Tinkering Toward Accolades:
School Gaming under a Performance Accountability System[†]

Julie Berry Cullen[a,b,*], Randall Reback[c]

[a]*Department of Economics, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0508, USA*
[b]*National Bureau of Economic Research, Cambridge, MA, USA*
[c]*Department of Economics, Barnard College, 3009 Broadway, New York, NY 10027-6598, USA*

Abstract

We explore the extent to which schools manipulate the composition of students in the test-taking pool in order to maximize ratings under Texas' accountability system in the 1990s. We first derive predictions from a static model of administrators' incentives given the structure of the ratings criteria, and then test these predictions by comparing differential changes in exemption rates across student subgroups within campuses and across campuses and regimes. Our analyses uncover evidence of a moderate degree of strategic behavior, so that there is some tension between designing systems that account for heterogeneity in student populations and that are manipulation-free.

JEL classification: D82; H39; I28

Keywords: School accountability; Performance standard; Gaming; Caseload manipulation

[†] The title is inspired by *Tinkering toward Utopia: A century of public school reform* (Tyack and Cuban, 1995), a book describing the evolution of U.S. public schooling during the 20th century.

[*] Corresponding author. Tel.: +1-858-822-2056; fax: +1-858-534-7040.
E-mail address: jbcullen@ucsd.edu (J.B. Cullen).

1. Introduction

Advocates of elementary and secondary education reform believe that the current system does not provide adequate checks and incentives to ensure that teachers and school administrators maximize student learning. Increased accountability is seen as a necessary condition for improving the quality of public schools. The two leading current reform movements to improve accountability involve expanding the educational choices of students and establishing performance standards. In this paper, we examine a problematic design issue associated with the latter type of reform by analyzing behavioral responses to incentives to exempt students from taking exams under Texas' system during the 1990s. While the goal of establishing standards is to improve school efficiency and student outcomes, standards are inevitably imperfect instruments.

Performance-based incentive systems are not new to the public sector and there is a well-developed literature addressing the potential pitfalls.[1] First, public agencies typically pursue multiple goals, of which only some produce measurable outcomes. Since rewards are necessarily tied to measurement, agents may divert resources toward these and away from other valuable outcomes. This issue may be particularly problematic in the education setting where teachers multi-task and where desirable outcomes, such as social adjustment, are often not easily measured.

Second, when outcomes do have empirical counterparts, the performance measures may only be weakly correlated with progress toward program goals. For example, by teaching specifically to the content of high-stakes exams, measured achievement may improve in the absence of general improvements in knowledge and ability in those subject areas.[2] Further, most states base school ratings on pass rates, so that schools may achieve higher ratings by targeting instruction to near-failing students, while not necessarily improving the performance of other students (Reback, 2005).

Finally, the performance measures themselves are often flawed, so that agents can improve reported performance without making progress on actual performance. We refer to behaviors that fall under this final potential pitfall as "gaming," and the form of gaming that we focus on is

[1] Ladd (2001) provides an overview of the theoretical issues associated with designing effective performance standards in the school accountability context.

[2] There is mixed evidence whether improvements on the test instruments used for accountability are matched by parallel gains on other exams (e.g., Hanushek and Raymond, 2005; Jacob, 2005; Klein et al., 2000).

the exclusion of low-achieving students from the test-taking pool.[3] When bureaucracies use heterogeneous inputs, as in the education sector, there is inevitably a trade-off between designing an accountability system that is "fair," in terms of accounting for this heterogeneity, and "manipulation-proof," in terms of ensuring that measured performance represents real accomplishment. For example, some students with pre-existing academic limitations should be legitimately excluded from high-stakes exams, or at least given alternative treatment. On the other hand, when an accountability system allows student exemptions from exams, schools are then able to improve measured performance by manipulating the composition of students taking the exams. Officials at a strategic school might classify additional students in exempt categories, such as special education or limited English proficient, misreport students' statuses, or "encourage" absences primarily to improve aggregate outcomes.

Although the potential for this dysfunctional response is well recognized, there has been relatively limited systematic analysis of the link or its practical relevance.[4] Haney (2000) and Deere and Strayer (2001a) provide descriptive evidence by documenting increases in special education classification rates particularly for minority and low-achieving students, respectively, following the introduction of the Texas policy. In the presence of strong secular upward trends in special education placements,[5] Deere and Strayer (2001b) present more convincing evidence by showing a reversal in the growth in the rate at which special education students sit for achievement exams following a recent policy change that counts scores of special education test-takers. More recent studies use within-state and within-school control groups. For example, Figlio and Getzler (2002) find that schools at risk of failing reclassify previously low-performing students as disabled at higher rates following the introduction of the testing regime in Florida,[6] and Figlio (2005) finds that schools also alter the test-taking pool by strategically assigning long suspensions to low-performing students subject to disciplinary action near the test-taking

[3] Jacob and Levitt (2003) find evidence that the accountability system in place in Chicago has led to a more direct form of gaming—cheating, likely by both students and teachers.

[4] The most extensive previous literature on an analogous form of caseload manipulation under performance-based incentive systems has examined "cream-skimming" of participants by local job training agencies (e.g., Heckman et al., 2002; Heckman et al., 1996; Anderson et al., 1993).

[5] Hanushek and Raymond (2005) caution against drawing strong inferences from pre-post analyses in this setting, and do not find any relationship between special education classification rates and the introduction of strong state accountability policies in a state-level panel analysis that accounts for flexible time trends.

[6] Even when controlling for time trends, Figlio and Getzler (2002) find that the introduction of high-stakes testing led to about a 50 percent increase in the rate that students from low-income families were exempted from test-taking due to special education classifications.

period—and in both cases the behavior is stronger for students in tested grades.[7,8]

Our paper differs from prior studies in that it exploits the specific structure and evolution of the performance targets of the Texas system to formulate more precise school-level incentives. We begin by developing a model of the marginal benefits to administrators from exempting students in order to increase the probability that their schools attain higher ratings. We then conduct empirical tests of two theoretical predictions that arise from the model. First, given that the pass rate standards apply not only to all students but separately to students within race/ethnicity and economic disadvantage subgroups, we analyze changes in exemptions rates among subgroups within the same schools. We predict that exemption rates should increase most for groups whose pass rates are expected to be below the required threshold, because these groups are most likely to constrain the school from attaining the next-highest rating. Our findings confirm this prediction; exemption rates are inflated by up to 7 and 14 percent for Hispanic and Black students, respectively, in years when these students are under-performing relative to peers in the same campus. Although the requirements for separate subgroups are meant to ensure that schools target improvement efforts toward all students, this policy also appears to encourage targeted exemptions.

In our second test, we use student-level test score data to explicitly compute the marginal benefit to a school from exempting additional students between consecutive years. Using these explicit incentives, we analyze whether changes in the level of exemptions at the same school during consecutive years are related to changes in incentives. Strong short-run incentives to exempt additional students are found to raise the likelihood that a school has a one-year increase in the fraction of students classified as exempt by 11 percent. This finding for overall exemptions appears to be driven by more aggressive special education placements and higher absenteeism.

The remainder of the paper proceeds as follows. In the next section, we provide detailed information about the Texas accountability system. Section 3 then presents a conceptual

[7] Jacob (2005) also finds evidence of strategic special education placements using a similar triple-differences strategy that compares trends in special education classification rates for low- versus high-performing students in low- versus high-performing schools in Chicago before and after the high-stakes testing policy.

[8] In addition to students enrolled in tested grades being exempted, low-achieving students can be excluded by being retained in untested grades or encouraged to drop out. Jacob (2005) reports higher retention rates in untested grades following the implementation of the accountability system in Chicago. Haney (2000) reports a similar finding for minorities in Texas, although Carnoy et al. (2001) show that this trend was pre-existing. The literature linking testing standards to dropout rates has focused on the negative impact of grade retention and failure on completion (e.g., Reardon, 1996; Kreitzer et al., 1989), rather than on the interaction with school incentives.

framework for measuring variation in schools' incentives to exempt students that arises from the structure of the accountability system. Sections 4 and 5 describe our empirical strategies and results, respectively. Section 6 offers a brief set of implications and conclusions.

2. Texas Accountability System

2.1 Background

The Texas accountability system originally applied only at the student level in the form of exit exams required for graduation. School-level accountability began in 1993, with schools classified into four rating categories based on student pass rates.[9] The rating categories are: low-performing, acceptable, recognized, and exemplary.[10] The same base indicators for determining ratings were used through 2000, after which the number of tests and grades tested were expanded and an entirely new set of assessments were introduced in 2003.[11] During our sample period, a school's rating depends on the fraction of students who pass Spring Texas Assessment of Academic Skills (TAAS) achievement exams in reading, mathematics, and writing. The reading and math exams are given in grades 3-8 and 10, while the writing exams are given only in grades 4, 8, and 10.[12] For the tested grades combined, all students and four separate student subgroups (White, Hispanic, Black, and economically disadvantaged) must demonstrate pass rates that exceed year-specific standards for each category.[13] In addition, prior year dropout rates must be below and prior year attendance rates must be above threshold levels.

Table 1 displays the year-specific standards for each category for the years 1993 to 1998. There was a planned phase-in of the pass rate standards for the acceptable and recognized categories to 50 percent by 2000 and 80 percent by 1998, respectively. Those two categories

[9] Throughout, we refer to school years by the year associated with the final term, or the fiscal year. That is, we refer to school year 1993-94 as year 1994.

[10] Districts are also assigned one of four ratings, based on identical indicators defined at the district level. Throughout we focus on campus-level incentives, since incentives at both levels are likely to be closely aligned and the kinds of behaviors that could affect exemptions occur at the school level.

[11] Detailed accountability manuals are available for each year at http://www.tea.state.tx.us/perfreport/account/.

[12] There is little interaction between student and school incentives under the accountability system—student performance on the high-stakes tests does not directly affect retention decisions, and graduation requirements can be met by passing end-of-course exams or eventually passing the exit-level TAAS exam.

[13] In the first year of this system, 1993, the subgroups were not held separately accountable. In all years, subgroup pass rates only count when the number of students contributing scores exceeds specified minimum levels. Economically disadvantaged students are those who are eligible for free- or reduced-price lunch or for other public assistance.

also depend on improvement in pass rates from the previous year during the phase-in period.[14] Although the thresholds have evolved, the standards for passing have remained the same.

A school's rating can have real consequences. The ratings are easily understood and are made public. The rating a school attains may affect how attractive it is perceived to be, which could affect its student population, property values, and local support for funding.[15] In addition, ratings can affect the regulatory burden placed on schools. Those placed in the lowest category undergo an evaluation process and may be reconstituted or otherwise sanctioned, including an allowance for students to transfer to better-performing schools inside or outside the district. Schools with the highest rating become exempt from some regulations and requirements. Finally, in most years there have been financial awards for schools that are either high performing or show substantial improvement.

