# DESIGNING INCENTIVES FOR CALIFORNIA SCHOOLS

**Thomas J. Kane, UCLA Professor of Policy Studies & Economics**
**Douglas O. Staiger, Dartmouth College**

During the 1990's, many states began using test scores to hold public schools accountable for their students performance. The increasing emphasis on test-based accountability represented a fundamental transformation of the department of education in many states. Rather than serving primarily as a financial conduit, funneling state and federal resources to local school districts, they began to play a more active role in developing standards and providing incentives to schools and local districts to focus on student performance. The federal No Child Left Behind Act of 2001 (NCLB) has raised the stakes even further.

Even before the passage of the NCLB, California had constructed an elaborate set of incentives for schools using student test scores. The incentives were designed to encourage schools to focus on improving performance and to close the large gaps in performance by race and ethnicity. In this chapter, I first describe the incentive system used in California and identify some of the unintended implications of that system. I then describe the portion of the No Child Left Behind Act that deals with school accountability and describe the implications for California. Finally, I suggest several changes to the California accountability system to ameliorate the shortcomings identified in this essay.

## The California School Accountability System

Beginning with the 1999-2000 school year, the California Department of Education provided financial incentives for schools and teachers to focus on student performance.[1] In fact, there were three separate award programs: the School Site Employee Bonus Program ($350 million), the Certificated Staff Bonus Program ($100 million) and the Governor's Performance Award ($227 million). In the first year, the financial incentives totaled $677 million, or approximately $114 per student attending a public school in California.[2]

California's budget crisis subsequently led to substantial cuts in all three programs. The School Site Employee Bonus program was originally authorized for just one year and was not renewed. The Certificated Staff Bonus program was not funded in the second or third year. In contrast, the Governor's Performance Award program did survive a second year, providing $144 million in awards for improvements between 2000 and 2001. Although funding may be restored in the future, the Governor's Performance Award program was not funded for the Spring 2002 and Spring 2003 testing seasons.

Nonetheless, schools are still being recognized with certificates and letters of recognition. Although the financial value of such a designation is difficult to determine, many of the award-winning schools still use the "Governor's Performance Award" logo on their web sites.[3]

All of the awards are based on an "Academic Performance Index" (API), which is calculated separately for each school, for each racial/ethnic subgroup, and for "socioeconomically disadvantaged" students in the school. The API is a weighted average of the proportion of youth scoring in each quintile in a nationally normed sample. The proportion of students in the bottom quintile receive a weight of 200, with higher weights for the proportion of youth in higher quintiles (500 for the 2nd, 700 for the third, 875 for the fourth and 1000 for the top quintile).

Weights in the API were constructed to provide more points for improvement in low-scoring schools. Moving 10 percent of a school's students from the bottom quintile to the second quintile means an improvement in the API of 30 points (since the difference in weights between the bottom and the second quintile was 500 minus 200 or 300). Moving 10 percent of a school's students from the fourth to the top quintile implies an improvement of 12.5 points. Although it is not clear which challenge is more difficult – moving very low performing kids up one notch or moving moderately high-achieving youth to the top of the distribution – the increases in API scores have tended to be larger for low performing schools.

**Brief Description of the Incentives**

The California school incentives were designed with two goals in mind– to provide schools with an incentive to improve and to ensure that schools do not ignore disadvantaged minority groups when they do so. In pursuit of the first goal, one option would have been to provide financial awards as a function of schools' absolute level of performance on student test scores in the spring. The designers of the accountability system decided against this approach, because it would have provided a windfall to schools in more affluent districts, which had higher baseline test scores.

To avoid giving an advantage to schools in more affluent districts, and to focus the incentives on improvements in performance, the state based the financial rewards on the *change* in performance from one year to the next. Each school is given a "growth target" based upon its API score from the previous spring. The growth target is 5 percent of the difference between the baseline API score and the statewide target of 800. Schools with a baseline score over 800 are expected to raise their API scores by a minimum 5 points in order to qualify for the monetary awards.

As noted above, the growth targets apply not only to the school overall, but to each of up to seven racial/ethnic subgroups: African American (not of Hispanic origin), Native American, Asian, Filipino, Hispanic/Latino, Pacific Islander and White (not of Hispanic Origin). In addition, schools are expected to achieve growth for students who are "socioeconomically disadvantaged", defined as those students whose parents have not graduated from high school or who are eligible for the free or reduced price lunch program. Only those subgroups that are "numerically significant" count. The state defines "numerical significance" as those subgroups which represent at least 15 percent of the tested students *and* more than 30 students, or more than 100 students regardless of their percentage. Each subgroup is expected to reach 80 percent of the schoolwide growth target. (If the schoolwide target is 5 point growth, each subgroup must achieve at least 4 point growth for the school to quality for the Governor's Performance Award.)

