.279 (.250)	.258 (.257)
Limited-English proficient	.105 (.087)	.107 (.088)	.130 (.100)	.142 (.103)
Special education	.042 (.173)	.140 (.174)	–.049 (.223)	.142 (.238)
At risk	.015 (.034)	–.012 (.036)	.007 (.039)	–.013 (.041)
Δ Ninth-grade student progress (%)				
Disappearance	—	.024 (.044)	—	.018 (.053)
Retained	—	.146*** (.050)	—	.141** (.060)
Withdrawal	—	.041 (.057)	—	.034 (.069)
Dropout	—	–.136 (.501)	—	–.109 (.236)
R^2	.046	.152	.042	.148
n	94	94	94	94

Note. Numbers in parentheses are standard errors.

*$p < .05$. **$p < .01$. ***$p < .001$.

TABLE 11
Multinomial Logistic Regression of Texas Education Agency (TEA) Accountability Changes (1997–2002): Coefficients and Odds Ratios

Predictor	Random effects: Model A		Random effects: Model B		Fixed effects: Model A		Fixed effects: Model B	
TEA rating	Same	Rise	Same	Rise	Same	Rise	Same	Rise
Δ School capacity (%)								
Fully certified	0.083	0.045					0.086	0.244*
	(0.054)	(0.075)	—	—	—	—	(0.073)	(0.124)
	0.129	1.048					1.090	1.277
Novice teacher	–0.022	–0.043					–0.161	–0.084
	(0.076)	(0.106)	—	—	—	—	(0.110)	(0.173)
	0.774	0.958					0.851	0.919
Teacher turnover	0.047	0.017					0.101	0.073
	(0.055)	(0.077)	—	—	—	—	(0.097)	(0.131)
	1.048	1.018					1.107	1.076
Δ Ninth-grade student progress (%)								
Disappearance			–0.112	–0.003			–0.061	0.133
	—	—	(0.060)	(0.072)	—	—	(0.079)	(0.111)
			0.894	0.997			0.941	1.143
Retained			0.215**	0.294**			0.221**	0.437**
	—	—	(0.071)	(0.094)	—	—	(0.109)	(0.156)
			1.240	1.342			1.248	1.548
Withdrawal			–0.149	–0.066			–0.147	0.044
	—	—	(0.077)	(0.097)	—	—	(0.099)	(0.147)
			0.054	0.936			0.864	1.045
Dropout			–0.150	–0.661*			–0.439	–1.845**
	—	—	(0.240)	(0.316)	—	—	(0.296)	(0.627)
			0.861	0.516			0.645	0.158

(continued)

TABLE 11 (continued)

Predictor	Random effects: Model A		Random effects: Model B		Fixed effects: Model A		Fixed effects: Model B	
TEA rating	Same	Rise	Same	Rise	Same	Rise	Same	Rise
Δ School demographic (%)								
White					0.198	–0.005	0.315	0.680
	—	—	—	—	(0.215)	(0.303)	(0.298)	(0.523)
					1.219	0.995	1.371	1.973
Limited-English proficient					–0.138	–0.261	–0.185	–0.164
	—	—	—	—	(0.126)	(0.179)	(0.208)	(0.310)
					0.871	0.770	0..831	0.849
Special education	—	—	—	—	–0.161	–0.638	–0.030	–1.203
					(0.215)	(0.337)	(0.289)	(0.549)
					0.851	0.528	1.031	0.300
At risk					0.195**	0.040	0.189*	–0.040
	—	—	—	—	(0.068)	(0.085)	(0.093)	(0.138)
					1.215	1.041	1.209	0.960

Note. Numbers in parentheses denote standard errors.
*$p < .05$ **$p < .01$.

that escalated ninth-grade retention maintained or increased their ratings in the years that they escalated their ninth-grade retention. In relation to the school's reference group, the reverse is also true: A decrease in retention portends a decrease in rating. Further study will shed light on whether high schools that responded to the press of accountability by excluding students from 10th-grade testing (and even from school) were able to continue to increase test scores over the long term as they adjusted to the changing formulas and requirements of the state's second-generation accountability policies.