The opportunity for schools to attain a higher rating by gaming this system is related to the way the accountability subset is defined. In order to safeguard schools against the risks of serving disadvantaged populations, there are a number of categories of students that are not included in calculating the school's aggregate pass rates. During our sample period, fiscal years 1993-98, there are four possible reasons why a student would not be tested at all: i) the student is in special education with a severe enough handicap to limit the usefulness of testing, ii) the student is limited English proficient (LEP) and the Spanish test was not offered that year, iii) the student was absent that day, or iv) some other reason (e.g., illness, cheating). There are three possible reasons why a tested student would not have his/her performance contribute to the school pass rate: i) the student is tested but is in a special education program, ii) the student took a Spanish test, or iii) the student was mobile, i.e. not in the district as of October of the school year. Classifying additional students as special needs or LEP or "helping" them to fall into any of these other categories could improve measured performance if these students would likely fail their tests.

Policymakers and practitioners understand the role that differential exemptions can play in upsetting the validity of the ratings categories. The Texas accountability system has evolved to

[14] The ability to substitute a minimum amount of growth in pass rates for a minimum pass rate level could, in theory, introduce important dynamics to schools' incentives. However, this more lenient alternative is only available for the acceptable category and, as Figure 3 shows, only a handful of campuses are ever unable to meet the standards for that ratings category.

[15] Figlio and Lucas (2004) find empirical evidence that the arbitrary distinctions made by Florida's school report card system are capitalized into housing values, particularly in the short run.

address these types of concerns. Since 1999, special education and Spanish TAAS test-takers have been included in the accountability subset, and more recently an assessment specifically designed for disabled students has been incorporated to further increase participation. Our analysis is based on data from the years leading up to these reforms when gaming is likely to have been more prevalent.[16]

2.2 Descriptive trends

Figure 1a presents mean school pass rates by subject and year. For all subjects, there was a dramatic, steady increase in pass rates between 1994 and 1998.[17] Furthermore, this rise in performance occurred for all of the various accountability subgroups. For example, Figure 1b reveals that the average White, Hispanic, Black, and economically disadvantaged pass rates on the math exam increased from 71, 52, 41, and 50 percent to 90, 81, 73, and 79 percent, respectively. Though the pass rate thresholds for various accountability ratings increased over this period (see Table 1), these pass rate improvements were significant enough to cause an upward trend in school ratings. Figure 2 displays the distribution of school ratings by year. Since the TAAS exams are supposed to be comparable over time, on the surface, it appears that there has been tremendous academic improvement.

Figure 3 presents mixed evidence as to whether alterations in the test-taking pool can explain any part of this “miracle.” After rising between 1993 and 1995, mean overall exemption rates have since fallen. The fraction of students exempted due to special education placement increased each year, particularly for disabled students who were able to take the test. The fraction LEP exempt also increased rapidly in the early years, before leveling off. The slowing in the rate of growth in each of these categories could be related to the pending inclusion of special education and Spanish test-takers in the accountability system, which started in 1999 and was first announced in 1996. The overall decrease is driven by pronounced declines in the fraction exempted due to mobility status. The drop in mobility exemptions in 1997 was due to a policy change that tightened the process for identifying a student as new to a district.[18] Thus, the overall decrease in exemptions does not dispel the idea that the accountability system, prior to

[16] Although it would be revealing to track gaming before and after the tightening of restrictions on exemptions, the student-level data that underlie our analysis are not available to us for these more recent years.

[17] The 1993 pass rates are not directly comparable since students were tested in fewer grades in that year.

[18] A student’s information reported on the Spring test document was no longer required to perfectly match Fall enrollment data for the student to be identified as having been in the district. For instance, prior to 1997, a student would have been classified as mobile if his name was entered as Juan Garcia in the Fall and as John Garcia on the Spring exams.

becoming more inclusive, may have led to higher rates of placement in excepted categories than would have otherwise prevailed.

3. Conceptual Framework

3.1 Basic model

In this section, we develop a simple model of the marginal benefits and costs of exempting students under a pass rate threshold system. We do not model the politics and competing objectives of agents within schools, and treat a school's exemption decisions as being made by a single agent, or administrator. Another important simplification is that we treat exemption decisions as depending on current-year incentives, and thus ignore potential dynamic behavioral responses. The framework generates testable predictions that motivate the empirical analyses that we conduct. We focus our theoretical discussion exclusively on the benefits and costs that are directly related to the accountability system, but in our empirical analyses we do control for the important role of fiscal incentives to place students in special programs that arise from the structure of the school finance system (Cullen, 2003).

Though there are three separate subject exams and separate hurdle requirements for student subgroups, start by assuming there is only one exam and that only the overall pass rate matters. Suppose that the administrator at school j takes the current level of exemptions (E_j^0) as given and is deciding how many students to exempt from among the set of students who are in the accountability pool ($N_j - E_j^0$). Define $\hat{P}_{ij}$ to be the administrator's expectation concerning the probability that student i will pass the exam. Let P_{ij} equal one if the student passes the exam and zero if the student fails, so that $P_{ij} = \hat{P}_{ij} + \varepsilon_{ij}$ with $E[\varepsilon_{ij}] = 0$. Assume that the disturbance terms associated with the administrator's predictions are independent, or at least perceived to be independent by the administrator.[19] Then, the administrator's expectation of the overall pass rate is based on the sum of independent binomial outcomes and is approximately normally

[19] This assumption, $Cov(\varepsilon_{rj}, \varepsilon_{sj}) = 0 \; \forall r, s \in \{1, N_j - E_j^0\}; r \neq s$, affects the specific shape of the marginal benefit curve and facilitates the simulation of incentives in our empirical work. It is possible that the surprises to students' pass rates are not independent, such as if students in an entire classroom are not prepared to correctly answer certain questions. Generalizing the framework to account for correlated disturbances would lead to the same basic predictions unless the individual error terms are correlated with each other in such a way that the distribution of the expected overall pass rate becomes severely skewed or multi-peaked.

distributed, with mean equal to the mean student pass probability, $\hat{R}_j = \frac{1}{N_j - E_j^0} \sum_{i=1}^{N_j - E_j^0} \hat{P}_{ij}$, and standard deviation $\sigma_j = \frac{\sqrt{\sum_{i=1}^{N_j - E_j^0} \hat{P}_{ij}(1-\hat{P}_{ij})}}{N_j - E_j^0}$. The standard deviation of the expected pass rate is greater when the underlying predicted probabilities for students are more uncertain (i.e. closer to 0.5) and when there are fewer test-takers.

The distribution of the expected pass rate determines the potential benefits to reclassifying students or otherwise exempting students from the test-taking pool. Let T equal the closest pass rate threshold to $\hat{R}_j$ (either from above or below) and define Z_j to be the marginal benefit associated with attaining the higher rating. If the administrator has risk neutral preferences, the marginal benefit of an additional exemption equals Z_j times the probability that this exemption causes the aggregate pass rate to be above T . The marginal benefit from increasing exemptions by $\Delta E_j = E_j^1 - E_j^0$ can be written as:

$$MB(\Delta E_j) = Z_j \times \left[\Phi\left(\frac{\hat{R}_j(N_j - E_j^1) - T}{\sigma_{j(N_j - E_j^1)}}\right) - \Phi\left(\frac{\hat{R}_j(N_j - E_j^0) - T}{\sigma_{j(N_j - E_j^0)}}\right)\right], \qquad (1)$$

where Φ is the cumulative standard normal density function. We refer to the term in square brackets as the raw (i.e., non-normalized) marginal benefit.

Both the change in the expected overall pass rate and in its standard deviation will influence the raw marginal benefit from increased exemptions. Assuming that administrators shrink the accountability pool by exempting students with the lowest pass probabilities, the expected pass rate will improve so that $\hat{R}_j(N_j - E_j^1) \geq \hat{R}_j(N_j - E_j^0)$. Selective new exemptions will also typically decrease the standard deviation, which on its own has an ambiguous effect on the probability that the overall pass rate exceeds the required threshold, depending on whether the mean expected pass rate is above or below the threshold. Regardless, changes in the expected pass rate will generally dominate, so that additional exemptions are beneficial.

In order to provide intuition for when incentives for additional exemptions are likely to be greatest, it is helpful to graph the marginal benefit curve under different scenarios. Consider an example where an administrator expects 68 percent of the accountability pool to pass this year's exam. Suppose the nearest pass rate threshold is 70 percent, so that the school is deemed “recognized” if the pass rate is at least that high. If the administrator exempts an additional (approximately) 2 percent of students in tested grades, and these students all had near zero probabilities of passing, then the likelihood of reaching the next highest rating increases to 50 percent. Figure 4a shows the shift in the distribution of the expected pass rate, with the associated increase in the probability of meeting the criterion represented by shaded area A. Note that, abstracting from evolutions in the standard deviation that accompany additional exemptions, the experiment could equivalently be carried out by hypothetically sequentially shifting the threshold down. From this perspective, as the administrator increases exemptions starting from the status quo, the raw marginal benefit curve would track the original pass rate distribution curve in reverse starting from where it crosses the relevant pass rate threshold, and the two shaded areas in the figure are identical.

Figure 4b illustrates marginal benefit curves for a school under two pass rate regimes. In Figure 4b, the *x*-axis shows the percentage point increase in the exemption rate (as a fraction of total enrollment), starting at the status quo. These illustrations are derived from the notion established above that the raw marginal benefit curve is closely related to the initial probability density function for the expected overall pass rate, adjusted for changes in the standard deviation as additional students are exempted and scaled by Z_j. As shown, the marginal benefit to exempting additional students when the initial pass rate is above the threshold (e.g., a threshold of 0.66) will generally be steadily decreasing. In contrast, the marginal benefit schedule when the initial pass rate is below the threshold, as in the case shown in Figure 4a, will at first increase and then decrease.

In order to determine how many students a given school would optimally choose to exempt, the marginal benefits have to be traded off against marginal costs. We presume that marginal costs are an increasing function of the level of exemptions, although the testable predictions that we develop in the next section do not depend on this. Exempting additional students likely involves short-run costs associated with reclassification, as well as long-run costs associated with providing a different type of service to the student. Schools are discouraged from classifying students as special education or LEP on test-day without actually assigning special services by the threat of audit if the number exempted exceeds service caseloads. The state also threatens to audit schools with excessively high absenteeism relative to annual attendance rates or high rates of “other” exemptions.

Figure 4b also depicts a representative marginal cost curve. In order to interpret the marginal cost curve appropriately, realize that the *x*-axis equivalently indexes the overall number of exemptions, with the origin set to the current exemption rate. In the example that is depicted, the marginal benefit to additional exemptions exceeds the marginal cost regardless of whether the threshold is high or low. The administrator will in one case want to improve the chances of attaining the higher rating, and in the other reduce the risk of receiving the lower rating. If the thresholds are equidistant from the mean expected pass rate across the two regimes, the administrator will generally behave more aggressively in the high-threshold regime. Obviously, administrators will also often have incentives to decrease exemptions, which in this framework would occur when policy or other changes cause marginal costs to exceed marginal benefits at the existing level of exemptions.

3.2 Testable predictions

Our first empirical test uses within-school variation in whether subgroup pass rates are expected to be above or below the closest relevant threshold. Consider the case where there are two mutually exclusive subgroups in a campus. If one group's expected rate is above the threshold (group A) while the other's is below (group B), exempting students in the former group is associated with muted benefits. The reduction in the risk that group A falls below the threshold translates into only a fractional reduction in the risk that the school is rated in the lower category.

