Racial Subgroup Rules
in School Accountability Systems

by

Thomas J. Kane	and	Douglas O. Staiger
UCLA		Dartmouth College
tomkane@ucla.edu		doug.staiger@dartmouth.edu

September 27, 2002
(First draft: May 31, 2002)

Preliminary draft
Please do not cite without permission

Paper prepared for the conference "Taking Account of Accountability: Assessing Politics and Policy," Kennedy School of Government, Harvard University, June 10-11, 2002. Support from the Andrew W. Mellon Foundation is gratefully acknowledged.

I. Introduction

While designing accountability systems for schools, state policymakers have been forced to confront large and longstanding differences in test performance by race and ethnicity. Some states have wielded test-based accountability as a tool with which to try to close the gaps in performance-- setting performance goals not only for students overall, but for subgroups of students defined by race and ethnicity as well. As reflected in the legislation's title, the federal No Child Left Behind Act of 2001 (NLCBA) aspires to leave no group behind, setting goals for subgroups defined by race/ethnicity, economic disadvantage, disability and English language learner status. However, as in many other areas of policy design, that which seems reasonable at first glance often has unintended consequences.

In this paper we have four goals: to describe the types of incentives that have been established, to analyze some of the perverse effects these subgroup rules have on schools, to provide preliminary evidence on the impact of such rules on student performance, and to make some suggestions regarding how such rules could be re-designed. Our bottom line is that subgroup rules are counter-productive in test-based accountability systems. Although well intentioned, subgroup rules result in fewer resources and more sanctions targeted on diverse schools simply because of their diversity, and do not appear to have any impact on the test score performance of students from minority groups.

II. Overview of Subgroup Rules

In order to encourage schools to raise the performance of all youth, seventeen states report performance separately for certain subgroups of students, including minority, low-income, and limited-English-proficient students. In many of these states, schools are held accountable only for their performance overall and do not face separate expectations for each of their subgroups. However, a growing number of states are setting performance targets for schools as well as for subgroups of students within schools.

States have used two basic strategies for incorporating racial subgroups into a school accountability system. First, some states, including Texas, have set a single performance expectation for the absolute level of performance, that applies to schools overall and to subgroups of students within schools. For example, in order to reach an "exemplary" rating, schools in Texas are expected to achieve a 90 percent proficiency rate for their school overall as well as for all subgroups, including white non-Hispanic youth, Hispanic students, African American students and economically disadvantaged students. In order to reach an "academically acceptable" rating, schools must achieve at least 55 percent proficiency rates for all subgroups of students (raised from 50 percent in 2001). Like the Texas plan, the No Child Left Behind Act of 2001 would require all states to establish a single minimum proficiency rate which would apply to all schools as well as to all subgroups of students within schools.

However, given large differences in test performance by race, states using such systems face a trade-off between setting a low standard for proficiency and accepting high failure rates for schools containing students from disadvantaged subgroups. This trade-off is more stark in more integrated states, where a large proportion of schools enroll significant numbers of minority youth. An alternative approach, adopted in California, is to set a uniform standard for the *growth* in performance and apply the standard to the school overall as well as to all subgroups in the school. One advantage of the latter approach is that it avoids the problem of large differences in baseline performance by race by focusing on changes in performance. However, focusing on annual changes in performance exacerbates other problems– such as those created by the imprecision of test score measures, since a large portion of the change in test scores from one year to the next could be expected to be due to sampling variation and other non-persistent causes.   We discuss both approaches in more detail below.

III. Holding All Subgroups to the Same Absolute Standard

The No Child Left Behind Act of 2001 (NCLBA) requires schools to achieve a minimum level of proficiency for its students overall, as well as for each subgroup in a school defined by race/ethnicity, socioeconomic disadvantage, disability status and English language learner status. The legislation allows states to create their own definition of "proficiency", based upon their own curriculum standards. However, the legislation circumscribes states' flexibility by specifying the manner in which the minimum proficiency rate for schools is to be determined.[1] Once a state defines proficiency, the minimum proficiency rate for each school and subgroup is set at the maximum of the proficiency rate of the twentieth percentile school or the proficiency rate of the lowest scoring subgroup. In states with more lenient definitions of proficiency, the minimum proficiency rate will be higher, since the proficiency rate of the 20th percentile school and for all subgroups will be higher.[2] Regardless of the initial proficiency level, the minimum proficiency level must be raised at regular intervals until it reaches one hundred percent at the end of twelve years. States which set a high standard, such that a small fraction of students achieve proficiency at the baseline, will be expected to achieve larger improvements over the next twelve years.

In most states, the minimum proficiency rate will be defined by the twentieth percentile school's proficiency rate. It will rarely be equal to the proficiency rate of the lowest scoring subgroup– simply because the twentieth percentile school's proficiency rate is likely to be higher. The reason, illustrated in more detail below, is that the racial gap in performance is quite large relative to the between-school dispersion in test scores.

The left panel of Figure 1 reports the black-white differences in 4th grade math scores on the National Assessment of Educational Progress by state in 2000 plotted against the same differences by state 8 years before in 1992. The line in Figure 1 is drawn at 45 degrees. States with points below the line have experienced a closing of the gap in test scores, while states above the line have witnessed a widening of their racial gap. Both gaps are reported in student-level standard deviations units (31.2 points in 4th grade math). There are two points worth highlighting in Figure 1. First, the black-white gap in mean math performance in 4th grade is quite large. In the year 2000, the gap ranged from .6 standard deviations in West Virginia to over 1.2 standard deviations in Michigan. Second, the racial gaps by state are also remarkably stable over time. The states with wide racial gaps in fourth grade mathematics in 1992 also tended to have wide gaps in 2000. There was some closing of the gap between 1992 and 2000. For instance, Texas and North Carolina have been identified as having had particularly rapid closing of the racial gap over time. However, any improvement has been modest relative to the size of the remaining gap.

The right panel of Figure 1 reports the gap in 4th grade math scores for Latino students in 1992 and 2000. The gaps are also large, between .4 and 1.2 standard deviations. Many of the points are above the 45 degree line, suggesting some widening of the gap in math performance between whites and Hispanics between 1992 and 2000. However, such widening may simply reflect recent immigrant flows into the United States.

Figure 2 portrays the distribution of grade 3 through 5 math scores in North Carolina at the individual level as well as at the school level, for African American students and for students overall. The dotted lines portray the distribution of test scores for individual students; the solid lines portray the distribution of test scores when aggregated up to the school level. Even though there is a difference in mean performance by race, there is a considerable amount of overlap at the individual student level. Even though the mean performance for African American students is .5 student level standard deviations below the statewide mean (.8 standard deviations below the white mean), 30 percent of individual African American students have test scores above the statewide mean.

However, as portrayed in Figure 2, moving from the level of the individual student to the level of school means greatly reduces the extent of overlap in the distributions for schools overall and for African American subgroups within schools. The distribution of school means collapses toward the overall mean, while the mean for African American subgroups within schools collapses toward the African American mean. Whereas 30 percent of individual African American students scored above the overall mean, only 2 percent of African American students were in schools where the *mean* performance of African American students exceeded the statewide mean.