**Unintended Consequences**

The power of any incentive system to transform schools depends upon the nature and strength of the incentives provided. As is usually the case, the "devil is in the details." In this section, I briefly describe some of the weaknesses of the California accountability system.

*First, single-year improvements in school test scores are very imprecise.* Much of the variance in the change in performance from one year to the next is due to two factors: sampling variation and other one-time factors affecting school performance. As reported in the U.S. Department of Education's Common Core Data for public schools, the median California elementary school contained 88 students per grade level.[4] With such a small sample size, a few particularly rowdy or particularly bright students can have a large impact on test scores in a given year.[5]

The importance of sampling variation is illustrated in Figure 1, which plots the growth in test scores between 1999 and 2000 (minus the targets), for schools and for African American and Latino subgroups by the number of students tested. There are three facts worth noting in Figure 1. First, most of the points are above zero on the y-axis, reflecting the fact that during the 1999-2000 school-year most schools and subgroups exceeded their growth targets. Second, the variance in growth is wider for smaller schools and narrows for larger schools. This reflects the impact of sampling variation-- there are wider swings in mean performance in small schools than in large schools. Small schools were more likely to have large swings in either direction-- positive or negative. Third, many of the subgroups were quite small, meaning that the distribution of subgroup growth tends to have a wider variance.

Sampling variation is only one source of short-term fluctuations in test scores. Other factors, such as a dog barking in the parking lot on the day of the test, a severe flu season, or interactions between the test and the examples used by the teachers in a particular school, could also lead to non-persistent changes in test scores. Using data from North Carolina, Kane and Staiger (2002b) estimate that nearly three quarters of the variance in the change in fourth grade test scores is due to sampling variation and other one-time causes. Only about a quarter of the variance reflected changes that persisted more than one year.

The imprecision of the single year changes meant that many awards were given for short-term, non-persistent changes. Moreover, the likelihood of winning an award was related to school size. Given that the target improvement in the first year was below the mean improvement, small schools were at a disadvantage in winning the Governor's Performance Awards. (Small schools were more likely to have improvements in the left tail of the distribution. The opposite would have been true if the minimum goal were set above the mean. Small schools would have been more likely to be positive outliers.) Even though the mean improvement in performance was slightly larger in small schools, only 68 percent of the smallest quintile of elementary schools won Governor's

Performance Awards in the first year, compared to 74 percent of the largest quintile of schools.

There was even more of a difference between large and small schools at the extremes of the distribution. The Certificated Staff Bonus program provided bonuses of $25,000 per teacher for 1000 teachers in schools with the largest changes in performance between 1999 and 2000, $10,000 bonuses to 3500 teachers in schools with the next largest increases and $5,000 bonuses to 7500 teachers in schools with the next largest changes. To be eligible, a school must have had an API score in 1999 below the statewide median and must not have had a decrease in performance between 1998 and 1999. Within this subset of schools, Table 1 reports the proportion of schools winning the Certificated Staff Bonus awards by school size decile. Schools in the smallest decile were roughly 8 times more likely to win a bonus of $5,000 or more under the Certificated Staff Bonus program than schools in the largest decile. (Their advantage for the $10,000 and $25,000 bonuses was even larger.)

Table 1 also reports the proportion of schools with declines in API scores in the subsequent year, from 2000 to 2001. The proportion is reported separately for those schools which won Certificated Staff Bonus awards based upon their improvement between 1999 and 2000 and those that did not. Schools were considerably more likely to have test score declines in the subsequent year if they had won the teacher bonuses based upon their performance in 2000 (26 percent versus 15 percent).

*Second, for purely statistical reasons, the subgroup rules disadvantage schools containing many different subgroups.* 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. This point is illustrated in Figure 2, which plots API growth (in excess of targeted growth) for white subgroups against African American (left plot) and Latino (right plot) subgroups in the same school (for the 1999-2000 growth cycle). 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. The probability of exceeding their growth target in 1999-2000 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). 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 unreliably measured, 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 (drawn from Kane and Staiger (2002b)), 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.