For example, suppose a school's rating is only influenced by the pass rates of these two mutually exclusive subgroups. Using the notation developed above, the marginal benefit of exempting more students in *g*, given that there is another group *h*, is:

$$\begin{aligned} MB_{gj}(\Delta E_{gj}) = Z_j \times & \left[\Phi\left(\frac{\hat{R}_{gj}(N_{gj} - E^1_{gj}) - T}{\sigma_{gj}(N_{gj} - E^1_{gj})} \right) - \Phi\left(\frac{\hat{R}_{gj}(N_{gj} - E^0_{gj}) - T}{\sigma_{gj}(N_{gj} - E^0_{gj})} \right) \right] \\ & \times \Phi\left(\frac{\hat{R}_{hj}(N_{hj} - E^0_{hj}) - T}{\sigma_{hj}(N_{hj} - E^0_{hj})} \right) \end{aligned} \qquad (2)$$

The last term represents the probability that the school satisfies the pass rate requirement for the other group. The marginal benefit to increasing exemptions for group B by a given amount will be greater than for group A as long as: $\frac{\varphi(k_{Bj})}{\Phi(k_{Bj})} > \frac{\varphi(k_{Aj})}{\Phi(k_{Aj})}$, where $k_{gj} = \frac{\hat{R}_{gj} - T}{\sigma_{gj}}$ and ϕ is the standard normal probability density function. If the variances are the same, this condition always holds whenever $k_{Bj} < k_{Aj}$, since log-concavity of the normal density and distribution implies the hazard function $(\phi/(1-\Phi))$ is everywhere increasing. We empirically test the prediction that, controlling for factors related to the variance of the test score distribution and to the cost of exempting students on the margin, there should be larger increases in exemption rates for student groups with expected pass rates below the relevant threshold ($k_{Bj} < 0$) than for student groups with expected pass rates exceeding this threshold ($k_{Aj} > 0$).

Our second empirical test uses student-level data to explicitly calculate shifts in schools' marginal benefit curves between years. Figures 4c and 4d illustrate the relationship between changes in marginal benefits and changes in exemption levels. Figure 4c shows three different potential marginal benefit curves for a school where the mean expected pass rate is in all cases above the relevant threshold, and each curve is associated with a threshold of differing stringency. These marginal benefit curves do not cross, so that one can readily determine whether incentives have increased or decreased. In Figure 4d, where similar curves are shown for a campus with the mean expected pass rate below the threshold values, which regime induces the greatest increase in exemptions depends on the school's cost function. If there are steep marginal costs associated with higher exemption levels, then schools may increase exemptions by less when their expected pass rate is relatively far from the required threshold.

We structure the empirical analysis in order to test predictions that are robust to these underlying ambiguities. First, we characterize the marginal benefit curve local to the observed exemption level in the prior year, and then determine whether the curve shifts up or down over a range of incremental exemptions. Assuming that marginal cost curves are fairly stable across consecutive years, schools with local upward shifts in the marginal benefit curve should be more likely to increase exemption rates than schools with local downward shifts. Further, the greater is the upward shift, the greater is the likelihood of an increase in exemptions from the prior year. Although we also estimate the relationship between changes in the level of exemptions and these changes in incentives, the predictions for the continuous outcome measure are not similarly independent of the shape of the marginal cost curve.

4. Data and empirical strategies

4.1 Data

We combine several administrative data sets, each collected and provided by the Texas Education Agency. These data sets include school- and student-level panel data. The school-level data for the years 1993 through 1998 come from the Academic Excellence Indicator System (AEIS), and are publicly available for download from the Texas Education Agency (http://www.tea.state.tx.us/perfreport/aeis/). The AEIS data are used to determine school ratings and include overall and student subgroup pass rates, attendance and dropout rates, and exemption rates broken down by type of exemption and student subgroup. These data also include a wide variety of campus demographic and financial variables. We purchased the restricted-use student-level data from the Public Education Information Management System (PEIMS) for the same years. These data include test scores, exemption status, and basic demographics for all students in tested grades. Individual students can be tracked across years and campuses through student- and campus-specific identifiers.

The total number of schools in Texas during our sample period ranges from 6,184 in 1993 to 7,053 in 1998. We analyze the 88 percent of schools (representing 97 percent of students) that are rated based on the standard accountability system and their own test scores. Schools that are excluded from our analysis fall into three categories: i) campuses that are not rated because they do not serve students in grades 1-12 (e.g., pre-kindergarten centers and special education schools), ii) alternative education campuses that are evaluated under a different system, and iii) schools that are assigned the test score performance of the school with which they have a feeder relationship since they do not serve students in tested grades (e.g., K-2nd grade and 9th grade centers). Although information from the early years is used in the calculation of incentives for other years, our regression samples exclude 1993, since not all of the relevant grades were tested in that year, and exclude 1994, to allow us to model changes in exemption rates from the prior year.[20]

4.2 Comparison of subgroups within schools

Our first empirical analysis focuses on schools in which there is heterogeneity in the achievement of student subgroups relative to the relevant pass rate standard. As previously shown, we would generally expect a strategic school administrator to increase exemptions by more (or decrease exemptions by less) for groups that are predicted to perform below the

[20] In 1993, only students in grades 4, 8, and 10 were tested.

threshold. We test this proposition by investigating whether these groups exhibit greater one-year changes in exemption rates than other groups within the same school and year.

Since actual student performance is endogenous with respect to exemptions, we determine overall and subgroup-specific expected pass rates by using the prior year performance of students and adjusting for upward trends in achievement. In particular, we find the statewide percentile associated with the lowest-scoring student in a given grade who passed during year *t*, and then calculate the fraction of students in that grade in each school (and each subgroup at the school) that scored at that percentile or better in year *t–1*. School administrators and teachers likely expect an achievement distribution similar to that of the previous year's cohort, adjusted for upward trends in achievement.

Given these expected pass rates, we then determine the target rating for each school by identifying the lowest rating for which at least one relevant pass rate is expected to be below the standard required threshold. We classify a group as "under" if the subgroup contributes to the school's rating and its expected pass rate (on either reading or math) is below this threshold and fails to satisfy any alternative less-binding criteria (such as required improvement).[21] Subgroups that are not classified as "under" either have expected pass rates that satisfy the requirements for the relevant ratings category or are not held separately accountable because there are too few students represented in the group.[22]

A campus may have up to three mutually exclusive race/ethnicity subgroups: White, Hispanic, and Black. For those campus-years with multiple subgroups and across-subgroup variation in "under" status, we estimate the following baseline regression model:

$$\begin{aligned}\Delta e_{gjt} = \alpha_{jt} + \delta_{g} + \theta_{1} \times Under_{gjt} + \theta_{2} \times d_{gjt-1} + \theta_{3} \times d_{gjt-1} \times Under_{djt} \\ + \mathbf{T}_{gjt-1}\boldsymbol{\Omega} + \mathbf{X}_{gjt}\boldsymbol{\Gamma} + \varepsilon_{jgt}\end{aligned}, \qquad (3)$$

where *g*, *j*, and *t* denote the subgroup, school, and year, respectively. The dependent variable is the change in the exemption rate from the prior year. $Under_{gjt}$ is a dummy variable equal to one if the group's pass rate is expected to hold the campus back from attaining the next highest rating. To incorporate any differential incentives to exempt economically disadvantaged

[21] We ignore writing since only rarely do subgroups fail to satisfy that requirement if the reading requirement is met. Math is almost always the binding subject, and in only 10 percent of cases does the subgroup meet the math requirement but fall short of the reading requirement.

[22] Pass rates for student subgroups with either fewer than 30 students tested or fewer than 200 students tested and less than 10 percent of all tested students do not factor into campus ratings under the Texas system. A subgroup is predicted to fail minimum size requirements based on its status in *t–1*.

students and differing degrees of overlap between this subgroup and the race/ethnicity subgroups, we include the fraction of the subgroup that is economically disadvantaged in the prior year (d_{gjt-1}), and an interaction between this fraction and an indicator for whether the economically disadvantaged subgroup is predicted to be "under" ($Under_{djt}$). Importantly, the specification includes campus-year fixed effects (α_{jt}), so that the model identifies the effect of "under" status relative to other student groups in the same campus and year.

The remaining variables included in the control set attempt to eliminate the potential for confounding factors that may be correlated with a group's expected status. First, groups that are "under" will tend to have relatively low academic ability, so that there may be secular differences in exemption patterns for this reason. Also, campuses may face differential marginal costs of exemptions through special program placements for students from different race/ethnicity subgroups for unrelated reasons. To address these concerns, we control for the fraction of students in the subgroup in the prior year with math and reading test scores in various failing ranges ($\mathbf{T}_{gjt-1}$) and subgroup fixed effects (δ_g).[23] The distribution of prior scores also helps to control for differences in the standard deviations of the expected pass rates across groups, as do the control variables related to the relative and absolute size of group *g* included in the vector $\mathbf{X}_{gjt}$.[24]

Although our baseline model controls for any shared school-wide changes in exemptions between years, since we include controls for campus-year, we also report specifications that allow there to be random growth that differs by subgroups within schools. In these cases, the effect of being below the threshold is identified by subgroups that are "under" in some years and not in others.

4.3 Relationship between changes in campus-level exemption rates and incentives

Our second empirical test explores whether schools are more likely to increase exemptions when the potential impact on the likelihood of attaining the relevant rating from additional exemptions increases between back-to-back years. Our approach is to first estimate the raw

[23] For both math and reading, we include the fraction of students in the group scoring in the following failing ranges (i.e., scores below 70) during the prior year: below 50, 50-59, 60-64, and 65-69.

[24] The vector X_{gjt} includes prior year level and changes in the subgroup's share of enrollment in tested grades and the number of students in tested grades (measured by a five-part spline with cut-points defined by quantiles of the size distribution). Also included is the change in the share of students economically disadvantaged.

marginal benefit curve associated with additional exemptions in *t–1*. Then, artificially holding the set of students in the accountability subset fixed, we simulate the raw marginal benefit curve in *t* by adjusting predicted pass probabilities for statewide upward trends and allowing the accountability rules to update. This allows us to isolate exogenous variation in changes in incentives for the same school that arises from changes in the required pass rate thresholds and the general upward trend in student performance over time.

We begin by describing how we estimate the raw marginal benefit curve associated with additional exemptions in *t–1*. First, each student in the accountability subset in *t–1* is assigned a pass probability for each exam equal to the statewide fraction of students with the same test score in *t–2* who actually pass the exam in *t–1*.[25] Second, we calculate the expected pass rate for all students and each student subgroup on each exam ($\hat{R}_{gj}$), as well as the standard deviation in these rates (σ_{gj}), in *t–1* by aggregating the student-level pass probabilities.