The vertical line in Figure 2 portrays the mean math score for the 20th percentile school. The twentieth percentile school has a mean test score .27 standard deviations below the overall mean. However, relatively few African American students– just 12 percent-- attended schools where the mean African American student scored above this threshold. In other words, 88 percent of African American students are in schools where the mean for African American students is below the mean for the 20th percentile school. While not reported here, results are similar for Latino subgroups.[3]

The between-school variance in mean test performance is small relative to the racial gap in North Carolina. However, North Carolina is unlikely to be anomalous in this regard. Although it depends upon a number of factors such as the test being used, school size and the extent of racial integration in a state, the between-school variance in student test scores generally represents between 10 to 15 percent of the variance in student test scores. Similar findings have been reported at least since the analysis by James Coleman and his colleagues in 1966.[4] If the distribution of school mean test scores is roughly normal (as a casual inspection of Figure 2 would confirm) with a variance of .10 to .15 of the student level variance, then the 20th percentile is likely to be .27 to .33 student level standard deviations below the overall mean– much less than the typical gap in performance between whites and blacks and whites and Latinos. Therefore, although the result may vary somewhat by test and by state, given the magnitude of the racial gaps, the 20th percentile school is likely to have scores quite a bit higher than the average score for African Americans and Hispanic students.

Anticipating School Failure Rates

The definition of minimum proficiency virtually ensures that twenty percent of schools will have proficiency rates below the minimum initially.[5] However, the proportion of schools failing to meet this new definition of "adequate yearly progress" is likely to be much higher than twenty percent. The reason is that a school is defined as failing if *any* of the racial subgroups within the school fails to achieve the minimum proficiency rate. As we saw above, given the definition of minimum proficiency in the law, the vast majority of African American and Latino subgroup mean scores at the school level are likely to fall short. As a result, a vast majority of the schools containing African American or Latino subgroups are also likely to fail.

How many students are likely to be in schools with African American or Latino subgroups? The NCLBA does not define subgroup status beyond stating that subgroup means could be excluded where "the number of students in a category is insufficient to yield statistically reliable information." Such language is open to interpretation, since there is no magical sample size above which subgroup means are likely to be "statistically reliable". In this paper, we apply the definition of subgroup status used by California, requiring any of the categories above to contain at least 30 students and 15 percent of the students in a school or greater than 100 students, regardless of their percentage representation to constitute an official subgroup. This definition results in somewhat fewer schools with subgroups than the definition currently used in Texas, requiring a subgroup to contain at least 10 percent of the student body and more than 30 students or more than 50 students regardless of the percentage to count.

The proportion of schools containing an African American or Latino subgroup varies widely by state, depending upon the representation of African American and Latino youth in the resident population and the degree of integration. Table 1 reports results from the Common Core Data set for the 1999-2000 school year, to provide a rough sense of the proportion of public schools in each state likely to be affected. These data are weighted by school size. Several states, including Idaho, Tennessee and Washington did not report complete racial representation data and were dropped. The data in Table 1 are sorted by the proportion of students in a state that are black or Hispanic.

Several results in Table 1 are particularly striking. First, a majority of the public schools nationwide (54 percent) contain an African American or Latino subgroup, using the definition of subgroup status described above. Moreover, in the South and West, the percentages are generally much higher. More than 80 percent of the public schools in seven states (TX, MS, NM, CA, LA and SC) and the District of Columbia contain an African American or Latino subgroup. An additional seven states (VA, NC, NV, FL, GA, AL and AZ) contain African American or Latino subgroups in more than 60 percent of their public schools. Therefore, given the fact that a majority of the African American and Latino subgroups are likely to fail given the manner in which the minimum proficiency rates are to be calculated, a very large share of the schools in these states are likely to fail to achieve adequate yearly progress.

Second, while 92 percent of African American youth and 91 percent of Latino youth attend a school where black or Hispanic students are sufficiently numerous to constitute a separate subgroup, the proportion of white students likely to be affected varies widely. For example, New York and Alabama have similar percentages of African American and Hispanic students in their public schools, but 20 percent of white youth in New York and 50 percent of the white students in Alabama attend schools with an African American or Latino subgroup. North Carolina and Illinois have similar percentages of black or Latino youth overall, yet white students in North Carolina are nearly *three times* as likely as white students in Illinois to attend schools containing an African American or Latino subgroup– 62 versus 23 percent. The more integrated a state's schools are, the higher proportion of their schools are likely to be affected by the NCLBA.

Minimum Proficiency Rates in Texas

The use of the same minimum proficiency rate for schools as well as for subgroups of students within schools is similar in spirit to the accountability system in Texas. However, many more schools are likely to fail under the NCLBA requirements than the 2 percent of schools rated as "academically unacceptable" in Texas in 2000[6]. The reason is that the minimum proficiency rate required by the NCLBA will be much higher than the minimum used in that state in the past. Figure 3 portrays the distribution of mean proficiency rates for schools overall, for African Americans and for Hispanic students grouped by school in Texas in the 1999-2000 academic year. Texas used a fairly lenient definition of proficiency, with a median proficiency rate in math of 89.5 percent in 2000. As a result, the 20th percentile school has a proficiency rate in math of slightly higher than 80 percent. This is 30 percentage points higher than the minimum proficiency rate the state used in the 1999-2000 academic year.[7]

The Importance of the Definition of Subgroup Status

Under the NCLBA, the stakes for schools are potentially quite high (although it remains to be seen how serious these consequences will be in practice). A school failing to achieve adequate yearly progress for two consecutive years– whether because of its overall mean or because of any of its subgroup means-- will be required to submit a "school improvement plan", and students in the school must be given the choice of attending another school in the district (if that schools is not also failing), with the district bearing the transportation expense. A school failing for three consecutive years will be required to offer vouchers to low-income students to be used for supplemental educational services such as after-school tutoring programs. A school failing for four consecutive years must institute one of several "corrective" actions, such as implementing a new school curriculum. A school failing for five years is subject to "restructuring", and must either be converted to a charter school, turned over to a private operator, or have "most or all" of its staff replaced.

Therefore, the definition of subgroup status is likely to be an important determinant of success or failure. For instance, in the academic year 1999-2000 in Texas, a racial or ethnic subgroup was required to represent at least 10 percent of the student body and 30 students or at least 200 students to count as a separate subgroup.[8] In order to achieve "exemplary" status, a school in Texas was required to have a 90 percent proficiency rate in reading, writing and mathematics for the school overall and for each subgroup. Given the racial differences in proficiency rates in Texas, relatively few of those schools with an African American or Latino subgroup were able to achieve an exemplary rating.

The impact of the subgroup size threshold is seen clearly in Figure 4, which portrays the proportion of schools achieving an exemplary rating, by the percentage of their students who were Latino. The graph is limited to the schools that had between 300 and 2000 students, where the percentage will solely determine subgroup status. The sample was also limited to those schools that did not also have an African American subgroup. Between 40 and 80 percent of such schools with fewer than 10 percent of their students Latino achieved exemplary status, whereas only 10 to 20 percent of the schools with more than 10 percent of students Latino achieved exemplary status. Moreover, the discontinuity is striking right at the 10 percent threshold: 42 percent of schools with 9 percent of students Latino were rated exemplary, while less than 20 percent of the schools with 10 percent of students Latino were rated exemplary. Therefore, given the large racial differences in performance, the designation of minimum size requirements for subgroups of students will largely determine the success or failure of schools near the thresholds.

IV. Requiring Improvements for All Groups

California, like many of the largest states, rewards schools that demonstrate improvement in student performance compared to the prior year. In 2001, California provided over $570 million in aid and teacher bonuses to schools whose improvement in test scores exceeded an annual growth target. All "numerically significant" racial and ethnic subgroups were also required to exceed their growth target in order for the school to receive an award. (The state also provided $100 million in teacher bonuses to low-scoring schools in 1999 that achieved the largest improvements by 2000, while also meeting their subgroup targets.) By focusing on improvement in test scores rather than the absolute level of performance for each subgroup, the California approach is intended to level the playing field for schools serving different student populations. However, the imprecision of changes in test scores exacerbates other problems, as we discuss below.