Table 2 has at least two important implications. First, under such rules, a district would have a strong incentive to segregate by race/ethnicity. Consider a district with 4 small schools, each being 25 percent African American, 25 percent Latino, 25 percent Asian American and 25 percent white, non-Hispanic. According to the results in Table 2, the district could nearly double each school's chance of winning an award simply by segregating each group and creating four racially homogeneous schools.

Second, because minority youth are more likely to attend heterogeneous schools than white non-Hispanic youth, the rules have the ironic effect of putting the average school enrolling minority students at a statistical disadvantage in the pursuit of award money. 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 2-3 subgroups, while most African Americans attend schools with 3 or more 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 (1,2,3,4+) of school 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, African Americans are nearly 5 percent less likely to be in an award winning school solely because of 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.[6]

*Third, because the rewards are based upon the improvement from one year to the next, any improvement in one year raises the bar for future years.* In designing any incentive system, identifying the standard against which to judge performance is often a challenge. When that standard is a function of one's own prior performance, improvements today can actually raise the cost of earning rewards later. The problem is known in managerial economics as the "ratchet effect".[7] By ratcheting up the standard based upon a school's most recent performance, such a system could actually lead schools to under-invest in improving their productivity because, on the margin, a larger improvement this year lowers the probability of an award in all future years.

## Implications of Federal NCLBA

The federal No Child Left Behind Act (NCLBA) will present a new set of challenges for California schools. Rather than being based on the annual change in performance, the federal law requires schools and subgroups within schools to meet a standard based upon their level of performance in the most recent year. Because the state uses changes and the federal law uses levels, many of the schools that are awarded by the state based upon their improvements, will be sanctioned under the federal law because their level of performance is below the threshold.

Under NCLBA, each state is allowed to define "proficiency" using its own standards. However, once a state settles on a definition of proficiency, the minimum acceptable proficiency rate in the state will be set at the proficiency rate of the 20th percentile school. In other words, every school will be required to achieve a higher rate of proficiency in reading and math than did the bottom 20 percent of schools in their state in 2001-02. States are required to raise that minimum proficiency rate at uniform intervals every few years to ensure that within twelve years, the minimum proficiency rate for all schools nationally is one hundred percent.

With this definition, the federal government has ensured that states can not exempt themselves by choosing a lenient definition of proficiency. At least initially, states which choose to define proficiency leniently will simply be required to achieve a *higher* minimum rate of proficiency, since the 20th percentile school will have a higher proficiency rate. However, laxness in the definition of proficiency does have its rewards, since it will be much easier to achieve one hundred percent proficiency if a state starts out with more than eighty percent of their students proficient (such as Texas) than if one starts out with forty percent of one's students proficient.

Given the description above, one would expect about twenty percent of schools in every state to fail on the basis of their school-wide scores, at least initially. Yet newspapers in many states have been reporting much higher failure rates.[8] The reason is that the minimum proficiency rate will apply not only to the school overall, but to every subgroup in a school defined by race/ethnicity, socioeconomic disadvantage, disability status and English language learner status.

Schools differ in their student performance much less than many people realize. Because very high-achieving and very low-achieving schools are most salient, we may infer a wide dispersion in performance. However, there are many schools in the middle. The 20th percentile school in most states will have a mean test score only about one-third of a standard deviation below the mean. (Kane and Staiger (2003)) Such a difference is small relative to the racial/ethnic differences in performance. In most states, the black-white differential in performance is usually more than three quarters of a standard deviation. As a result, a large share of the schools which contain a disadvantaged minority subgroup will fail to achieve "adequate yearly progress" (AYP) as defined by Congress.

**Implications of the Subgroup Rules in No Child Left Behind**

The subgroup rules will generate failure rates of more than fifty percent in many states. Kane and Staiger (2003) used data on math scores for 3rd through 5th grade students in North Carolina elementary schools to illustrate this point. There is much overlap in test scores at the individual student level: 30 percent of individual African American students have test scores above the statewide mean. However, in North Carolina, as in many other states, despite the overlap in the distribution in performance at the individual student level, the difference in *mean performance* within schools remains quite substantial. Whereas 30 percent of individual African American students scored above the overall mean statewide, only 2 percent of African American students were in schools where the *mean* performance of African American students in the school exceeded the statewide mean. As a result, a large fraction of those schools which contain an African American or economically disadvantaged subgroup will fail to achieve AYP, even if their school-wide rates of proficiency exceed the minimum required.

The subgroup rules were intended to shine a harsh light on schools which have allowed the performance of minority youth to lag for decades and to provide incentives to schools to focus their efforts on closing those gaps. However, the subgroup rule suffers from several serious shortcomings, which will blunt the law's impact.