If these performance measures were independent, we could readily calculate the probability that a school meets the relevant targets for each rating by calculating the probability each student group exceeds the requirement for each subject and then multiplying these probabilities across subjects and groups. However, shocks to a student's performance may be correlated across exams, and all students contribute to the overall pass rate as well as possibly to that of a race/ethnicity and/or the economically disadvantaged subgroup. For tractability, we assume that math and reading performance are independent, but assume that writing requirements will always be satisfied if reading requirements are satisfied.[26] In aggregating across groups, we use the lowest of the following three probability measures for math and reading: (1) the product of the probabilities that each accountable racial subgroup meets the required threshold, (2) the

[25] All of the calculations are done separately by grade. For grades 4 though 8, we group students who earned identical scores in the prior year on the math, reading, or writing exam, depending which is the outcome of interest. If students are missing prior year scores for certain subjects, we use prior year scores on the other subjects if available. If all prior year scores are missing, we assign the average estimated probability of passing among students who earned the same score in the current year. For grade 10, since students are not tested in grade 9, we group students based on scores in grade 8 (two years earlier), and then apply the same procedures as for the other grades. A few 10th grade students automatically count as passing the exam, because they passed the exam previously, and we set these students' pass probabilities equal to one. For grade 3, since this is the first grade of testing and prior scores are never available, we assign the same pass probability to all students within a school, based on the scores of the previous year's cohort within that school. We use the same method for third graders as described in the previous section: finding the statewide percentile associated with the lowest-scoring third grade student who passed in the current year and calculating the fraction of students in each school's third grade that scored at that percentile or better in the prior year.

[26] This assumption holds very well in the data; in 95 percent of the cases in which the reading pass rate threshold is met, the writing pass rate threshold is also met.

product of the probabilities that the economically disadvantaged subgroup and the White subgroup, if accountable, meet the required threshold,[27] and (3) the probability that the overall campus pass rate meets the required threshold.[28] We are implicitly assuming that whenever the most binding subset of indicators satisfies the requirements, so will the remaining indicators. For example, if it is less likely that a school meets the required pass rate for all race/ethnicity subgroups than for the overall student population, then we presume that the overall pass rate exceeds this threshold in the event that the pass rates for all race/ethnicity subgroups do.

Given this method of using individual student pass probabilities to determine the likelihood that a school obtains a certain rating, we can easily calculate the change in this likelihood if one additional student becomes exempted. We determine the most advantageous exemption for each campus as that which generates the greatest increase in the probability of attaining any of the three ratings, and identify the ratings category most relevant to the campus as the one associated with the maximum increase. We then treat that student as exempted, determine which subgroups continue to meet the minimum size requirements, adjust the expected pass rates and standard deviations, re-calculate the probabilities that the school meets the relevant performance standards, and use these probabilities to determine the new likelihood that the school will obtain the rating. We repeat this process until we have traced out the raw marginal benefit to a campus for exempting various numbers of additional students.

We follow the same process in order to simulate the raw marginal benefit curve in the following year, *t*. We start with the accountability subset in *t–1* as before, and the only differences are that we adjust the student-level predicted pass probabilities for statewide trends and update the policy parameters.[29]

Figure 5 illustrates our resulting measures of the change in incentives for an example case. We determine whether the raw marginal benefits to exempting an additional one, two, and three percent of students in tested grades have all either increased or decreased.[30] We can also measure the amount by which the marginal benefit curve has shifted, shown by areas *a*, *b*, and *c*

[27] While nearly two-thirds of minority students are economically disadvantaged, less than one-third of White students are. Given that there is far less overlap, it is reasonable to treat the White subgroup as distinct.

[28] We also account for dropout and attendance rates—setting the probability of attaining any given rating to zero if these criteria (which are based on prior year outcomes) are not satisfied.

[29] We assign each student a predicted pass probability for each exam in *t* by upgrading the pass probability used in the prior calculations to that at the same percentile of the statewide distribution of predicted pass probabilities (calculated in exactly the same way as for students in *t–1*) for students in year *t*.

[30] We chose to consider exemptions up to an additional three percent since that is the increase at the 75th percentile of the distribution of one-year changes in exemptions.

in the figure. Assuming that the marginal costs of exempting additional students and the marginal benefits to attaining the next highest rating (Z_j) remain relatively constant across consecutive years, schools should be more likely to increase exemptions from the prior year when the raw marginal benefit has shifted up in the directly relevant range, as compared to when it has shifted down.

The baseline regression model that we estimate is:

$$y_{jt} = \beta_1 \times 1_{jt}^{inc} + \beta_2 \times 1_{jt}^{ambig} + \lambda_{ct} + \eta_{ct-1} + \mathbf{X}_{jt}\mathbf{\Pi} + T_{ijt} \times \delta_t + u_{jt}, \quad (4)$$

where *j* and *t* denote the school and year, respectively. The dependent variable is either a dummy variable indicating whether exemptions increased or the change in the exemption rate from the prior year.[31] Our key incentive measure is a dummy variable indicating that the raw marginal benefit curve is higher at all three levels of additional exemptions in the second year of the comparison (1_{jt}^{inc}). If schools behave strategically, then we expect $\beta_1 > 0$. The specification also includes an indicator for cases where the curve shifts up at some points and down at others (1_{jt}^{ambig}), so that the reference group includes cases where the curve shifts everywhere down. To account for the fact that the relevant ratings category varies across campuses and perhaps from the prior year, we include dummies indicating the ratings category predicted to be relevant in the current (λ_{ct}) and prior year (η_{ct-1}). Because our incentive measure is based on simulated changes in the raw marginal benefit curve, we also present results where the sample is restricted to campus-years when the relevant ratings category (and hence Z_j) is predicted to be the same as the prior year.

The vector $\mathbf{X}_{jt}$ includes a variety of campus-level control variables related to the size and demographic composition of students that may affect the marginal costs of exempting students.[32] The predicted overall math and reading pass rates based on the accountability subset in *t–1* are also included to allow for heterogeneity in trends in exemptions related strictly to average performance, so that the variation across campuses in the increase in incentives comes from

[31] We estimate both models using ordinary least squares to facilitate the inclusion of campus fixed effects in some specifications. We have estimated the models with the binary variable as the dependent variable that do not include campus fixed effects using a Probit specification as well. The estimated marginal effects evaluated at the mean are nearly identical to the point estimates that we report for our key variables from the linear model.

[32] We control for the prior year level and change in the following variables: the fraction of students who are in one of the tested grades, the grade and race/ethnicity distribution of students across tested grades, and a five-part spline measuring the number of students in tested grades at the school.

more subtle variation in the distribution of test scores and heterogeneity across student subgroups. Finally, we include year-specific tax base quintile effects to capture secular time effects and schools' fiscal incentives to place students in exempt categories, particularly special and bilingual education, that arise from the structure of the school finance system (Cullen, 2003).

In addition to estimating equation (4), we also present results from models that add campus fixed effects to allow for campus-specific random growth in exemptions. This model identifies strategic responses from differential behavior on the part of campuses across years when incentives have increased and years when they have decreased. Finally, we also estimate a variation that adds an interaction between the indicator for an increase in incentives and the magnitude of the increase, to test whether campuses are more likely to react when there is a greater upward shift in the marginal benefit curve.

5. Results

In this section, we present the empirical results for our two tests. The first test explores whether campuses broadly target exemptions to subgroups for which exemptions are most advantageous from the perspective of attaining or maintaining a higher rating. The second test considers strategic behavior at a finer level—whether campuses respond to short-run increases in incentives to exempt more students that are determined by expected positioning relative to the phased-in pass rate targets.

5.1 Comparison of subgroups within schools

The sample for this analysis includes campus-years that have variation in the "under" status of race/ethnicity subgroups over the period 1995 through 1998. We lose 36 percent of observations for which either only one subgroup is held separately accountable or all are predicted to exceed the less restrictive requirements for the most relevant rating. We exclude an additional 5 percent of observations for campuses that have variation across years in the set of race/ethnicity subgroups that are represented (i.e., at least 5 students in the subgroup in tested grades). This leaves us with 11,026 campus-year observations, representing 59 percent of regular education campuses in those years.[33]

Table 2 presents the summary statistics for this sample. White and Hispanic subgroups are

[33] Although we analyze a restricted sample, the mean rates at which students in the three race/ethnicity subgroups are exempted are nearly the same as in the full sample.

represented at most campuses in our analysis sample, while Black subgroups are represented at only 72 percent of campuses. The disparities in achievement shown in Figure 1b across subgroups underlie the differences in the rates at which the various subgroups are predicted to be holding the campus back from the next highest rating. When present, White, Hispanic, and Black students are predicted to be "under" 24, 64, and 54 percent of the time, respectively. Although Hispanics tend to outperform Blacks, Black subgroups are more likely to be too small to be separately accountable. When White students are "under" they are most commonly in a school with Hispanic and Black peers that are not necessarily out-performing them, but that are not separately accountable. For the majority of cases where the Hispanic and Black subgroups are classified as "under," the White subgroup is predicted to exceed the ratings requirements.

Table 3 presents the regression results for the independent variables of most interest from estimating variants of equation (3). In all cases, the dependent variable is the change in the subgroup's exemption rate from the prior year and the control set includes campus-year fixed effects. The findings confirm our hypothesis that campuses target exemptions toward student subgroups when they are likely to prevent the campus from earning a higher rating. The estimates presented in the first column suggest that membership in a subgroup that is "under" (i.e., a subgroup whose pass rate is expected to be below the required threshold) is associated with a statistically significant 0.6 percentage point increase in exemptions. However, there does not appear to be a differential effect on exemptions for subgroups that have a high fraction low-income when the low-income subgroup is predicted to be "under." There is also evidence of a secular trend toward relative decreases in exemptions for Hispanic and Black subgroups.[34] This underlying tendency may have been partly due to aforementioned longitudinal changes in the feasibility of certain types of exemptions; mobility exemptions became more difficult in the middle of the sample period and schools knew that limited English proficiency exemptions would disappear with the introduction of Spanish exams after our sample period.

Column 2 is a more demanding specification that adds fixed effects for subgroups by campus. These estimates are identified from the relative changes in exemptions, as compared to peers in the same campus, across years when the subgroup is predicted to be holding the campus

[34] Point estimates reveal similarly-sized (although statistically insignificant) negative trends in exemption rates for Hispanic and Black students in the sample of 355 campus-years for which there are no differential incentives to exempt low-performing students from any particular subgroup (i.e., no student subgroup is large enough to be separately accountable).

back and years when there is not an incentive to target exemptions to that subgroup.[35] Here the effect of being "under" rises to a 1.7 percentage point (or more than a five percent) increase in exemptions that is also highly statistically significant.

Columns 3 and 4 consider whether the effect varies by race/ethnicity subgroup, with the second set of estimates from the specification that includes campus-subgroup fixed effects. There is strong evidence of strategic targeting for Hispanic and Black students, but no similar shifting to White subgroups when the subgroup is "under." The net increases for Hispanic and Black subgroups are 1.1 and 1.2 percentage points, respectively, in the model without campus-subgroup fixed effects, and 2.3 and 3.9 percentage points when those are included. These statistically significant one-year changes are equivalent to .15 and .36 standard deviations of their respective mean exemption levels. Incentives related to subgroup performance appear to negate schools' underlying tendency to shift exemptions away from Hispanic and Black students and towards White students over our sample period.