Each year, California calculates a score (called the Academic Performance Index, or API) for each school and student subgroup. The API score is a weighted average of the proportion of all students in grades 3 and up scoring in each quintile of the national distribution on the reading, math, language and spelling sections of the Stanford 9 test. (The weights given to each quintile were 200, 500, 700, 875 and 1000, with an average score in 2000 of about 620.) The annual growth target for each school and subgroup is 5 percent of the difference between their initial API score and the statewide goal of 800. If a school or subgroup started out over 800, they were simply expected to keep their scores above 800. Schools that met their targeted improvements in performance between 1999 and 2000 received $63 per student funding from the Governor's Performance Award program. In addition, $591 per full-time equivalent teacher was awarded to both the school and teacher (for a total of about $59 per student) through the School Site Employee Bonus program.

In order to win these awards, a school must achieve a minimum improvement in performance at the school level, but also for each "numerically significant" racial or ethnic subgroup within the school. In order to be numerically significant, a group must represent at least 15 percent of the student body and contain more than 30 students, or represent more than 100 students regardless of their percentage. There are 8 different groups which could qualify as "numerically significant," depending upon the number of students in each group in a school: African American, American Indian (or Alaska Native), Asian, Filipino, Hispanic, Pacific Islander, White non-Hispanic or "socioeconomically disadvantaged" students.[9]

By focusing on changes in performance rather than the level of performance, California avoids the problems in the NCLBA and Texas systems caused by the lower level of performance in African American and Latino subgroups. In contrast to the difference across these groups in the distribution of average test scores (see Figures 2 and 3), the distribution of test-score growth is fairly similar across the groups (although subgroup performance is more variable for reasons we discuss below).

However, holding schools accountable for changes in subgroup performance introduces another important bias: Annual changes in test scores can be very noisy. The imprecision of test score measures arises from two sources. The first is sampling variation, which is a particularly striking problem in elementary schools. With the average elementary school containing only 68 students per grade level nationally, the amount of variation due to the idiosyncracies of the particular sample of students being tested each year is often large relative to the total amount of variation observed between schools. A second source of imprecision arises from one-time factors that are not sensitive to the size of the sample: a dog barking in the playground on the day of the test, a severe flu season, one particularly disruptive student in a class or favorable "chemistry" between a group of students and their teacher. Both small samples and other one-time factors can add considerable volatility to the change in average API scores, particularly for subgroups with relatively small numbers of students. In previous work, we estimate that between 50 and 80 percent of the variation in annual changes in test score measures is due to one of these two sources of non-persistent variation in test scores.[10]

The importance of sampling variation in the change in average API scores for a school or subgroup is immediately apparent in Figure 5. For all elementary schools in California, we plot the difference between API growth and target growth (between 1999 and 2000) against the number of students tested for all students, African American subgroups, and Latino subgroups. Points above the horizontal line at zero in each plot are those schools or subgroups for which API growth exceeded target growth. Figure 5 illustrates two facts which are important in the discussion of volatility. First, although the average small school exceeded its growth target by a similar amount as the average large school, small school performance was much more variable because of the noise in API measures based on small number of students. As a result, both small schools and small subgroups are more likely to have API growth that is below target. A second important fact from Figure 5 is that the distribution of performance for small subgroups is similar to that for small schools – but because subgroups tend to test a smaller number of students their performance is more volatile and more subgroups fail to achieve their growth target. Thus, for purely statistical reasons, subgroups may be less likely to pass the hurdle for financial awards in California.

Because of the importance of sampling variation in the change in average API scores, many schools will appear to excel in one subgroup but not another. But this is not necessarily the result of disparate improvement -- sampling variation would generate this pattern since fluctuations in one group would be expected to be largely independent of fluctuations in other groups. In fact, there is only a weak correlation in the magnitude of improvements for white and minority subgroups. Moreover, schools are about as likely to achieve the target for their minority subgroup but fail for the white subgroup as the other way around. Figure 6 illustrates this point using a Venn diagram for schools that had all three subgroups. The probability of exceeding their growth target was about equal for white (83 percent), African American (87 percent), and Latino (90 percent) subgroups, but only 69 percent of the schools with all three groups exceeded the target for all three groups simultaneously. The probability of exceeding the growth target for any one subgroup but not the other two was similar for whites (2 percent), African Americans (1 percent) and Latinos (3 percent). Moreover, eleven percent of schools exceeded their growth targets for African Americans and Latinos but failed for whites, suggesting that the subgroup rule is as likely to be binding on white subgroups as on minority subgroups.

When changes in API scores are this noisy, there will be a considerable amount of chance involved in whether a school or subgroup exceeds its growth target in a given year. As a result, California's subgroup rules are analogous to a system that makes every school flip a coin once for each subgroup, and then gives awards only to schools that get a "heads" on *every* flip. Schools with more subgroups must flip the coin more times and, therefore, are put at a purely statistical disadvantage relative to schools with fewer subgroups.

This statistical disadvantage is clearly seen in Table 2, which reports the proportion of California elementary school's winning their Governor's Performance Award by school size quintile and number of numerically significant subgroups in each school. Among the smallest quintile of elementary schools, racially heterogeneous schools were almost half as likely to win a Governor's Performance award as racially homogeneous schools: 47 percent of schools with 4 or more subgroups won a Governor's Performance Award as opposed to 82 percent of similarly sized schools with only one numerically significant group. This is particularly ironic given that the more integrated schools had slightly larger overall growth in performance between 1999 and 2000 (36.0 points versus 33.4 points). The statistical bias against racially heterogeneous schools is also apparent among larger schools, but somewhat less pronounced because subgroups in these schools are larger in size and, as a result, their scores are less volatile.

Because minority youth are more likely to attend heterogeneous schools than white non-Hispanic youth, the rules have the unintended effect of putting the average school enrolling minority students at a statistical disadvantage in the pursuit of award money. In California, nearly 30 percent of white students attend a racially homogenous school with only one subgroup, compared to about 5 percent of African Americans and Latinos. In contrast, most Latinos attend schools with 2 (47 percent) or 3 (41 percent) subgroups, while most African Americans attend schools with 3 (47 percent) or more (21 percent) subgroups. Based only on the number of subgroups in their schools, this makes minority students less likely to be in schools that win awards in California. For example, multiplying the proportion of white students in each type of school (1, 2, 3 or 4+ subgroups) by the probability that each type of school wins an award (from the last row of Table 2) yields an estimate that 76.5 percent of white students would be in an award winning school. In contrast, if white students attended schools with multiple subgroups at the same rates as African Americans, only 71.7 percent would be in an award winning schools. Thus, 5 percentage points of the difference in rates of award-winning for schools attended by African American and white, non-Hispanic youth is solely due to the statistical bias against schools with subgroups. A similar calculation suggests that Latinos are 2.5 percent less likely to be in an award winning school because of the subgroup bias. The dollar value of these awards was approximately $124 per student. Therefore, a rough estimate would suggest that the subgroup rules in California had the effect of reducing the average award to schools attended by African American and Latino youth by roughly $3 to $6 per student, for a total of over $6 million per year.[11]

V. Impact of Subgroup Rules on Minority Achievement

Despite the difficulties discussed above, racial subgroup rules may be worthwhile if they are effective in forcing schools to focus on the academic achievement of minority youth. Comparisons of states that do and do not use subgroup rules are inconclusive. For example, Texas closed the racial gap in the NAEP considerably between 1992 and 2000, but so did North Carolina (a state that does not use racial subgroup rules). In this section, we compare the performance of minority students in schools where they are sufficiently numerous to count as a separate subgroup to the performance of minority youth in schools just below the cut-off for numerical significance. To the extent that schools just below the cut-off are not affected by the subgroup rules, we can use this comparison to evaluate the impact of the subgroup rules on the test performance of minority students. This comparison is not perfect: a school's subgroup status may change year to year; and schools near the cut-off do not know for certain if they qualify for subgroup status until after test scores are reported. As a result, the incentives to focus on minority group achievement may not be so different for schools just below and above the cutoff, but rather all schools near the cut-off may face intermediate incentives. Nevertheless, if subgroup rules were effective we would expect to see rising test performance for minority youth as we moved from schools below the cut-off to schools above the cut-off.