First, the law requires states to define how many minority students it takes to "count" as a separate subgroup. The NCLBA does not define subgroup status beyond stating that a group counts as a separate subgroup when the number of students in a category is sufficient to yield "statistically reliable information." Since there is no such magical sample size which produces "statistically reliable information", states will define the minimum sizes differently. But wherever the threshold is drawn, the stakes will be very high for schools on either side of what must be an arbitrary threshold. For instance, in the academic year 1999-2000 in Texas, to count as a separate subgroup, 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 regardless of the percentage). Yet, in order to achieve "exemplary" status, a school in Texas was required to have a 90 percent proficiency rate for each group that met the minimum size requirements. Given the differences in performance by race and ethnicity, the stakes were quite high for schools with a percentage of minority students near the 10 percent minimum.

Among the schools that did not also have an African American subgroup, 42 percent of schools with exactly 9 percent of students Latino (where Latino students did not count as a separate subgroup) were rated exemplary, while less than 20 percent of the schools with exactly 10 percent of students Latino were rated exemplary. In other words, despite the fact that the mean performance overall was quite similar and the mean performance of the minority students was quite similar, a one-percentage point difference in the percentage of students in a particular racial/ethnic group meant a more than doubling of a school's chance of being recognized as "exemplary," because schools with 10 percent Latino students were held accountable for Latino scores separately and schools with 9 percent Latino students were not. Given the large racial differences in performance, the designation of minimum size requirements for subgroups of students will determine the fates of schools near the thresholds. Unfortunately, they will do so arbitrarily.

Interestingly, California raised the minimum threshold for "numerical significance" from that which it had been using for its own accountability system. To count as a separate subgroup under NCLBA, a subgroup must represent at least 15 percent of the student body and more than 50 students or more than 100 student regardless of the percentage. (Formerly, the threshold was 15 percent of the student body and more than 30 students. The same absolute threshold of 100 students regardless of the percentage was being used.)

Second, the subgroup rules will lead to very uneven failure rates in different parts of the country, depending upon the percentage of disadvantaged minorities in their schools and the degree of integration. Kane and Staiger (2003) used data on individual schools from 48 states in 1999-2000 and applied the definition of subgroup status used by California, which required a minority group to contain at least 30 students and 15 percent of the students in a school or greater than 100 students to constitute an official subgroup. The proportion of schools containing an African American or Latino subgroup varied widely by state, depending upon the representation of African American and Latino youth in the resident population and the degree of integration.

While a majority of the public schools nationwide (54 percent) contained an African American or Latino subgroup, the percentages were much higher in the South and West. More than 80 percent of the public schools in seven states (TX, MS, NM, CA, LA and SC) and the District of Columbia contained an African American or Latino subgroup. An additional seven states (VA, NC, NV, FL, GA, AL and AZ) contained African American or Latino subgroups in more than 60 percent of their public schools. A large share of these schools are likely to fail, simply because of their demographics.

Moreover, the more integrated a state's schools are, the higher proportion of their schools are likely to be affected by the NCLBA. 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.

On August 15, 2003, the California Department of Education released its preliminary list of schools failing the new federal standard. Roughly half (48 percent) of schools enrolling 52 percent of students of public school students statewide, failed to meet the definition of "adequate yearly progress" under the federal standard.

**Implications of School Failure**

The implications for failing schools are fairly lenient at the outset, but become increasingly severe. Failing schools will be required to submit a school improvement plan to describe their strategy for improving. After two years of failing, students in the schools receiving Title I funds must be given the option of attending a non-failing school in the district and a portion of the Title I funds are to be used to pay the transportation costs. After three years of failing, a portion of a school's Title I funds must also be made available to parents as vouchers to pay for "supplemental educational services." After four years of failing, schools receiving Title I funds will be subject to "corrective action," which will require a school to do one of the following "replace school staff relevant to the failure," "institute and implement a new curriculum," "significantly decrease management authority in the school," "appoint outside experts to advise the school," "extend the school year or school day" or "restructure internal organization of the school." The range of options allowed under "corrective action" gives states and districts some flexibility in dealing with failing schools. However, after 5 years of failure, the room for flexibility is considerably reduced. Such schools will be required to reopen as a public charter school, replace all or most of the school staff (including the principal), hire a private management company to operate the school, or be subject to a state takeover.