The final two columns of Table 3 test for an exacerbating effect on subgroups that are predicted to be low-performing, and that also have a high share low-income in campus-years when the low-income group is predicted to be "under." In these cases, exempting members of a particular race/ethnicity group may be particularly beneficial. This prediction appears to hold, particularly when the model includes campus-subgroup fixed effects. Consider a racial subgroup that has 50 percent of its students in the low-income category, which is approximately the mean rate. The estimates in column 6 imply that, compared to cases in which neither this race/ethnicity subgroup nor the low income subgroup is "under," when both are "under" exemptions increase by 2.1 percentage points, an effect that is 1.25 times as large as our baseline estimate in column 2.

In results that are not shown, we find that the estimates are quite similar if we split the sample into two periods: pre-1997 and post-1997. We also find that the results are robust to the inclusion of race-year fixed effects. We also do not find any evidence of differential responsiveness depending on whether subgroups are holding the campus back from escaping a low rating or from attaining a high rating.

5.2 Relationship between changes in campus-level exemption rates and incentives

The sample for this analysis is based on campuses for the years 1996 through 1998. We lose

[35] Approximately one-third of campuses have variation across years in the status of at least one group.

observations from 1995 since this is the first year for which there are two years of prior data available for calculating incentives (we require information on 8th graders in *t–2*), and this year serves as the basis for determining whether incentives have increased or not in 1996. After excluding five percent of observations where the relevant ratings category does not remain constant as we trace out the raw marginal benefit curve for increasing exemptions by up to three percentage points, 16,567 campus-year observations remain.

Table 4 presents summary statistics for this sample. On average, 25.3 percent of students are excluded from the accountability subset, and the majority of these students are exempt due to special education placement. Slightly less than half of campuses increase exemptions between years. The incentives to exempt students on the margin are also predicted to have increased from the prior year nearly half of the time. Conditional on increasing, the mean magnitude of the increase is 22.1 percentage points. To interpret this value, recall that we calculate the raw marginal benefit as the change in the probability that the campus meets the requirements for the most relevant rating associated with exempting an additional three percent of students, and that this variable captures the change in this marginal benefit from the prior year.

Table 5 presents ordinary least squares regression results from specifications based on equation (4). Each cell in the table reports the estimated coefficient on an indicator for whether incentives increased from the prior year from a separate regression. The top panel shows results for the full sample, while the bottom panel restricts the sample to campus-years when the relevant ratings category is predicted to remain the same as in the prior year. For the latter sample, there is less uncertainty about whether marginal benefits including the scaling factor Z_j have increased conditional on determining that the raw marginal benefit curve has shifted up.

The results in column 1 are based on specifications where the dependent variable is a binary indicator signaling whether the exemption rate increased from the prior year. In the full sample, campuses with increased incentives are 1.6 percentage points more likely to increase overall exemptions. The effect is larger in the sample with no change in the relevant ratings category. Here, the increase is 3.4 percentage points, or approximately seven percent of the sample mean. These effects for overall exemptions appear to be driven by special education exemptions and exemptions due to absences and other miscellaneous reasons. The specifications in column 2 control for campus fixed effects, allowing for differential growth rates across campuses. Although the point estimates are somewhat attenuated, the loss of statistical significance is primarily explained by loss of precision. Only slightly more than half of the campuses in our sample experience both years with increased incentives and years with decreased incentives.

The dependent variable in column 3 is the change in the exemption rate. While the point estimate is not statistically significant in the full sample, increased incentives lead to a statistically significantly 0.41 percentage point increases in overall exemptions in the restricted sample. Again, the effect can be attributed to special education and absences, which increase by 0.23 and 0.08 percentage points respectively. Only the statistical significance of the overall exemption rate survives the inclusion of campus fixed effects in column 4. These effects are relatively small, ranging from 0.04 to 0.08 standard deviations of the distribution of annual changes. Recall that the theoretical prediction for the size of the change in the level of exemptions is not as clear as for the likelihood that exemptions increase.

Table 6 presents results from specifications that add an interaction term between the indicator for an increase in incentives and the magnitude of the increase to the control set. A larger increase is associated with a greater probability of increasing exemptions between years. To interpret the magnitude of these estimates, compare two administrators: one that has no increase in incentives from the prior year, and one for whom the likelihood of meeting the ratings criteria through additional exemptions has increased by 50 percentage points (one standard deviation above the mean increase). The estimates for the full sample in columns (1a) and (1b) imply that the overall exemption rate would be 2.8 percentage points more likely to increase at the school where the administrator faces enhanced incentives. For the restricted sample, the predicted increase in the likelihood is 5.1 percentage points (or 11 percent). The estimated responsiveness is slightly larger for the second specification that controls for campus fixed effects.

Table 6 also reveals that there are no statistically or economically significant relationships between the magnitude of the change in incentives and the change in the level of exemptions. Although exemptions are more likely to increase at all when there is a large upward shift in the marginal benefit curve associated with incremental exemptions, relatively strong incentives measured locally do not systematically lead to relatively large increases in exemptions.

6. Conclusions

Given the difficulties associated with designing an accountability system that is both fair and manipulation proof, we investigate the extent to which administrators appear to exploit loopholes that allow for exemptions for students that are at an academic disadvantage. We first develop a simple model of the incentives to exempt marginal students in order to secure more favorable ratings under the system that was in place in Texas during the mid 1990s. The model generates two testable predictions that motivate our empirical analyses. In our first empirical test, we find that, regardless of whether incentives to exempt additional students increase or decrease from the prior year, campuses actively preserve or increase those exemptions that are most advantageous. This leads to a targeting of exemptions toward low-performing Hispanic and Black students. In the second test, we find that campuses respond to short-run increases in incentives to expand exemption rates between consecutive years by classifying more students as special needs and encouraging absences.

The gaming that we observe involves potential real costs—costs to the schools that expend resources to engage in this activity, and costs due to decisions that are made based on distorted measures of performance. To the extent the accountability ratings reflect arbitrary differences in classification practices, these misleading ratings can lead to inefficiencies such as misguided educational policy decisions, misguided enrollment decisions, and unwarranted changes in property values. There may be unintended, real (positive or negative) effects on student outcomes if the induced changes in classifications change the type of instruction a student receives or the student's label. The social welfare implications of classifying additional students in special programs depend on the level of pre-existing distortions. Cullen (2003) finds that financial incentives under the school finance system already likely lead to excessive classifications, so that the accountability system exacerbates rather than corrects existing distortions.

Texas' accountability program served as the blueprint for the *No Child Left Behind Act of 2001* (*NCLB*), a federal law requiring all states to adopt standardized testing for students in grades three to eight and to use student proficiency rates in order to rate schools. A key difference between current state systems and the early Texas system is the presence of a minimum participation requirement of 95 percent of students in the tested grades. Critics of this participation requirement are concerned that schools serving high fractions of high needs students are at a disadvantage, but sacrifices in vertical equity may be warranted if the more comprehensive participation requirement under *NCLB* reduces the inefficiencies described above.

Acknowledgements

The authors would like to thank the Communications and Student Assessment Divisions of the Texas Education Agency for providing the student-level longitudinal test documents data, as well as the Office of Tax Policy Research and the Department of Economics at the University of Michigan for providing funding for data acquisition. We also thank Nora Gordon and Jim Ziliak for particularly helpful discussions.

References

Anderson, K.H., Burkhauser, R.V., Raymond, J.E., 1993. The effect of creaming on placement rates under the Job Training Partnership Act. Industrial and Labor Relations Review 46(4), 613-624.

Carnoy, M., Loeb, S., Smith, T.L., 2001. Do higher state test scores in Texas make for better high school outcomes? Consortium for Policy Research in Education Research Report RR-047.

Cullen, J.B., 2003. The impact of fiscal incentives on student disability rates. Journal of Public Economics 87(7-8), 1557-1589.

Deere, D., Strayer, W., 2001a. Putting schools to the test: school accountability, incentives, and behavior. Private Enterprise Research Center Working Paper 114. Texas A&M University.

Deere, D., Strayer, W., 2001b. Closing the gap: school incentives and minority test scores in Texas. Texas A&M University mimeo.

Figlio, D.N., 2005. Testing crime and punishment. National Bureau of Economic Research Working Paper 11194. Cambridge, MA.

Figlio, D.N., Getzler, L.S., 2002. Accountability, ability, and disability: gaming the system. National Bureau of Economic Research Working Paper 9307. Cambridge, MA.

Figlio, D.N., Lucas, M.E., 2004. "What's in a grade? School report cards and the housing market. *American Education Review* 94(3), 591-604.

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

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

Heckman, J.J., Heinrich, C.J., Smith, J., 2002. The performance of performance standards. Journal of Human Resources 37(4), 778-811.

Heckman, J.J., Smith, J.A., Taber, C., 1996. What do bureaucrats do? The effects of performance standards and bureaucratic preferences on acceptance into the JTPA program, in: Libecap, G. (Ed.), Advances in the Study of Entrepreneurship, Innovation, and Growth, Volume 7. Greenwich, CT: JAI Press, 191-218.

Jacob, B.A. 2005. Accountability, incentives and behavior: evidence from school reform in Chicago. Journal of Public Economics 89(5-6), 761-796.

Jacob, B.A., Levitt, S.D., 2003. Rotten apples: an investigation of the prevalence and predictors of teacher cheating. Quarterly Journal of Economics 118(3), 843-877.

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) (http://epaa.asu.edu/epaa/v8n49/).

Kreitzer, A.E., Madaus, G.F., Haney, W.M., 1989. Competency testing and dropouts, in: Weis, L., Farrar, E., Petrie, H.G. (Eds.), Dropouts from School: Issues, Dilemmas, and Solutions. Albany, NY: State University of New York Press.

Ladd, H.F., 2001. School-based educational accountability systems: the promise and the pitfalls. National Tax Journal 54(2), 385-400.

Reardon, S., 1996. Eighth grade minimum competency testing and early high school dropout patterns. Presented at the annual meeting of the American Educational Research Association, New York, NY.

Reback, R., 2005. Teaching to the rating: school accountability and the distribution of student achievement. Unpublished working paper.

Tyack, D., Cuban, L., 1995. Tinkering Toward Utopia: A Century of Public School Reform. Cambridge, MA: Harvard University Press.