In Texas, between 1994 and 2000, a racial subgroup did not count separately in the accountability system unless the group contained 10 percent of the students in a school and 30 students or more than 200 students regardless of their percentage[12]. Therefore, subgroup status depended upon two dimensions– the percentage of all students in the group and the absolute number. To simplify the analysis and limit the determinants of subgroup status to one dimension, we focused on schools with 300 to 2000 students, where any group that contained more than 10 percent of the students would have counted as a separate subgroup and none of those subgroups representing less than 10 percent of the student body would have contained more than 200 students.

In California, the minimum size requirements for subgroup status are somewhat different. As defined by the Public Schools Accountability Act of 1999, a racial subgroup is not "numerically significant" unless the group represents 15 percent of the student body and 30 students or 100 students, regardless of their percentage. As we did in Texas, to simplify the analysis, we limited the analysis to those in schools with 200 to 667 students where any group satisfying the 15 percent threshold would have counted separately and any group with less than 15 percent would not have satisfied the 100 student minimum size for separate subgroup status.

Although the minimum thresholds for numerical significance are necessarily arbitrary, such arbitrariness is fortuitous from the point of view of the evaluation, since we would not expect those schools immediately above or immediately below the thresholds to be systematically different. There may be one exception, however. Because the rules are not a secret, the subgroup rules provide an incentive to schools near the thresholds– particularly those with low-scoring minority youth-- to reclassify students by race or to ensure that certain students are not present on the day of testing. If either of these practices were common, we would expect to see an unusual number of schools with minority youth just below the thresholds for numerical significance. While we saw some evidence of "clumping" of schools in Texas with percentages of African American youth just below the 10 percent threshold, the distribution of percentage Latino was quite smooth above and below the 10 percent threshold. We did not see any evidence of "clumping" in California.

Figure 7 reports math, reading and writing proficiency rates for Latino youth in schools with between 1 and 30 percent Latino youth. The sample of schools was limited to those schools who did not also have a African American subgroup, since the performance of Latino students would have been less decisive in determining a schools' status. We calculated the mean proficiency rate separately for all schools with a given percentage Latino. In other words, we calculated the proficiency rate separately for schools with "j" percent Latino students, where j ranged from 0 through 100. The size of the symbols in the graph reflects the number of students in schools with that percentage of Latino students. We also included the regression line that would have been fit by running separate regressions of proficiency on the percentage of Latino youth for those points below 10 percent and for those points at 10 percent and above. As might be expected, the proficiency rates do decline somewhat for Latino youth in schools where Latino youth represent a larger share of the student body – since the percentage Latino is probably related to the socioeconomic status of the students attending these schools. However, there is no evidence of any rise in proficiency for Latino youth in schools immediately above the threshold for numerical significance. In other words, it does not appear that the performance of Latinos is any better in schools where they constitute 10 percent of the student body and, therefore, do count as a separate subgroup than in schools where they represent 9 percent of the student body and, therefore, do not count as a separate subgroup.

We performed a similar exercise for African American and Latino youth in California in 2001. Because the state does not publish an API index for subgroups that are not numerically significant, we obtained Stanford 9 scaled scores for all subgroups of students consisting of at least 10 students. (These scores include the scores of some students who are excluded from the CA accountability system– for instance, students who are in the district for less than a year do not count in the API score, but were included in the scaled scores we use.) We calculated the mean math and reading scores for African American and Latino youth in schools where they represented “j” percent of the student body in 2000, where “j” ranged from 0 through 100. The top two panels of Figure 8 report mean test scores by single percentage point for schools with 5 to 30 percent African American students. (Recall that the minimum percentage required for “numerical significance” was 15 percent in California.) As in Texas, we see no evidence of a rise in performance around the subgroup threshold as would have been expected if the subgroup rules were forcing schools to focus on the performance of African American youth. The bottom two panels of Figure 8 report similar mean scores for Latino subgroups, with no apparent discontinuity at 15 percent. (We have also analyzed the data for the year 2000 with similar results.)

In Texas, the strength of the incentive is clouded somewhat by the fact that schools near the threshold may not know whether Latino students will count as a separate subgroup until after the tests are taken. Schools with less than 10 percentage points minority student enrollment in recent years do face weaker incentives to focus on minority youth performance than those with 10 or more percent minority enrollment, but the difference at the 10 percent threshold may not be dramatic. In California, however, schools know in the fall what proportion of the students taking the test the prior spring were in each racial and ethnic group. In Figure 8, none of those schools with less than 15 percent of the in a given minority group in 2000 will be held accountable for that group’s performance separately in 2001.

The failure to find an impact of the subgroup rules on minority performance is not necessarily evidence that test-based incentives are ineffective in general – only that the racial subgroup rules are not having their intended impact. Although the empirical literature is still developing in this area, there is some evidence that test-based accountability systems do improve test performance overall, although it is not clear whether such test score gains are achieved with broad learning, teaching to the test, or outright cheating.[13] Therefore, holding schools accountable for student test scores may well encourage schools to focus on the performance of all students in a school. However, there is no evidence that holding schools separately accountable for the test scores of minority groups encourages schools to focus more heavily on minority youth performance. It may simply be difficult for schools and teachers to single out one group of students and target their responses by race.

VI. Conclusion

Despite some closing of the gap in performance between whites and blacks between the mid-Seventies and the late Eighties, such gaps remain quite large.[14] Therefore, raising the academic achievement of minority groups with poor test score performance is an important goal. However, our analysis suggests that the use of subgroup targets in school accountability programs is not the answer. In current accountability systems, subgroup targets cause large numbers of schools to fail (as in the NCLBA), arbitrarily single out schools with large minority subgroups for sanctions and exclude them from awards (as in Texas), or statistically disadvantage diverse schools that are more likely to be attended by minority students (as in California). Moreover, while the costs of the subgroup targets are clear, the benefits are not. Although these targets are meant to encourage schools to focus more on the achievement of minority youth, we find no association between the application of subgroup targets and test-score performance among minority youth. Thus, taken together, the evidence suggests that the use of subgroup targets is counter-productive in test-based accountability systems.