These robust findings regarding the effects of ninth-grade retention on school test scores and accountability ratings support the contention made by Holmes (2006) that when large numbers of students are retained in a grade, the next grade's scores are higher because the low scorers are removed from that year's pool. Although school and district scores may go up, the long-term consequences for individual students are invisible in the yearly snapshot of high-stakes test-score reporting. As a result of retention practices and strategies resulting in the loss of low-achieving students, large gains in school and districtwide test scores can be obtained without improving educational opportunities for those students (Owens & Ranick, 1977).

Discussion

This study has sought to better understand student achievement and progress in Brazos City, a large urban district in the midst of first-generation Texas-style high-stakes testing and accountability. We examined trends in student performance and progression through school while investigating sources of potential gaming that have been identified in other studies: grade retention, student exclusion from testing and from school, and misreporting of indicators that are valued in the accountability system (e.g., drop-out rates).

A major strategy for avoiding the TAAS tests at the high school level was 9th-grade retention. At its peak, more than 30% of 9th-grade students were retained for 1 or more years. Of those who were retained, only 12% ever took the TAAS, and only 8% passed it. A majority of retained students left school as dropouts or disappearances. Although official drop-out rates were kept low—under the annual 3.5% threshold required for a *recognized* school accountability rating—the proportion of students withdrawing or disappearing reached more than 40% of the cohort. More than 90% of students who were coded as *withdrawn* or who disappeared from the data set had failed to pass the exit exam. We also found that some students were kept in the 9th grade for more than 1 year and then skipped to the 11th or 12th grade, thereby never taking the 10th-grade exit TAAS that was used in the accountability ratings.

Although BCSD reported soaring graduation rates and high pass rates on the exit TAAS in 10th grade, our high school cohort analysis for those entering 9th grade in 1997 documented that only 33% of the cohort had graduated within 5 years; that 49% had dropped out, withdrew, or disappeared from the data set (among these, about 8% likely transferred to schools outside the district); and that the remainder were still enrolled in school, trying to make up credits or pass the exit exam. African Americans, Latinos, and LEP students had the lowest graduation rates.

The large discrepancy between publicly reported graduation rates and micro-level student data can be reconciled only by considering the many ways that students are excluded from the enrollment data for calculating drop-out rates. In addition to the large number of disappearances of students from the data set with no codes, most withdrawals appear to be dropouts. This finding aligns with qualitative data that describe how low-achieving students were retained and how they were discouraged from entering and staying in school.

In addition to finding increases in 9th-grade retention rates over the period studied, especially for African American and Latino students, we found that high schools that retained greater numbers of students in the 9th grade—and those with more student disappearances—were able to boost their 10th-grade exit TAAS scores and state accountability ratings. The significant relationships between (a) increases in 9th-grade retention and student disappearances and (b) gains in test scores and ratings hold up before and after considering changing student populations,

teacher capacity, and other measures of student progress. The widespread use of these strategies is confirmed by interviews with students, teachers, and administrators.

In the Texas high-stakes accountability system that we studied, Brazos City schools were forced to organize their responses around snapshot accountability measures based on test scores and reported drop-out rates instead of a long-term measure of student learning and success in completing school. From an institutional theory perspective, this macro-level policy sought to build public confidence in education based on student achievement on standardized tests. As test scores improved, the state and district gained confidence from the media and political system. However, when students did not show test score improvement, the onus of accountability fell on them and their schools, instead of the state. Although many schools and students were handicapped by capacity and resource constraints, the state was able to transfer the consequences for failure to them. Improvements in the quality of education for the least advantaged students did not materialize. Thus, governmental agents, instead of students, became the primary beneficiary of accountability policy, whereas some students were clearly the losers.

It may be possible to construct a high-stakes accountability system without engendering some version of gaming. However, it is apparent that first-generation Texas-style accountability clearly created incentives for pushing out students from high schools and that schools responded to each shift in the incentives by finding new accountability loopholes to manipulate student placements and how data were reported about students. Students also detailed structural obstacles designed to encourage them to leave, such as excessive enforcement of attendance policies, repetitive class and grade-level assignment, and a generally nonsupportive environment for low-achieving students. An important question for the field is whether there is any way to protect low-income, low-achieving students—often students of color and recent immigrants—from bearing the brunt of accountability strategies that impose test-based sanctions on the schools they attend.