Figure 1a. Mean school TAAS pass rate by subject

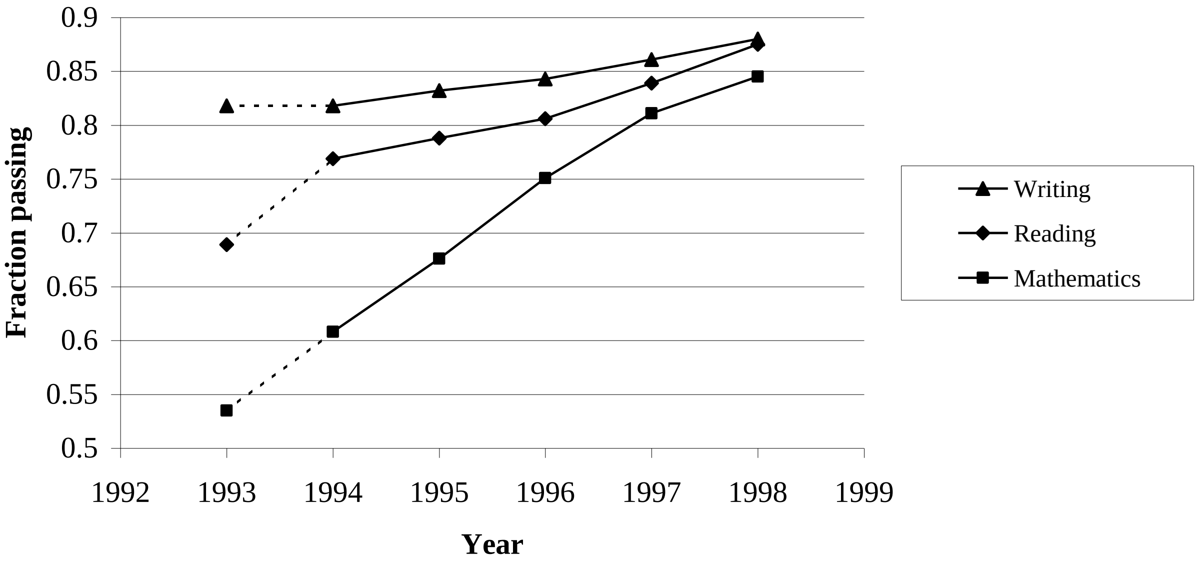


Figure 1b. Mean school mathematics TAAS pass rate by subgroup

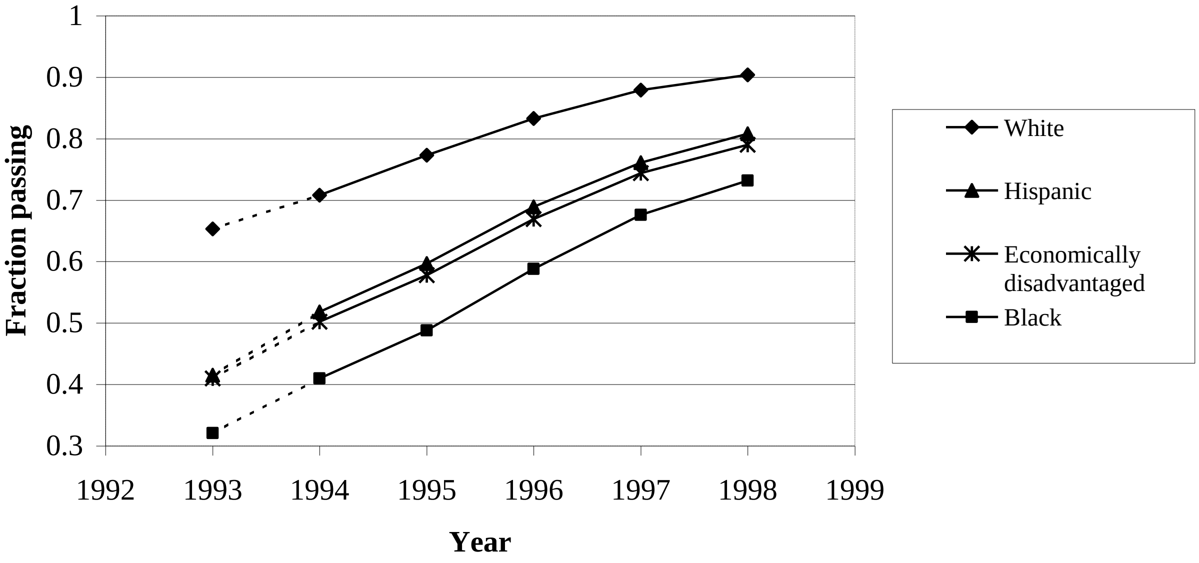


Figure 2. Percent of schools receiving each rating

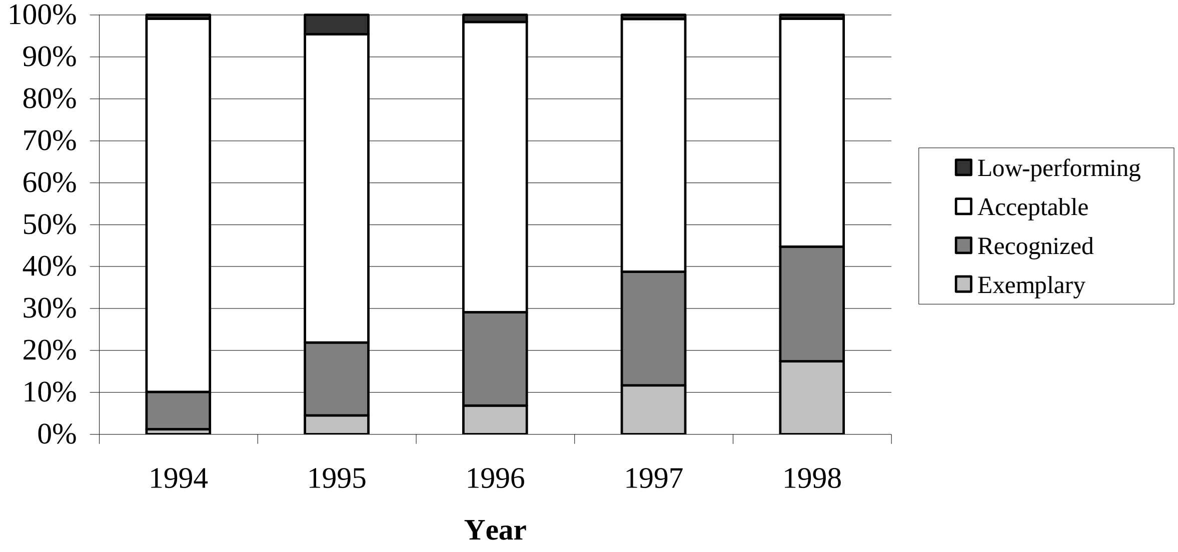
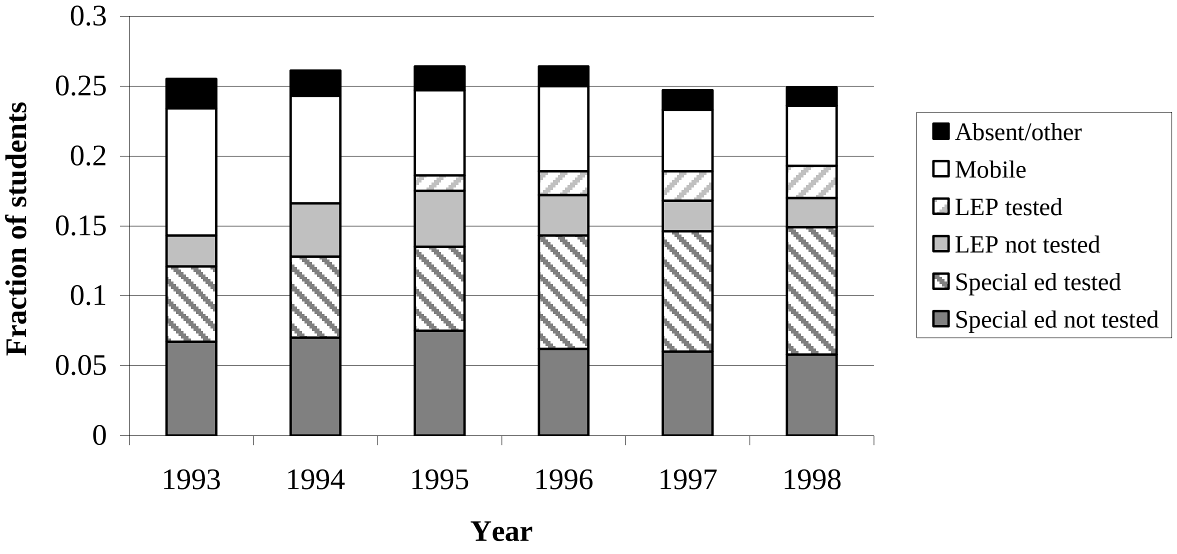
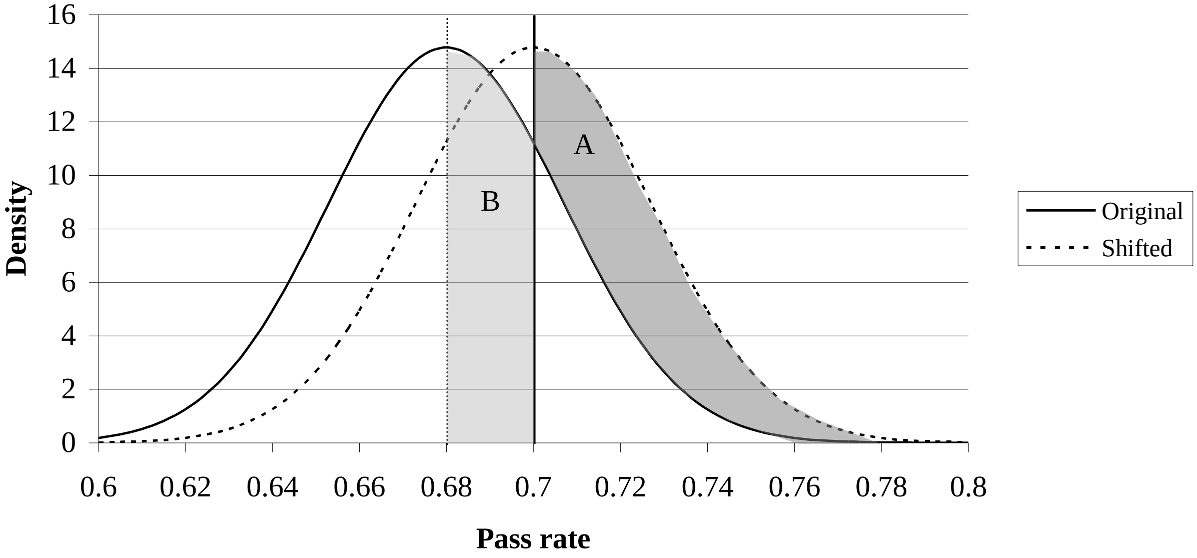


Figure 3. Mean share of students exempt by category



Notes: During this period special education students were exempt regardless of whether they took any of the exams, and LEP students were exempt unless they took the English exams. The shift in the composition of LEP exemptions toward tested students is due to the introduction of Spanish exams for grades 3-4 in 1995 and grades 3-6 in 1996. Mobile students are students who were not enrolled in the district as of October of the school year.

Figure 4a. Illustrative raw marginal benefit to increasing exemptions



Notes: The standard deviation in the illustration is set equal to 0.027—which is the average in our sample for campuses of typical size (enrollment in the range of 200 to 400 students) with simulated mean expected pass rates in the range of .65 to .75 in 1995.