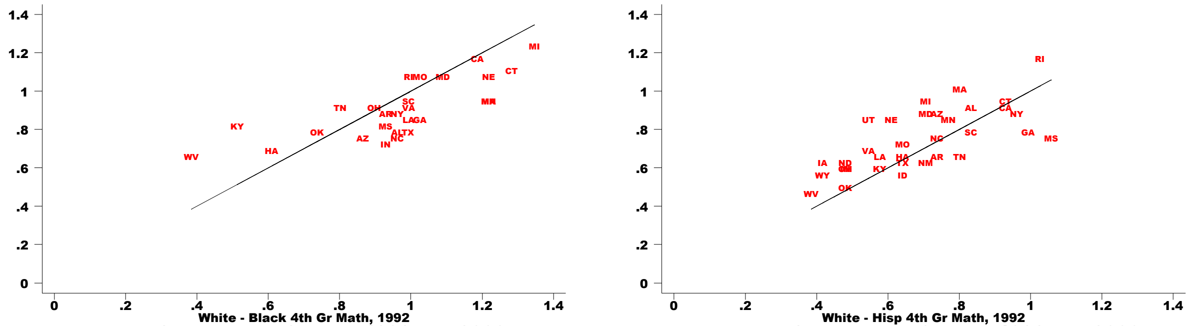
School accountability systems can certainly reduce some of the unintended consequences of subgroup targets by tinkering with the details of their application. For instance, the inequity in California's system could be lessened if the awards were made proportional to the percentage of students in subgroups achieving their targets (for instance, providing 75 percent of the award money if a school achieve's its target for the subgroups comprising 75 percent of their students), rather than requiring schools to achieve their targets for all subgroups to receive any award. Similarly, the NCLBA allows states to determine the minimum size threshold for subgroups and to fail only those schools whose subgroup performance lies significantly (in a statistical sense) below their target. This gives states some discretion in determining how binding the subgroup rules will be in practice (although it remains to be seen how large a minimum subgroup size the U.S. Department of Education will be willing to approve).

One might argue that the consequences for those schools more likely to fail as a result of subgroup rules are not serious enough to warrant tinkering with the subgroup rules. For example, while California's subgroup rules put schools attended by minority students at a statistical disadvantage in winning award money, we do not find that this had any obvious detrimental impact on minority achievement in schools with subgroups. Similarly, while the NCLBA may result in 50 to 80 percent of schools failing in many states, the consequences for failing schools (at least in the short term) – such as creating a school improvement plan or providing students with school choice – may not be very serious in practice. Moreover, until schools face restructuring after 5 years of failing to meet adequate yearly progress expectations, states will have the flexibility to apply different corrective actions in schools that fail because of a single subgroup than in schools that fail for all of their subgroups.

On the other hand, the schools in various states will face very different burdens imposed by the NCLBA, simply because of the subgroup rules. In the absence of the subgroup rules, an equal proportion– roughly 20 percent– of the schools in each state would fail to achieve adequate yearly progress in the initial years, given the way minimum proficiency is defined at the 20th percentile school. In other words, regardless of how they define proficiency, roughly 20 percent of the schools in a state with few minority students, such as New Hampshire, will fail to achieve adequate year progress in the first years-- somewhat less if schools succeed in raising performance. However, also without regard to how they define proficiency, the failure rate is likely to be two to four times higher in states in the South and West with large minority populations, because of the subgroup rules. Indeed, the single most important determinant of the difference in failure rates between states is likely to be the racial composition of their schools. While it is true that submitting a school improvement plan or being required to offer vouchers for supplemental educational services are not overwhelmingly onerous requirements, those requirements will be imposed at very different rates in the various states simply because of the racial composition of various states' schools. Moreover, the consequences will become even more severe in five when schools enter restructuring.

Although there may be room to soften the implications somewhat, identifying ways to tinker with subgroup rules misses the point: Since subgroup targets do not appear to be an effective way to improve the test scores of minority youth in the first place, one could simply eliminate such targets altogether. The fact that North Carolina (which reports subgroup results but does not set subgroup targets) experienced a narrowing in the racial test-score gap as test scores rose among all students, suggests that an explicit focus on racial and ethnic subgroups may be unnecessary. Test-based accountability systems are intended to shine a harsh light on low-performing schools and raise the stakes for improving student performance. Unfortunately, if a large share of schools are failing to achieve the new standards because of the racial subgroup rules, the law may simply make it easier for the lowest-performing schools to be lost in the crowd.

Figure 1



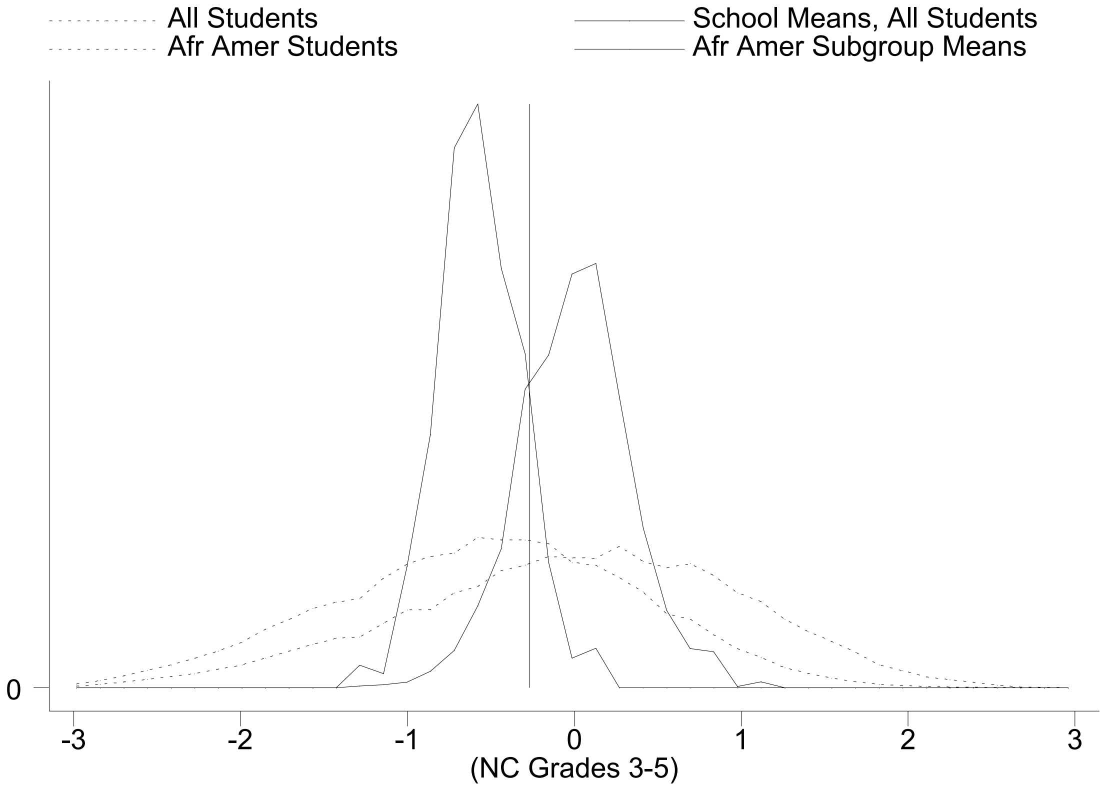
White-Black Gap in NAEP Scores, 1992 and 2000