During the first years of implementation, there will be two main dangers: First, the law will lead to an unprecedented amount of wasteful paper shuffling. School failing to achieve AYP will be required to draw up a school improvement plan. In California, this will be roughly half of the schools in the state initially. When so many schools are submitting school improvement plans at the same time, few states will have the resources to review those plans credibly.

Second, failing schools will also be required to make supplemental educational services available to students. States will be expected to publish a list of acceptable providers of such services. Any time public funds are available for private use, there is always a danger of fraud and abuse. We are likely to see a number of providers spring up, eager to accept those vouchers for non-productive uses while claiming an educational benefit. Particularly when the demand for those services blossoms in two years time, states will be hard pressed to distinguish worthwhile services from the not-so-worthwhile.

**Labor Market Payoffs for Better Performance**

The movement to hold schools accountable for student test scores spread quickly among state governments in the late 1990's. By the spring of 2002, virtually every state and the District of Columbia had implemented some form of accountability for public schools using test scores. The federal No Child Left Behind Act mandates even broader accountability, requiring states – including California – to test children each year in all grades three through eight and to intervene in schools failing to achieve specific performance targets.

Critics of school accountability worry that current systems already place too great a weight on imperfect measures of academic achievement and, on net, may do more harm than good-- by encouraging a narrowing of the curriculum or student and teacher cheating. To evaluate these concerns, one must have a sense of the potential value that we should place on an increase in student achievement. Calculations by Kane and Staiger (2002a) revealed that the monetary value of even a small improvement in academic achievement can have very large payoffs.

The first challenge was to come up with an estimate of the value of test performance for earnings. Two papers provide estimates of the impact of test performance on the hourly wages of young workers. Murnane, Willett and Levy (1995) estimate that a one-standard deviation difference in math performance is associated with an 8 percent hourly wage increase for men and a 12.6 percent increase in for women.[9] This pair of estimates probably understates the value of test performance, since the authors also control for years of schooling completed.[10] Neal and Johnson (1995), who do not condition on educational attainment, estimate that an improvement of one standard deviation in test performance is associated with a 18.7 and 25.6 percent increase in hourly wages for men and women, respectively.[11] With a discount rate of 3 percent, the present value at age 18 of an increase of one standard deviation in test performance is worth roughly $110,000 per student using the Murnane, Willett and Levy estimates and $256,000 per student using the higher estimates from Neal and Johnson.[12] Discounting these values back to age 9 – that is, 4th grade – would reduce the estimates to $90,000 and $215,000 per student. A 3 percent discount rate may be too low, since investment in human capital is not risk free. But even with a discount rate of 6 percent, these estimates are only reduced by about one third.

Such estimates are quite large relative to the rewards offered to schools for increasing student test performance. Even in the first year of the financial incentives, when they were most generous, California paid elementary schools and their teachers an average award of $122 per student if their school improved student performance by an average of at least 0.03 student-level standard deviations.[13] Based on the calculations in the preceding paragraph, the present value of an increase of .03 standard deviations is in the range of $2700 to $6400 per student (0.03 times $90,000 or $215,000). In other words, the labor market value of the test score increase would have been worth roughly 20 to 50 times the value of the incentive provided by California in the year of its most aggressive financial incentive program.

Because test score measures are imprecise and because the possibility for distorting school curricula is large, we would want to rein in the financial incentives and not pay the full value on the margin of a true test score increase. But, even when California was attaching very high financial stakes to the test score increases, it was only paying a small share of the value on the margin, if the test score increases were real. Therefore, while it would be worthwhile to make some incremental improvements in the design of the programs, it would not make sense to abandon the financial incentives entirely.

**Policy Options**

There are five ways in which the accountability system could be improved. First, schools could be ranked relative to their 1999 base, rather than updating the base each year. This approach would still be imprecise, but there is likely to be more true variance (more variance in the "signal") in the extent to which schools have improved over the longer time period. Moreover, using the 1999 base eliminates the "ratchet effect" of basing each school's new target on their most recent performance.

Second, as currently structured, the subgroup rules disadvantage schools with many subgroups. Moreover, there is little evidence that the subgroup rules lead schools to focus on disadvantaged student performance any more than they would have with only the schoolwide targets. Kane and Staiger (2003) study the performance of African American and Latino students in schools in California immediately above and below the cut-offs for numerical significance. For example, they compare the performance of African American students in schools where they represent 14 percent of the student body (and, therefore, do not count as a separate subgroup) to the performance of African American students in schools where they represent 15 percent of the student body and, therefore, do count as a separate subgroup. There is little evidence that disadvantaged minority youth perform better as a result of the subgroup designation.