Figure 4b. Illustrative comparison of incentives across regimes

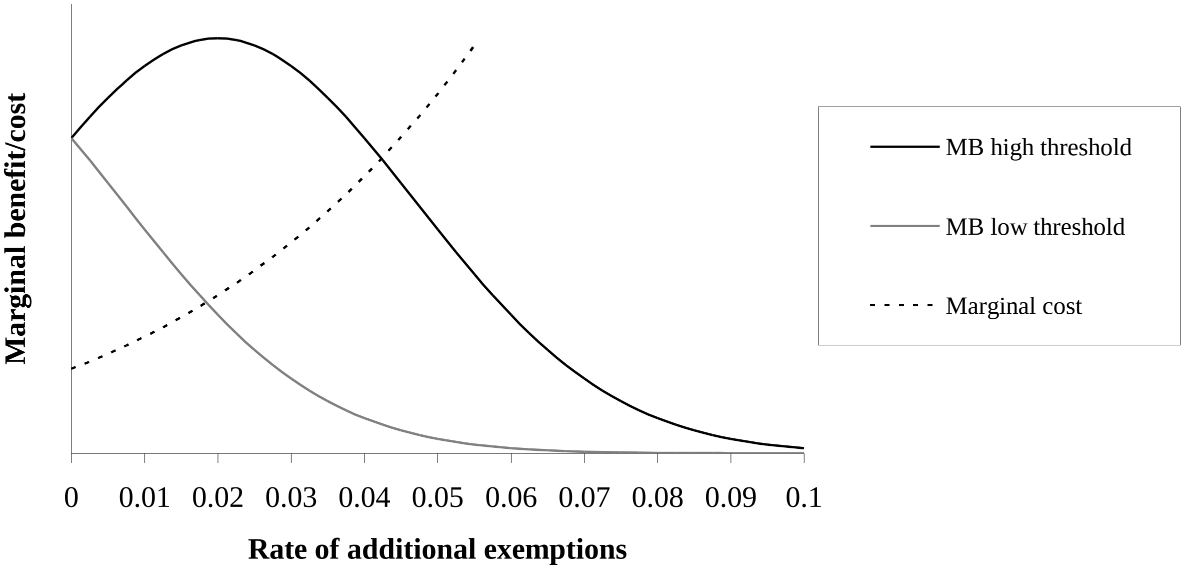


Figure 4c. Illustrative comparison of incentives across regimes

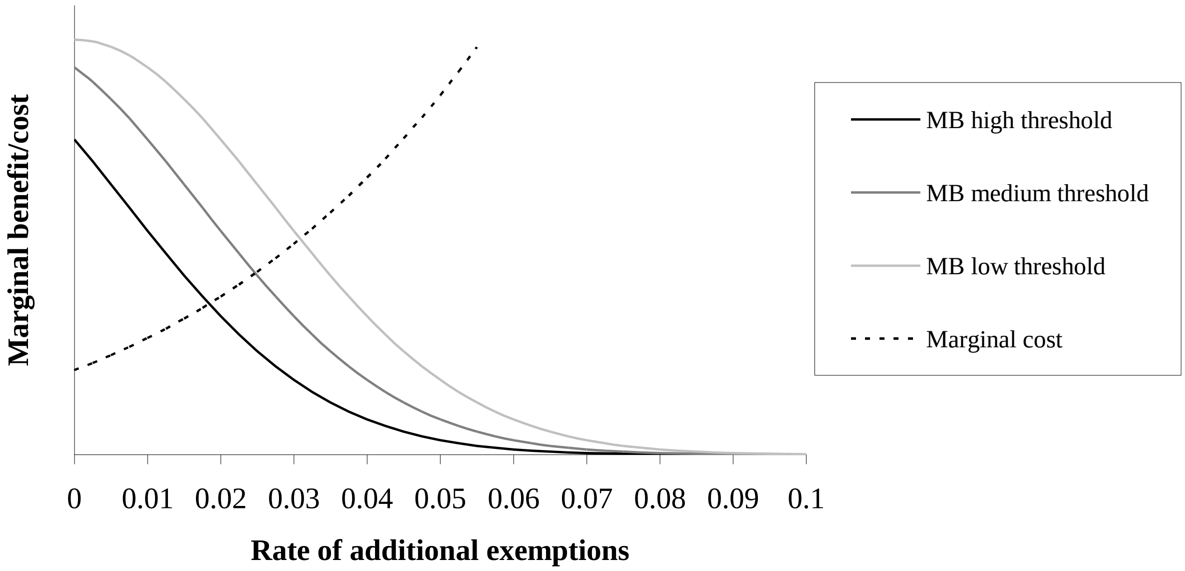


Figure 4d. Illustrative comparison of incentives across regimes

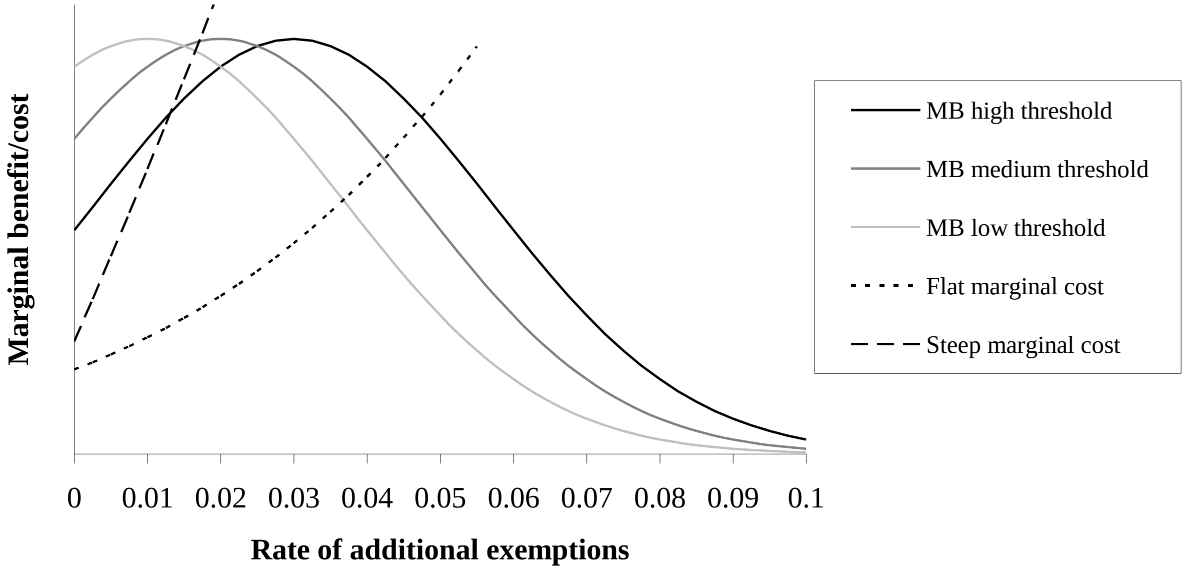


Figure 5. Illustrative increase in the "raw" marginal benefit curve

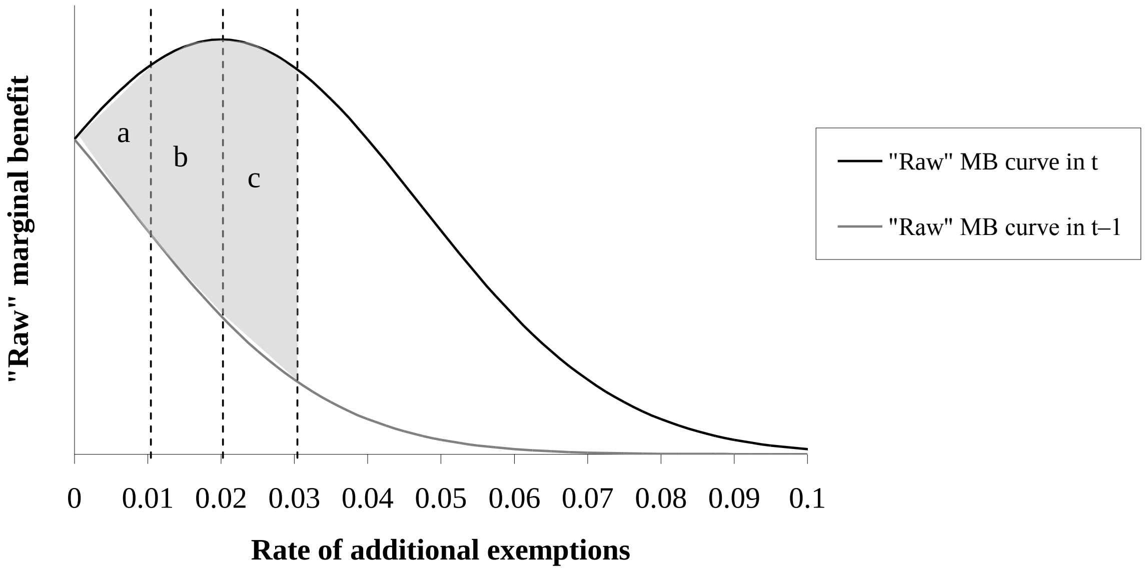


Table 1. Key provisions of the Texas accountability system

	Minimum TAAS Pass Rate			Maximum Dropout Rate			Minimum Attendance Rate		
	E	R	A	E	R	A	E	R	A
1993	90.0%	65.0%	20.0%	1.0%	3.5%	N/A	97.0%	95.0%	N/A
1994	90.0%	65.0%	25.0%	1.0%	3.5%	N/A	94.0%	94.0%	N/A
1995	90.0%	70.0%	25.0%	1.0%	3.5%	6.0%	94.0%	94.0%	N/A
1996	90.0%	70.0%	30.0%	1.0%	3.5%	6.0%	94.0%	94.0%	N/A
1997	90.0%	75.0%	35.0%	1.0%	3.5%	6.0%	94.0%	94.0%	N/A
1998	90.0%	80.0%	40.0%	1.0%	3.5%	6.0%	94.0%	94.0%	N/A

Notes: Schools are assigned one of four ratings: exemplary (E), recognized (R), acceptable (A), or low-performing. Schools are evaluated on three performance measures: current pass rates on the Spring TAAS exams for tested grades, dropout rates for grades 7-12 from the prior year, and the attendance rate for students in grades 1-12 from the prior year. All students and each separate student group (White, Hispanic, Black, and economically disadvantaged) must satisfy the test score and dropout requirements. Except for in 1993 when the requirement applied only to all tests taken combined, the pass rates apply separately to each subject area exam (mathematics, reading, and writing).

■ The dark shading indicates that there are additional requirements (such as sustained performance or required improvement) that mean a school could achieve the indicated standard and still not obtain the indicated rating.

□ The light shading indicates that there are alternative provisions (such as required improvement and single group waivers) that mean the minimum standards are not always binding.