White-Hispanic Gap in NAEP Scores, 1992 and 2000

Gaps reported in S.D. Units

cial and Ethnic Gaps in NAEP 4th Gr Math Scores, 1992-2(

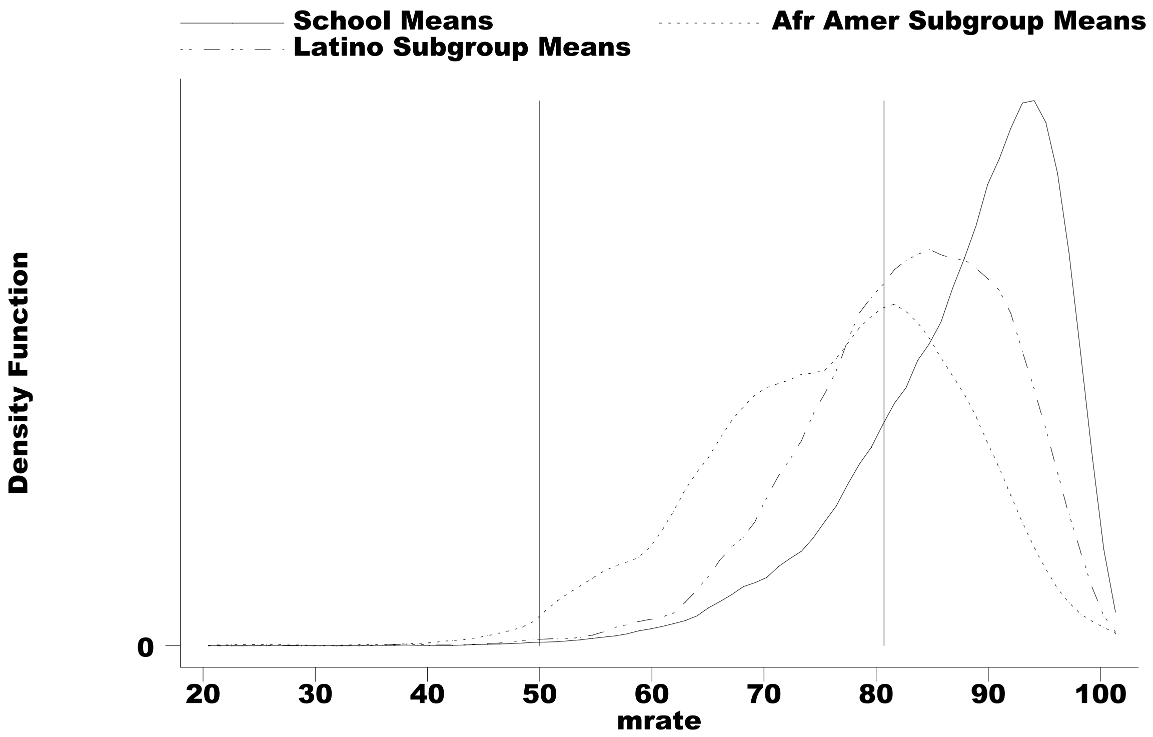
Figure 2.



Distribution of Math Scores at the School and Student Level

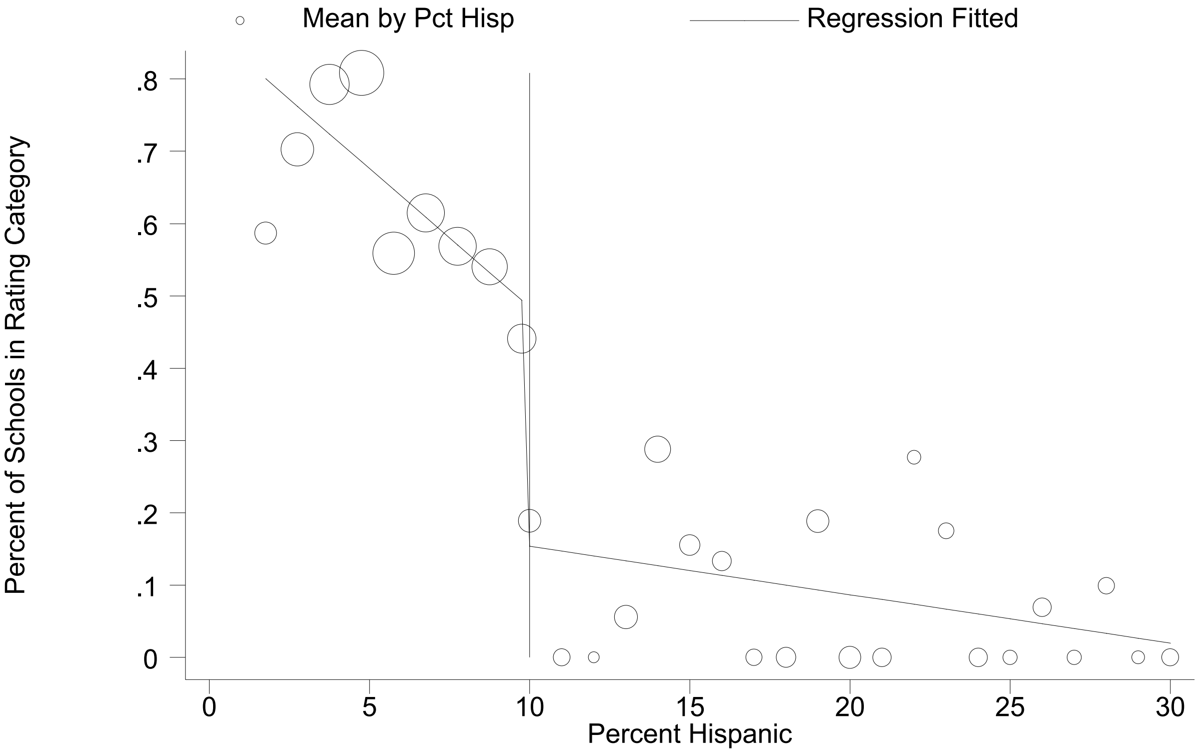
Note: The vertical line corresponds with the 20th percentile from the distribution of school means.

Figure 3.



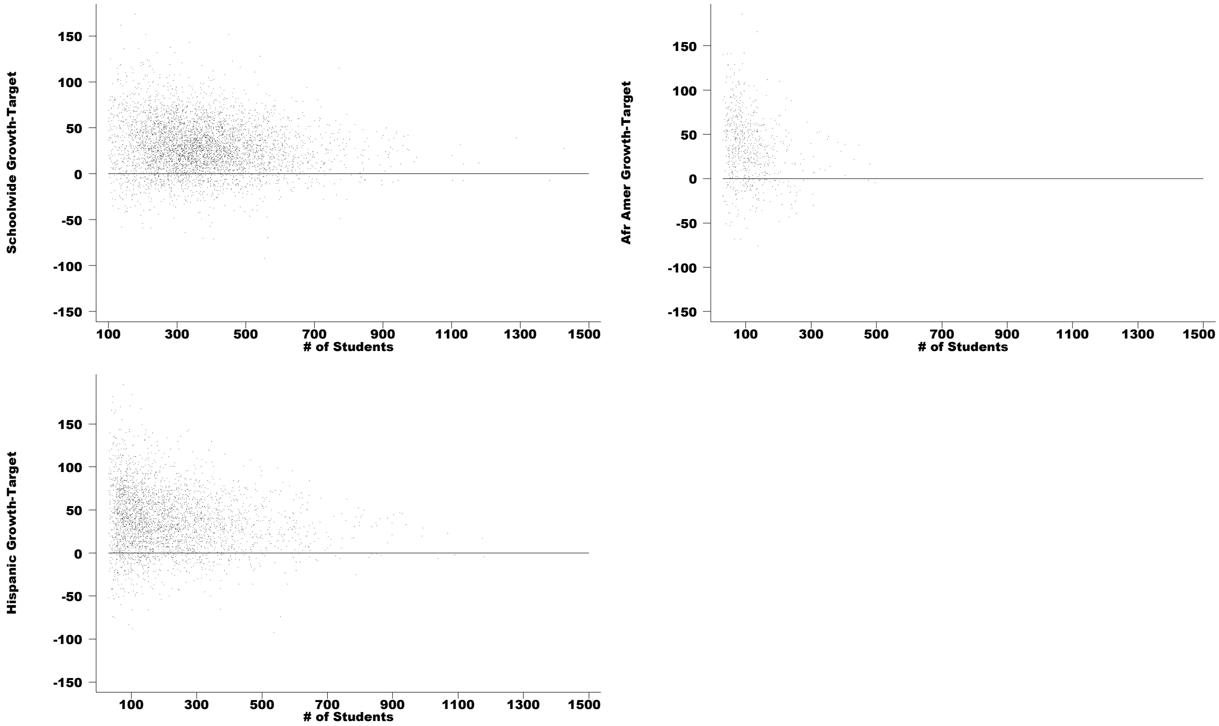
istrib of Math Proficiency for Schools and Subgroups in T