Appendix: Summary of Variables Used in School-Level Regression Analyses (1997–2002)

	n	Minimum	Maximum	*M*	*SD*
Δ Achievement scores[a]					
Reading	96	–5.93	6.85	0.92	2.30
Math	96	–4.92	10.02	1.16	2.37
Δ School capacity (%)					
Fully certified	94	–11.66	14.70	3.15	5.34
Novice teacher	94	–13.05	4.97	–4.94	3.76
Teacher turnover	95	–22.37	9.71	–1.47	5.02
Δ Student progress (%)					
9th grade					
Disappearance	96	–24.26	19.48	2.00	8.86
Retained	96	–14.12	15.53	0.86	5.45
Withdrawal	96	–12.88	23.02	0.73	6.81
Dropout	96	–6.14	4.22	0.05	1.27
10th grade					
Disappearance	72	–16.45	19.50	1.82	7.31
Retained	96	–15.27	15.96	2.33	5.31
Withdrawal	96	–12.07	24.43	1.90	6.14
Dropout	72	–4.82	2.60	0.02	1.15

(continued)

Appendix (continued)

	n	Minimum	Maximum	*M*	*SD*
Δ School demographic (%)					
White	96	–10.27	3.04	–0.31	1.48
Limited-English proficient	96	–10.36	12.57	0.11	3.20
Special education	96	–2.76	6.78	0.60	1.48
At risk	96	–12.86	28.05	4.90	7.98
School demographics (%)					
White	96	0.11	60.97	11.70	15.97
Limited-English proficient	96	0.00	42.38	10.59	9.01
Special education	96	0.28	27.77	12.33	5.45
At risk	96	2.13	88.63	65.51	20.59
Free lunch	96	5.22	91.11	39.22	17.23
School capacity (%)					
Certified teachers	95	47.27	86.36	69.23	7.88
Novice teachers	95	0.00	28.95	10.18	7.62
Teacher turnover	96	0.00	100.00	8.59	13.65

a. Tenth-grade exit Texas Assessment of Academic Skills–Texas Learning Index.

References

Advocates for Children. (2002). *Pushing out at-risk students: An analysis of high school discharge figures—A joint report by AFC and the Public Advocate*. Retrieved November 30, 2005, from http://www.advocatesforchildren.org/pubs/pushout11-20-02.html

Allington, R. L., & McGill-Franzen, A. (1992). Unintended effects of educational reform in New York. *Educational Policy, 6*(4), 397–414.

Amrein, A. L., & Berliner, D. C. (2002). High-stakes testing, uncertainty, and student learning. *Education Policy Analysis Archives, 10*(18). Retrieved November 16, 2003, from http://epaa.asu.edu/epaa/v10n18

Amrein-Beardsley, A., & Berliner, D. C. (2003). Re-analysis of NAEP math and reading scores in states with and without high-stakes tests: Response to Rosenshine. *Education Policy Analysis Archives, 11*(25). Retrieved November 16, 2003, from http://epaa.asu.edu/epaa/v11n25/

Carnoy, M., & Loeb, S. (2002). Does external accountability affect student outcomes? A cross-state analysis. *Education Evaluation and Policy Analysis, 24*(4), 305–332.

Carnoy, M., Loeb, S., & Smith, T. (2001). *Do higher state test scores in Texas make for better high school outcomes?* (CPRE Research Report No. RR-047). Philadelphia: Consortium for Policy Research in Education.

Clarke, M., Haney, W., & Madaus, G. (2000, January). *High stakes testing and high school completion*. Retrieved March 16, 2008, from the Web site of the National Board on Educational Testing and Public Policy: http://www.bc.edu/research/nbetpp/publications/v1n3.html

Darling-Hammond, L. (1991). The implications of testing policy for quality and equality. *Phi Delta Kappan, 73*(3), 220–225.