Nevertheless, because the subgroup rules are an important part of the federal legislation, it is unlikely that the subgroup rules will be dropped. As an alternative, the reward for schools could be made proportionate to the proportion of students in groups that achieve the standard. For instance, if the groups representing 75 percent of the student body meet their growth targets, the school could be given 75 percent of the award. This would at least eliminate the statistical disadvantage of requiring schools to hit their targets for all subgroups in order to receive the award money.

Obviously, such pro-rating may be useful in apportioning financial awards, but it would be more difficult to do when providing indivisible awards, such as simply being designated an award-winning or exemplary school.

Third, because of sampling variation and other one-time factors, it does not make sense to provide large awards at the extremes, where small schools would be at a large advantage (or disadvantage if penalties are being imposed at the bottom). California's Certificated Staff Bonus program – providing large bonuses of $5,000 to $25,000 – is an example of such policy. Fortuitously, that program has already lost funding. Hopefully, it will not be restored.

Fourth, the award competitions based upon Academic Performance Index (API) growth could be organized by school size, as is done in high school sports. Otherwise, either small or large schools will be advantaged (depending upon the design), purely as a result of sampling variation. In fact, the reason for doing so in high school sports is also due to the nature of sampling distributions: A school with 3000 students is much more likely to have a large number of students over 6'5" in height with which to field a basketball team than a school with 300 students.[14] As with basketball or football, schools should compete with other schools of similar size when measuring their performance with mean test scores.

Fifth, the California Department of Education must create clear guidelines for the approval of supplemental service providers, which will be eligible to receive student vouchers under the No Child Left Behind Act in the next couple of years. Without adequate safeguards to ensure that the funding is used for worthwhile purposes, there is a danger that the funds will be misused. This is an even more serious problem than the federal government faces with the Pell Grant program in post-secondary education– another example of an educational voucher program. Under the Pell Grant program, the federal government has struggled to ensure that only worthwhile educational services are funded. Because the program involves young adults who might otherwise be employed, there is a built-in safeguard: Despite the availability of a federal voucher, students are taking classes during time they could be working. The foregone earnings are a form of co-payment that gives students an incentive to avoid educational programs that are not worthwhile. For younger school-age children, there is less of a built-in safeguard against wasteful expenditures and a greater danger that the supplemental educational vouchers will be used to subsidize services that amount to child-care, with little educational content.

## Conclusion

Test-based accountability has become a fixture of elementary and secondary education in California and the rest of the United States. However, as with any incentive system, rewards based upon testing can create perverse results, depending upon the details of the design of the incentives. Given the apparent labor market payoffs to academic achievement, this chapter does not suggest an abandonment of the incentives based upon test results. However, some small changes in the nature of the incentives provided could reduce some of the perverse incentives and keep schools focused on improving student performance.

## References

Grogger, Jeffrey and Eric Eide "Changes in College Skills and the Rise in the College Wage Premium" Journal of Human Resources (Spring 1995) Vol. 30, No. 2, pp. 280-310.

Jencks, Christopher and Meredith Phillips, "Aptitude or Achievement: Why Do Test Scores Predict Educational Attainment and Earnings?" in Susan Mayer and Paul Peterson (eds.) Earning and Learning: How Schools Matter (Washington DC: Brookings Institution and Russell Sage Foundation Press, 1999), pp. 15-48.

Kane, Thomas J. and Douglas O. Staiger (2002a) "The Promise and Pitfalls of Using Imprecise School Accountability Measures" Journal of Economic Perspectives (Fall, 2002), Vol. 16, No. 4, pp. 91-114.

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

Kane, Thomas J. and Douglas O. Staiger, "Unintended Consequences of Racial Subgroup Rules" in Paul E. Peterson and Martin R. West (eds.) No Child Left Behind? The Politics and Practice of Accountability (Washington, DC: Brookings Institution Press, forthcoming 2003).

Kane, Thomas J. and Douglas O. Staiger "Improving School Accountability Measures" National Bureau of Economic Research Working Paper No. 8156, March 2001.