Figure 4



Exemplary School Ratings and Percent Hispanic in Texas

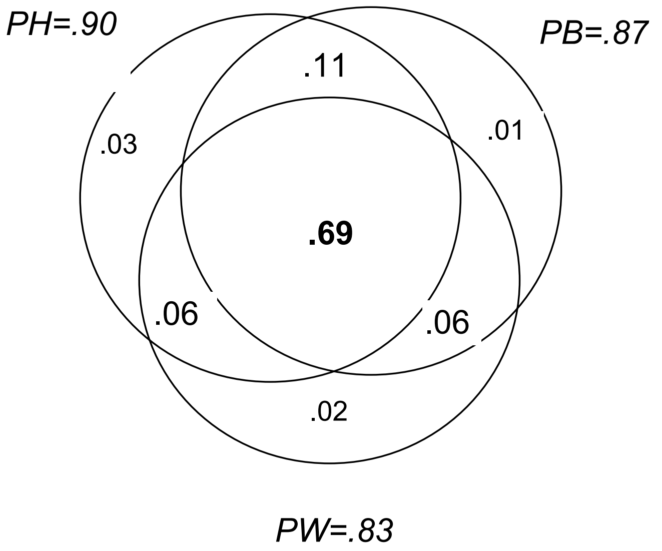
Figure 5



Distribution of Growth in Excess of Target by School Size

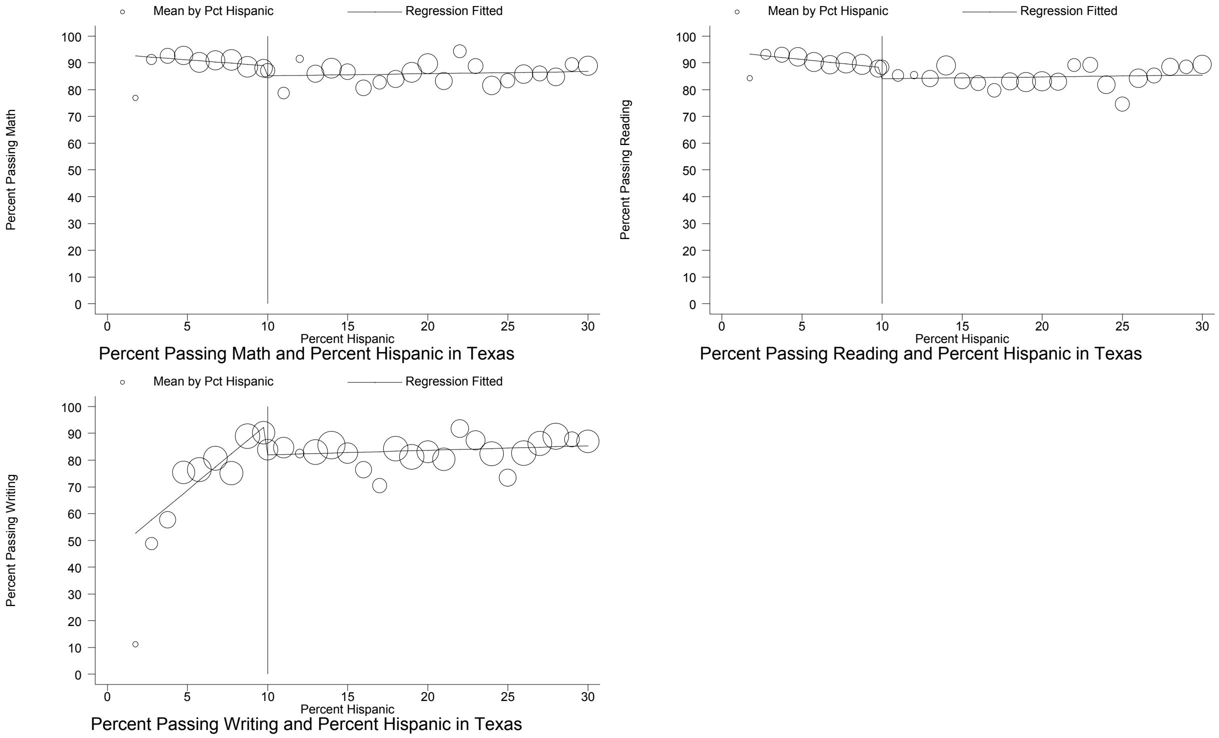
Figure 6

Probability of Achieving Expected Growth for Schools with Black, Latino and White



Note: Growth between 2000 and 2001 for elementary schools in CA with an African American, Latino and White subgroup.

Figure 7.



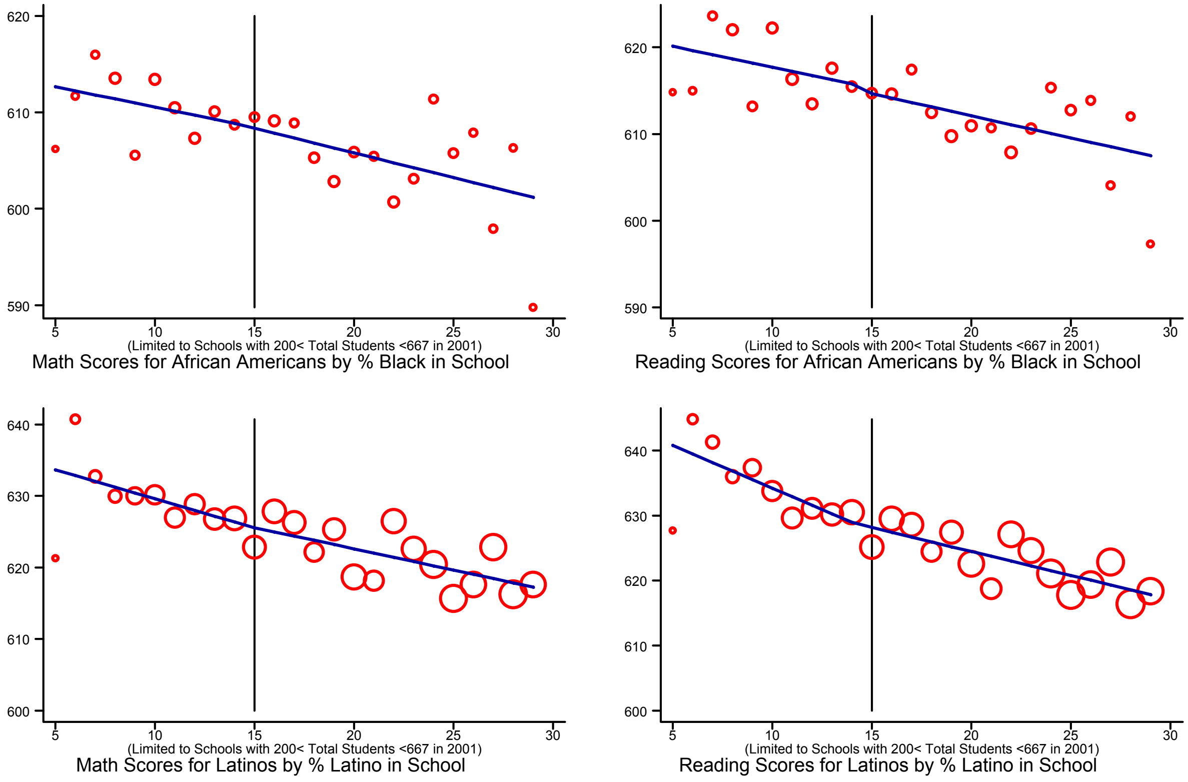
Percent Passing Math and Percent Hispanic in Texas