Darling-Hammond, L. (2000). Teacher quality and student achievement: A review of state policy evidence. *Education Policy Analysis Archives, 8*(1). Retrieved July 30, 2004, from http://epaa.asu.edu/epaa/v8n1/

Darling-Hammond, L., & Sykes, G. (2003). Wanted: A national teacher supply policy for education: The right way to meet the "Highly Qualified Teacher" challenge. *Education Policy Analysis Archives, 11*(33). Retrieved April 15, 2007, from http://epaa.asu.edu/epaa/v11n33/

DeBray, E., Parson, G., & Woodworth, K. (2001). Patterns of response in four high schools under state accountability policies in Vermont and New York. In S. H. Fuhrman (Ed.), *Annual yearbook of the National Society for the Study of Education: Vol. 2. From capitol to the classroom: Standards-based reform in the states* (pp. 170–192). Chicago: University of Chicago Press.

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

Figlio, D. N., & Getzer, L. S. (2002, April). *Accountability, ability, and disability: Gaming the system?* Cambridge, MA: National Bureau of Economic Research.

Greene, J. P. (2002). *High school graduation rates in the United States: Revised.* New York: Manhattan Institute for Policy Research. Retrieved December 1, 2007, from http://www.manhattan-institute.org/pdf/cr_baeo.pdf

Grissmer, D. W., Flanagan, A., Kawata, J., & Williamson, S. (2000). *Improving student achievement: What do state NAEP test scores tell us?* Santa Monica, CA: RAND.

Hamilton, L. M., Stecher, B. M., & Klein, S. P. (2002). *Making sense of test-based accountability in education.* Santa Monica, CA: RAND.

Haney, W. (2000). The myth of the Texas miracle in education. *Education Policy Analysis Archives, 8*(41). Retrieved April 15, 2004, from http://epaa.asu.edu/epaa/v8n41/

Hanushek, E., & Raymond, M. (2003). Improving educational quality: How best to evaluate our schools? In Y. Kodrzycki (Ed.), *Education in the 21st century: Meeting the challenges of a changing world* (pp. 193–224). Boston: Federal Reserve Bank of Boston.

Heubert, J., & Hauser, R. (Eds.). (1999). *High stakes: Testing for tracking, promotion, and graduation.* Washington, DC: National Academy Press.

Holmes, C. T. (2006). Low test scores + high retention rates = more dropouts. *Kappa Delta Pi Record, 42*(2), 56–58.

Jacobs, B. A. (2001). Getting tough? The impact of high school graduation exams. *Educational Evaluation and Policy Analysis, 23*(2), 99–122.

Jordan, H. R., Mendro, R. L., & Weerasinghe, D. (1997, June). *Teacher effects on longitudinal student achievement: A preliminary report on research on teacher effectiveness.* Paper presented at the National Evaluation Institute, Indianapolis, IN.

Klein, S. P., Hamilton, L. S., McCaffrey, D. F., & Stecher, B. M. (2000). "What do test scores in Texas tell us?" *Education Policy Analysis Archives, 8*(49). Retrieved February 27, 2006 from http://epaa.asu.edu/epaa/v8n49/

Lee, J., & Wong, K. (2004). The impact of accountability on racial and socioeconomic equity: Considering both school resources and achievement outcomes. *American Educational Research Journal, 41*(4), 797–832.

Lilliard, D., & DeCicca, P. (2001). Higher standards, more dropouts? Evidence within and across time. *Economics of Education Review, 20*(5), 459–473.

Linton, T. H., & Kester, D. (2003). Exploring the achievement gap between White and minority students in Texas: A comparison of the 1996 and 2000 NAEP and TAAS eighth grade mathematics test results. *Education Policy Analysis Archives, 11*(10). Retrieved July 20, 2007, from http://epaa.asu.edu/epaa/v11n10/

McCombs, J. S., Kirby, S. N., Barney, H., Darilek, H., & Magee, S. (2005). *Achieving state and national literacy goals: A long uphill road.* Santa Monica, CA: RAND.

McNeil, L. (2005). Faking equity: High-stakes testing and the education of Latino youth. In A. Valenzuela (Ed.), *Leaving children behind: How "Texas-style" accountability fails Latino youth* (pp. 57–112). Albany: State University of New York Press.