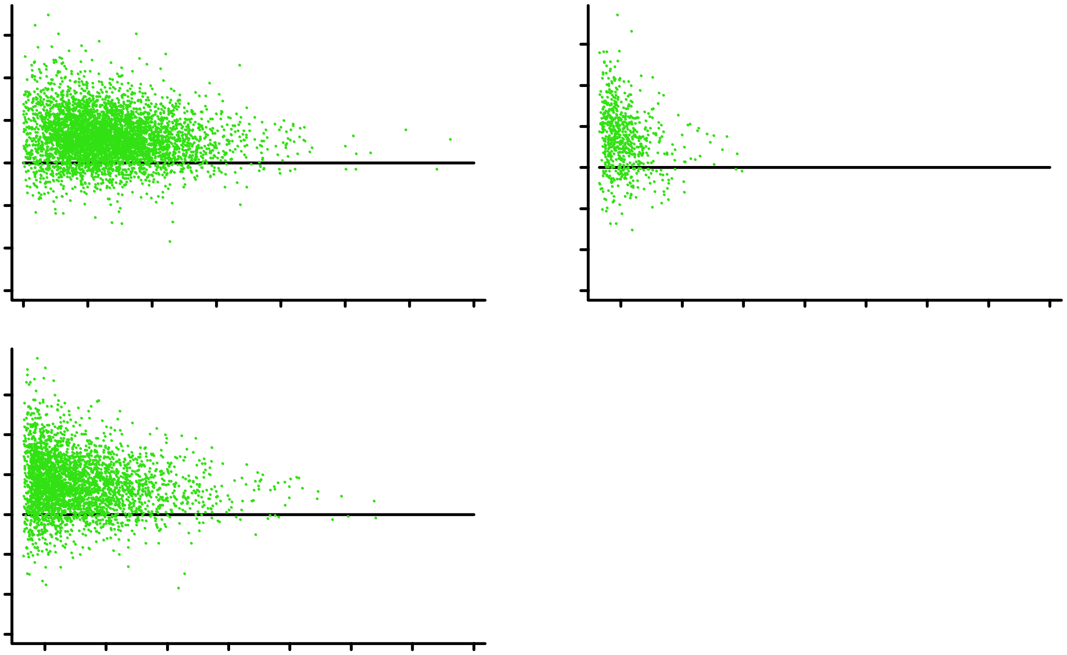
Milgrom, Paul and John Roberts Economics, Organizations and Management (Englewood Cliffs, New Jersey: Prentice Hall, 1992).

Murnane, Richard J., John B. Willett and Frank Levy, "The Growing Importance of Cognitive Skills in Wage Determination" Review of Economics and Statistics Vol. 77, No. 2 (May 1995): 251-266.

Neal, Derek and William Johnson "The Role of Premarket Factors in Black-White Wage Differentials" Journal of Political Economy (1996) Vol. 104, pp. 869-895.

Robelen, Eric "State Reports on Progress Vary Widely" Education Week September 3, 2003.

**Figure 1.**



Note: Based upon growth in California API scores between 1999 and 2000. Figure drawn from Kane and Staiger (2003).