Percent Passing Reading and Percent Hispanic in Texas

Percent Passing Writing and Percent Hispanic in Texas

Hispanic Test Performance and Percent Hispanic in Texas

Figure 8



Math Scores for African Americans by % Black in School

Reading Scores for African Americans by % Black in School

Math Scores for Latinos by % Latino in School

Reading Scores for Latinos by % Latino in School

Table 1. Proportion of Students in Public Schools with an African American or Latino Subgroup by State

State	% in State Black or Latino	Percent in Schools with a Black or Latino Subgroup			
		Total	Blacks	Latinos	Whites
Total	34	54	92	91	33
DC	94	97	100	97	49
TX	54	80	94	96	61
MS	54	86	98	84	71
NM	53	89	96	98	87
CA	51	81	94	96	63
LA	50	80	97	86	63
SC	44	86	97	89	76
FL	43	78	95	94	66
GA	42	73	96	82	56
MD	40	63	94	85	41
AL	40	68	95	61	50
NY	38	50	93	92	20
AZ	38	65	85	92	50
DE	37	90	96	97	86
IL	36	48	94	88	23
NC	35	73	95	84	62
NV	34	72	91	90	62
VA	32	63	92	77	50
NJ	32	45	89	86	24
AR	27	48	94	63	32
CO	27	49	83	82	35
CT	26	37	87	85	18
MI	23	29	91	57	11
RI	20	29	75	87	13
MA	19	29	77	83	15
PA	19	27	87	78	12
MO	19	31	88	45	18
OH	18	27	89	59	13
KS	18	27	74	70	17
OK	16	26	76	58	18
IN	15	29	90	61	19
NE	15	27	83	69	18
WI	15	19	86	60	9
OR	12	20	51	54	15
KY	11	26	80	54	19
MN	10	17	72	46	8
UT	10	18	46	57	13
AK	9	13	42	25	9
WY	8	8	20	26	7
IA	7	11	55	44	8
HI	7	4	37	6	10
WV	5	5	37	6	3
NH	2	1	6	17	1
ND	2	1	0	10	1
MT	2	0	2	3	0
SD	2	0	0	0	0
VT	2	0	0	0	0
ME	2	0	5	2	0

Note: Based upon author's tabulation of the Common Core Data for 1999-2000 for public schools, grades 3 thro

Table 2.
Proportion of California Elementary Schools Winning Governor's Performance Awards by School Size and Number of Numerically Significant Subgroups

Proportion Winning
(Average Growth in API 1999-2000)
[# of Schools in Category]

	# of Numerically Significant Subgroups				Total:
	1	2	3	4+	
Smallest	.824 *(33.4)* [204]	.729 *(45.6)* [343]	.587 *(42.2)* [349]	.471 *(36.0)* [51]	.683 *(41.2)* [947]
2nd	.886 *(29.9)* [158]	.769 *(42.6)* [337]	.690 *(42.2)* [358]	.670 *(43.9)* [94]	.749 *(40.5)* [947]
3rd	.853 *(26.8)* [156]	.795 *(36.3)* [308]	.708 *(38.9)* [390]	.667 *(44.6)* [93]	.756 *(36.6)* [947]
4th	.903 *(28.0)* [144]	.823 *(41.8)* [328]	.776 *(39.5)* [379]	.656 *(40.8)* [96]	.799 *(38.7)* [947]
Largest	.876 *(29.5)* [89]	.776 *(37.9)* [370]	.726 *(36.9)* [387]	.686 *(40.5)* [102]	.755 *(37.0)* [948]
Total:	.864 *(29.8)* [751]	.778 *(40.9)* [1686]	.699 *(39.9)* [1863]	.647 *(41.7)* [436]	.749 *(38.8)* [4736]

Note:Reflecting the rules of the Governor's Performance Award program in 1999-2000, the above was limited to elementary schools with more than 100 students**.**

Endnotes

[1] There are at least two reasons why states will still have an incentive to define proficiency at a low level: First, the minimum proficiency rate must be raised from its baseline level to 100 percent within 12 years. Although a lenient definition of proficiency does not guarantee high passage rates in the first year, the rate of required increase in subsequent years is slower in states with lenient definitions. Second, subgroups that close 10 percent of the gap between their proficiency rate last year and 100 percent in a single year are counted as having achieved "adequate yearly progress" even if their proficiency rate falls below the expected level. It would be easier for schools to benefit from this "safe harbor" provision if their proficiency rate starts out at 80 percent than if their proficiency rate starts at 20 percent.

[2] However, after 12 years, the NCLBA requires all schools and all subgroups to achieve 100 percent proficiency. Therefore, while a lenient definition of proficiency may not provide many advantages during the first year or two, a lenient definition would be much easier to satisfy in the coming years.

[3] More detailed results are in an earlier draft of this paper available from the authors.

[4] Coleman, James S., E.Q. Campbell, C.J. Hopson, J. McPartland, A.M. Mood, F.D. Weinfeld, and R.L. York Equality of Educational Opportunity (Washington, DC: U.S. Department of Health, Education and Welfare, 1966).

[5] There could be fewer than 20 percent of schools failing if there are a large number of schools closing more than 10 percent of the gap between their baseline and the goal of 100 percent proficiency, who would be protected by "safe harbor" provisions.

6. As in other sections of this paper, the statistics for school level characteristics are weighted by school size unless otherwise noted.

7. In 2001 the minimum proficiency rate in Texas was raised slightly to 55 percent.

8. In 2001, the 200 student minimum was lowered to 50 students.

9. A socioeconomically disadvantaged student is a student of any race neither of whose parents completed a high school degree or who participates in the school's free or reduced price lunch program.

10. Kane, Thomas J., and Douglas O. Staiger "Volatility in School Test Scores: Implications for Test-Based Accountability Systems" Brookings Papers on Education Policy, 2002 (Washington, DC: Brookings Institution, 2002).

11. This rough approximation was calculated using the number of students with valid scores used in calculating API scores in California– approximately 300,000 African American students and 1.4 million Latino students.

12. In 2001, the absolute threshold was dropped from 200 to 50 students.

13. Grissmer, David and Ann Flanagan, "Exploring Rapid Achievement Gains in North Carolina and Texas" Paper written for the National Education Goals Panel, (November, 1998); Jacob, Brian A. "The Impact of High-Stakes Testing on Student Achievement: Evidence from Chicago" Unpublished paper, Kennedy School of Government, Harvard University, June 2001; Jacob, Brian A. and Steven D. Levitt, "Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating" Unpublished paper, Kennedy School of Government, Harvard University, April 2002; Koretz, Daniel. "Limitations in the Use of Achievement Tests as Measures of Educators' Productivity" Unpublished paper, Rand Corporation, May 2000. Paper initially presented at "Devising Incentives to Promote Human Capital", National Academy of

Sciences Conference, December 1999 (Forthcoming in the *Journal of Human Resources*).

14. For a summary of the evidence on the racial gap in test performance, see Christopher Jencks and Meredith Phillips (eds.), The Black-White Test Score Gap (Washington, DC:Brookings Institution, 1998).