Meyer, J., & Rowan, B. (1977). Institutionalized organizations: Formal structure as myth and ceremony. *American Journal of Sociology, 83,* 340–363.

Mintrop, H. (2003). The limits of sanctions in low-performing schools: A study of Maryland and Kentucky schools on probation. *Education Policy Analysis Archives, 11*(3). Retrieved November 8, 2004, from http://epaa.asua.edu/epaa/v11n3.htm

National Center for Education Statistics. (2003). *Characteristics of the 100 largest public elementary and secondary school districts in the United States: 2001–02*. Washington, DC: U.S. Department of Education.

Nichols, S. L., Glass, G. V., & Berliner, D. C. (2006). High-stakes testing and student achievement: Does accountability pressure increase student learning? *Educational Policy Analysis Archives, 14* (1). Retrieved on August 5, 2007 from http://epaa.asu.edu/epaa/v14n1/

Oliver, C. (1991). Strategic responses to institutional processes. *Academy of Management Review, 16*(1), 145–179.

Orfield, G., & Ashkinaze, C. (1991). *The closing door: Conservative policy and Black opportunity.* Chicago: University of Chicago Press.

Orfield, G., Losen, D., Wald, J., & Swanson, C. B. (2004). *Losing our future: How minority youth are being left behind by the graduation rate crisis.* Cambridge, MA: Civil Rights Project at Harvard University.

Owens, S. A., & Ranick, D. L. (1977). The Greensville program: A commonsense approach to basics. *Phi Delta Kappan, 58*(7), 531–533.

Roderick, M., Bryk, A., Jacob, B., Easton, J., & Allensworth, E. (1999). *Ending social promotion: Results from the first two years.* Chicago: Consortium on Chicago School Research.

Rosenshine, B. (2003). High-stakes testing: Another analysis. *Education Policy Analysis Archives, 11*(24). Retrieved March 16, 2008, http://epaa.asu.edu/epaa/v11n24/

Rumberger, R. W., & Larson, K. A. (1998). Student mobility and the increased risk of high school dropout. *American Journal of Education, 107*(1), 1–35.

Rustique-Forrester, E. (2005). Accountability and the pressures to exclude: A cautionary tale from

England. *Education Policy Analysis Archives.* Retrieved March 16, 2008, http://epaa.asu.edu/epaa/v13n26/

Schiller, K., & Muller, C. (2000). External examinations and accountability, educational expectations, and high school graduation. *American Journal of Education, 108*(2), 73–102.

Smith, F. (1986). *High school admission and the improvement of schooling.* New York: New York City Board of Education.

Stock, J. H., & Watson, M. W. (2003). *Introduction to econometrics*. Boston: Addison-Wesley Higher Education.

Texas Education Agency. (2000). *TAAS technical digest.* Retrieved July 20, 2007, from http://www.tea.state.tx.us/student.assessment/researchers.html

Texas Education Agency. (2002). *Three-year follow-up of a Texas public high school cohort* (Working Paper No. 6). Austin, TX: Author.

Texas Education Agency. (2003). *Secondary school completion and dropouts in Texas public schools, 2001–02*. Austin, TX: Author.

Wheelock, A. (2003). *School awards programs and accountability in Massachusetts: Misusing MCAS scores to assess school quality.* Cambridge, MA: Fair Test.

Wright, S. P., Horn, S. P., & Sanders, W. L. (1997). Teacher and classroom context effects on student achievement: Implications for teacher evaluation. *Journal of Personnel Evaluation in Education, 11*(1), 57–67.

Authors

JULIAN VASQUEZ HEILIG is an assistant professor of educational policy and planning in the Department of Educational Administration at the University of Texas at Austin. His current research includes quantitatively and qualitatively examining how high-stakes testing and other accountability-based reforms and incentive systems impact minority students. Other research interests include issues of access, diversity, and equity in higher education.

LINDA DARLING-HAMMOND is Charles E. Ducommun Professor of Education at Stanford University. Her research, teaching, and policy interests focus on school reform, teaching quality, and educational equity.

Manuscript received February 8, 2007
Final revision received January 2, 2008
Accepted January 23, 2008