## Figure 2. Changes in API Scores for Subgroups Attending the Same Elementary Schools (1999-2000)

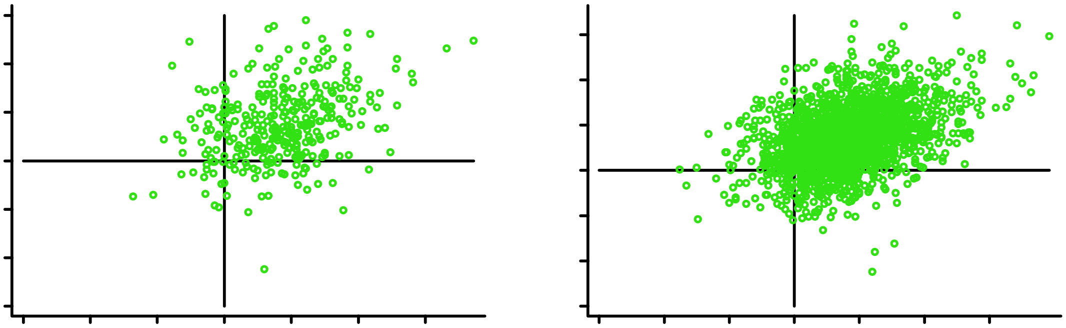


**Table 1.**
**Proportion of Eligible Elementary Schools Winning**
**Certificated Staff Bonus Awards by School Size**

| School Size Decile | Proportion winning teacher bonuses: | | | Percent with API Decline In Subsequent Yr (2000-01) | |
|---|---|---|---|---|---|
| | ∃$5,000 per teacher | ∃$10,000 per teacher | ∃$25,000 per teacher | if No CSB Bonus | if Won CSB Bonus |
| Smallest | .253 | .111 | .037 | 25 | 29 |
| 2nd | .176 | .103 | .039 | 23 | 18 |
| 3rd | .108 | .028 | 0 | 22 | 45 |
| 4th | .166 | .046 | .017 | 18 | 21 |
| 5th | .128 | .039 | .006 | 15 | 48 |
| 6th | .109 | .041 | 0 | 15 | 19 |
| 7th | .075 | .032 | 0 | 16 | 21 |
| 8th | .063 | .018 | .004 | 11 | 14 |
| 9th | .050 | .021 | .004 | 12 | 8 |
| Largest | .031 | .007 | 0 | 9 | 11 |
| Total: | .105 | .039 | .008 | 15 | 26 |

Note: Based upon author's tabulation of data from the 1999-2000 school year. The sample of elementary schools was limited to those who were eligible for the certificated staff bonus program: those with scores in bottom 5 deciles in 1999, with no decline in test scores between 1998 and 1999.

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

Note: Drawn from Kane and Staiger (2002). The above was limited to elementary schools with more than 100 students to reflect the rules of the Governor's Performance Award program in 1999-2000**.**

## Endnotes

[1] For more details on the Public Schools Accountability Act of 1999, see http://www.cde.ca.gov/psaa/sb1x.htm.

[2] There were 5,951,612 students attending public elementary and secondary schools in California in 1999-2000. California Department of Education, *Fact Book 2003*.

[3] Joetta Sack, "California Drops Its Bonuses to Schools" *Education Week*, October 30, 2002.

[4] Based upon author's calculation using the Common Core of Data for 1999-2000.

[5] Middle schools and high schools tend to be larger: the median school enrolling 7th and 9th grade students in California enrolled 128 and 99 students respectively. However, as discussed in Kane and Staiger (2002b), variance in school size tends to be much larger in middle schools and high schools, reflecting the rural/urban differences in community sizes. (Elementary schools are more uniform in size, due to more uniformity in within-district catchment areas.) As a result, the difference in the advantage/disadvantaged associated with size is also larger for middle and high schools.

[6] 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.

[7] Milgrom and Roberts (1992), p. 232.

[8] Robelen (2003).

[9] Using similar data, but conditioning on high school grades as well as educational attainment, Grogger and Eide (1995) find that a standard deviation in math scores was associated with a 5 percentage point wage increase for men and 7.5 percentage point increase for women in 1986.

[10] The Murnane, Willett and Levy estimates may also differ because they include only the math test score measure and not the composite measure of reading and math skills.

[11] The correlation between test scores and earnings is not simply reflecting the payoff to innate abilities, since improvements in test scores are also associated with higher earning prospects. Jencks and Phillips (1999) find that a one standard deviation improvement in math scores between 10th and 12th grade was associated with a 26 percent increase in earnings 10 years after high school graduation.

[12] Kane and Staiger (2002a) used the following calculation, incorporating productivity growth as suggested by Krueger (2002):

where: ∃ is the proportional rise in wages associated with a given test score increase; $w_i$ represent wages from age 18 through 64 estimated using full-time, year-round workers in the 2000 CPS; ( represents the general level of productivity growth, assumed to equal 0.01; and r is the discount rate, assumed to equal 0.03.

[13] The School Site Employee Bonus program provided $591 per full-time equivalent teacher to both the school and teacher, or $59 per student based on an average of 20 students per teacher. The Governor's Performance Award (GPA) program provided an additional $63 per student. The growth target for the average elementary school was 9 points on the state's Academic Performance Index (API). Because the state did not publish a student-level standard deviation in the API scores, we had to infer it. A school's API score was a weighted average of the proportion of students in each quintile of the national distribution on the reading, math, language and spelling sections of the Stanford 9 test. For elementary schools, the

average proportion of students across the four tests in each quintile (from lowest to highest) was .257, .204, .166, .179 and .194 and the weights given to each quintile were 200, 500, 700, 875 and 1000. Under the assumption that students scored in same quintile on all four tests, we could calculate the student-level variance as $.257(200-620)^2+.204(500-620)^2+.166(700-620)^2+.179(875-620)^2+.194(1000-620)^2=89034$, implying a standard deviation of 298. This is nearly 5 times the school-level variance, which is roughly consistent with expectations.

[14] The rationale for organizing leagues in high school sports involves the sampling distribution for an order statistic, rather than a sampling distribution for a mean.