Table 2. Summary statistics for analysis of race/ethnicity subgroups

Variable	Student subgroup		
	White	Hispanic	Black
Exemption rate	0.211	0.310	0.272
	(0.089)	(0.153)	(0.109)
Change in exemption rate	-0.003	-0.003	-0.003
	(0.072)	(0.103)	(0.104)
Subgroup's enrollment share (for tested grades)	0.466	0.381	0.225
	(0.262)	(0.277)	(0.208)
Number of students in tested grades in subgroup	184	163	92
	(174)	(192)	(112)
Fraction of students low-income in subgroup	0.309	0.683	0.643
	(0.208)	(0.213)	(0.227)
Fraction scoring 0-49 on the math exam in the prior year	0.038	0.096	0.144
	(0.051)	(0.087)	(0.115)
Fraction scoring 50-59 on the math exam in the prior year	0.064	0.121	0.153
	(0.056)	(0.082)	(0.094)
Fraction scoring 60-64 on the math exam in the prior year	0.048	0.077	0.088
	(0.040)	(0.056)	(0.064)
Fraction scoring 65-69 on the math exam in the prior year	0.068	0.096	0.103
	(0.046)	(0.059)	(0.067)
Fraction of campus-years:			
Subgroup is represented	0.957	0.948	0.724
Subgroup's pass rate is below the standard	0.233	0.605	0.392
When subgroup's pass rate is below the standard:			
White subgroup's pass rate is above the standard	0	0.669	0.741
White subgroup's pass rate is below the standard	1	0.093	0.055
White subgroup is not separately accountable	0	0.219	0.121
Hispanic subgroup's pass rate is above the standard	0.030	0	0.178
Hispanic subgroup's pass rate is below the standard	0.242	1	0.387
Hispanic subgroup is not separately accountable	0.617	0	0.366
Black subgroup's pass rate is above the standard	0.003	0.015	0
Black subgroup's pass rate is below the standard	0.093	0.250	1
Black subgroup is not separately accountable	0.565	0.402	0
Number of observations	10,552	10,449	7,983

Notes: There are 11,026 campus-years from 1995-98 that have variation across subgroups in their expected performance compared to the relevant standard. For these campus-years, at least one subgroup's expected pass rate is below the standard, and at least one other subgroup either is expected to exceed the standard or has too few students to be separately accountable. Each column in the table reports statistics for the campus subgroup indicated in the column heading. There are a total of 28,984 campus-year-subgroup observations, with the number of observations contributed by each race/ethnicity subgroup indicated in the final row. The top panel shows the mean (and standard deviation in parentheses) for the variable indicated by the row heading. The bottom panel shows the fraction of campus-years represented by alternative combinations of relative performance for each subgroup.

Table 3. Ordinary least-squares regression results for analysis of race/ethnicity subgroups

Independent variable	Dependent variable = Change in exemption rate					
	(1)	(2)	(3)	(4)	(5)	(6)
Subgroup is “under”	0.006** (0.001)	0.017** (0.003)	-0.001 (0.002)	0.002 (0.005)	-0.001 (0.002)	0.006 (0.006)
Subgroup is Hispanic and “under”	—	—	0.012** (0.004)	0.021** (0.007)	—	—
Subgroup is Black and “under”	—	—	0.013** (0.004)	0.037** (0.008)	—	—
Share low-income in subgroup × low-income subgroup is “under”	-0.003 (0.007)	0.018 (0.012)	-0.002 (0.007)	0.015 (0.012)	-0.008 (0.009)	0.002 (0.017)
Share low-income in subgroup × subgroup is “under”	—	—	—	—	0.004 (0.007)	0.008 (0.013)
Share low-income in subgroup × subgroup is “under” × low-income subgroup is “under”	—	—	—	—	0.013** (0.006)	0.019* (0.010)
Subgroup is Hispanic	-0.013** (0.002)	—	-0.017** (0.003)	—	-0.015** (0.002)	—
Subgroup is Black	-0.014** (0.003)	—	-0.019** (0.003)	—	-0.016** (0.003)	—
Share low-income in subgroup	0.006 (0.009)	-0.010 (0.027)	0.007 (0.009)	-0.008 (0.027)	0.005 (0.010)	-0.003 (0.028)
Includes campus × year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Includes campus × subgroup fixed effects	No	Yes	No	Yes	No	Yes

Notes: The sample is restricted to campus-years for which there is variation across subgroups in their expected performance compared to the relevant standard. There are 28,984 observations defined at the level of the campus, year, and race/ethnicity subgroup for the years 1995-1998. Each column presents the results from a separate ordinary least-squares regression. The dependent variable in each case is the change in the exemption rate from the prior year for the subgroup. Standard errors that are robust to unspecified correlation across observations from the same campus are shown in parentheses. In addition to the variables shown, all specifications include the share of students in the subgroup scoring in various failing ranges on the math and reading exams in the prior year (the measures included for reading are for the same ranges as for the math exam), the change in the share low-income from the prior year, the prior level and change in the subgroup’s enrollment share, and the prior level and change in the number of students in tested grades (measured by a five-part spline with cut-points defined by quintiles of the size distribution). All specifications also include campus-year fixed effects, and the specifications in the even columns add campus-subgroup fixed effects.

**Indicates significance at the 5-percent level * Indicates significance at the 10-percent level

Table 4. Summary statistics for campus-level analysis

Variable	Mean	Variable	Mean
Dependent variables		*Student characteristics*	
Fraction of students exempt	0.253	Fraction Hispanic	0.347
	(0.103)		(0.316)
	[0.461]	Fraction Black	0.135
Fraction special education exempt	0.146		(0.199)
	(0.058)	Fraction low-income	0.491
	[0.562]		(0.269)
Fraction LEP exempt	0.044	Fraction in tested grades	0.580
	(0.090)		(0.263)
	[0.274]	Enrollment in tested grades (3-8,10)	351
Fraction mobility exempt	0.049		(269)
	(0.027)	Predicted overall math pass rate	0.816
	[0.395]		(0.090)
Fraction absent/other exempt	0.014	Predicted overall reading pass rate	0.744
	(0.019)		(0.126)
	[0.402]		
raw marginal benefit		*Relevant ratings category*	
Increased from prior year	0.477	Acceptable	0.066
Decreased from prior year	0.398	Recognized	0.667
Ambiguous change	0.125	Exemplary	0.267
Magnitude conditional on	0.221	District per pupil tax base wealth in	192
an increase	(0.260)	prior year (thousands of 1998 $)	(223)

Notes: The summary statistics are based on the sample of regular schools for the fiscal years 1995 through 1998 described in the text, and includes a total of 15,657 campus-year observations. (We exclude the five percent of campus-year observations where the most relevant ratings category changes as we trace out the marginal benefits to exempting an additional three percent of students). We show the mean for the variable indicated in the row heading, with the standard deviation in parentheses. For the exemption variables, the fraction of campuses with increases from the prior year is also shown in square brackets. The predicted pass rates are based on the accountability subset in the prior year, accounting for statewide increases in pass rates between the prior and current year as described in the text. The relevant ratings category is the one we predict to be the "nearest" from either above or below given the predicted overall and subgroup pass rates.

Table 5. Estimated relationships between one-year changes in exemptions and incentives

Sample and exemption type	Dependent variable			
	Indicator for exemptions increased		Change in exemptions	
	(1)	(2)	(3)	(4)
Full sample [N=15,657]				
Total exemptions	1.55^{*}	1.27	0.16	0.08
	(0.91)	(1.22)	(0.10)	(0.14)
Special education exemptions	1.52^{*}	0.84	0.13^{*}	0.08
	(0.93)	(1.22)	(0.07)	(0.10)
LEP exemptions	0.14	-0.76	-0.06	-0.10
	(0.77)	(0.98)	(0.06)	(0.08)
Student mobility exemptions	0.02	0.14	0.01	0.04
	(0.89)	(1.20)	(0.06)	(0.08)
Absent/other exemptions	1.48^{*}	1.77	0.07^{**}	0.06
	(0.92)	(1.26)	(0.04)	(0.05)
No change in relevant ratings category from prior year [N=12,847]				
Total exemptions	3.40^{**}	3.22^{**}	0.41^{**}	0.32^{**}
	(1.02)	(1.50)	(0.11)	(0.17)
Special education exemptions	2.00^{*}	1.43	0.23^{**}	0.17
	(1.05)	(1.50)	(0.08)	(0.12)
LEP exemptions	1.15	-0.17	0.02	0.02
	(0.89)	(1.24)	(0.07)	(0.10)
Student mobility exemptions	0.40	0.47	0.07	0.10
	(0.99)	(1.47)	(0.06)	(0.10)
Absent/other exemptions	1.79^{*}	1.55	0.08^{**}	0.04
	(1.04)	(1.56)	(0.04)	(0.06)
Includes campus fixed effects	No	Yes	No	Yes

Notes: Each cell presents the coefficient (multiplied by 100) on an indicator for an increase in incentives from the prior year from a separate ordinary least-squares regression. The top panel shows results for the sample of regular schools for 1995-98 described in the notes to Table 4. The bottom panel shows the results for the sub-sample with "no change in ratings category" that excludes the 18 percent of campus-year observations where the ratings category that is relevant for calculating incentives changes from the prior year. The rows indicate the type of exemption the dependent variable is based on. The first two columns present results when the dependent variables are defined to be indicators for whether the relevant exemption rate increased from the prior year. Columns 3 and 4 present results when the dependent variables are expressed as changes in the rate from the prior year. Columns 2 and 4 also control for campus fixed effects. Standard errors (multiplied by 100) that are robust to unspecified correlation across observations from the same campus are shown in parentheses. All specifications include the prior level and change in the student demographic characteristics shown in Table 4 (with enrollment in tested grades captured by a five-part spline with cut-points defined by quintiles of the size distribution), the prior level and change in the grade distribution of students in tested grades, predicted overall math and reading pass rates, indicators for the relevant ratings category in the prior and current year, prior year per pupil tax base wealth quintile interacted with an indicator for the year, and an indicator for an ambiguous change in incentives from the prior year.

**Indicates significance at the 5-percent level * Indicates significance at the 10-percent level

Table 6. Estimated relationships between one-year changes in exemptions and incentives

Sample and dependent variable	Independent variable			
	1^{inc}	$1^{inc}\times\Delta MB$	1^{inc}	$1^{inc}\times\Delta MB$
	(1a)	(1b)	(2a)	(2b)
Full sample [N=15,657]				
Indicator for total exemptions increased	0.59	4.46^{*}	-0.74	8.94^{**}
	(1.03)	(2.37)	(1.41)	(3.14)
Change in total exemptions	0.11	0.21	-0.01	0.39
	(0.11)	(0.19)	(0.16)	(0.27)
No change in ratings category [N=12,847]				
Indicator for total exemptions increased	2.14^{*}	5.93^{**}	0.94	10.44^{**}
	(1.18)	(2.70)	(1.72)	(3.81)
Change in total exemptions	0.36^{**}	0.24	0.29	0.17
	(0.13)	(0.23)	(0.19)	(0.33)
Includes campus fixed effects	No		Yes	

Notes: The top panel shows results for the sample of regular schools for 1995-98 described in the notes to Table 4. The bottom panel shows the results for the sub-sample with "no change in ratings category" that excludes the 18 percent of campus-year observations where the ratings category that is relevant for calculating incentives changes from the prior year. The rows indicate whether the dependent variable is an indicator for an increase in exemptions or the change in exemptions. Columns 1a and 1b present estimated coefficients on an indicator for an increase in incentives from the prior year (1^{inc}) and that indicator interacted with the size of the increase ($1^{inc}\times\Delta MB$) in specifications that include the full set of control variables describes in the notes to Table 5. Columns 2a and 2b present the same results from specifications that add campus fixed effects as well. Standard errors that are robust to unspecified correlation across observations from the same campus are shown in parentheses. **All coefficients and standard errors have been multiplied by 100**.

**Indicates significance at the 5-percent level * Indicates significance at the 10-percent level