NBER WORKING PAPER SERIES

BEYOND INCENTIVES: DO SCHOOLS USE ACCOUNTABILITY REWARDS PRODUCTIVELY?

Marigee Bacolod
John DiNardo
Mireille Jacobson

Working Paper 14775
http://www.nber.org/papers/w14775

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
March 2009

Beyond Incentives: Do Schools use Accountability Rewards Productively?
Marigee Bacolod, John DiNardo, and Mireille Jacobson
NBER Working Paper No. 14775
March 2009
JEL No. H0,I0,I2,J0,J24

ABSTRACT

"Accountability mandates" -- the explicit linking of school funding, resources, and autonomy to student performance on standardized exams -- have proliferated in the last 10 years. In this paper, we examine California's accountability system, which for several years financially rewarded schools based on a deterministic function of test scores. The sharp discontinuity in the assignment rule -- schools that barely missed their target received no funding -- generates "as good as random" assignment of awards for schools near their eligibility threshold and enables us to estimate the (local average) treatment effect of California's financial award program.

This design allows us to explore an understudied aspect of accountability systems -- how schools use their financial rewards. Our findings indicate that California's accountability system significantly increased resources allocated to some schools. In the 2000 school year, the average value of the award was about 60 dollars per student and 50 dollars in 2001. Moreover, we find that the total resources flowing to districts with schools that received awards increased more than dollar for dollar. This resource shift was greatest for districts with schools that qualified for awards in the 2000 school year,the first year of the program, increasing total per pupil revenues by roughly 5 percent.

Despite the increase in revenues, we find no evidence that these resources increased student achievement. Schools that won awards did not purchase more instructional material, such as computers, which may be inputs into achievement. Although the awards were likely paid out as teacher bonuses, we cannot detect any effect of these bonuses on test scores or other measures of achievement. More worrisome, we also find a practical effect of assigning the award based in part on the performance of "numerically significant subgroups" within a school was to reduce the relative resources of schools attended by traditionally disadvantaged students.

Marigee Bacolod
Department of Economics
3151 Social Science Plaza
University of California-Irvine
Irvine, CA 92697-5100 USA
mbacolod@uci.edu

Mireille Jacobson
UC Irvine
202 Social Ecology I
Irvine, CA 92697-7075
and NBER
mireille@uci.edu

John DiNardo
440 Lorch Hall
Ford School of Public Policy
University of Michigan, Ann Arbor
Ann Arbor, MI 48109-1220
and NBER
jdinardo@umich.edu

1 Introduction

"Accountability mandates" – the explicit linking of a public school's funding, resources, and autonomy to student performance on standardized tests – have proliferated in the last 10 years. A major impetus to this proliferation was the No Child Left Behind Act (NCLB) of 2001, which requires states test their students in reading and math annually from third to eighth grade. Accountability mandates can be crudely divided into those that rely primarily on "carrots" – money and recognition awarded to schools, its teachers, staff, and/or students for performing well on these tests – and those that rely on "sticks" – withholding of funds, intervention, or outright takeover for low performing schools.[1]

The accountability reforms in the California Public School system that we study are an example of the first type: schools and teachers within those schools that made adequate progress or attained a "passing grade" were rewarded with cash bonuses. Three programs provided mechanisms for rewarding schools and teachers in high performing California schools: the Governors Performance Award Program (GPAP), the Certificated Staff Performance Incentive Award (CSPIA), and the Schoolsite Employee Performance Bonus (SEPB). While one of these programs (GPAP) "was envisioned as money to be used for school site purposes [e.g., purchasing computers]," the California Department of Education found it was in many cases "awarded to certificated staff [i.e., teachers] in the way of bonuses or stipends..."[2] In effect, the three programs combined to substantial teacher bonuses. GPAP was funded for up to $150 per pupil at winning schools. Assuming, as the California Department of Education suggests, that these funds were paid out as teacher bonuses,

[1] While deferrals and withholding of funds is stipulated as a possible intervention in NCLB, to our knowledge it has not been practiced in any state. Most systems that rely on "sticks" threaten state takeover and/or outright closure of a school.

[2] See page 4 of the document by Patrick J. Chladek, http://www.wcer.wisc.edu/CPRE/conference/nov02/chladek.pdfAs we explain further below, schools had discretion on how to spend these awards.

distributed funds amounted to an additional $1300 per teacher. Explicit bonuses paid out under the CSPIA ranged from $5,000 to $25,000 per teacher and the SEPB paid on average $591 per full-time equivalent (FTE). Thus, teachers at winning schools could have earned up to $27,000, although $2,000 was probably more typical.

In this paper, we evaluate the effect of these three programs with particular focus on the questions of immediate policy relevance: what happens when (through an accountability system) we increase a school's resources? How does the school spend these additional resources? Given that the awards ended up mainly as teacher bonuses, did the awards and teacher bonuses increase student achievement?

An important challenge to evaluations of interventions based on student performance on standardized tests is that these tests are necessarily imperfect. As Kane and Staiger (2002) note, since exam scores contain both "noise" and "signal," measures of aggregate student performance at smaller schools will be noisier than those at large schools with similar students. The presence of noise can have important consequences for naive before and after comparisons of schools facing accountability mandates. Chay, McEwan, and Urquiola (2005), for example, document that accounting for mean reversion in exam performance substantially reduces the estimated impact of a Chilean accountability-like program on test score gains. A related challenge is that since student bodies change, assessing which changes in exam performance are merely the "spurious" consequence of changing student characteristics rather than changes in school practices *per se* can be difficult.[3]

The deterministic nature of California's awards program allows us to circumvent many of the difficult issues involved in evaluating the use of these resources and their impacts on student achieve-

[3]Figlio and Rouse (2006), for example, find that much of the estimated test score gains associated with Florida's school accountability system can be accounted for by changes in the characteristics of students in the state rather than noise.

ment. Schools and teachers within those schools that met a pre-determined threshold for improvement in exam performance were rewarded with a significant bonus. The sharp discontinuity in the assignment rule - schools that barely missed the performance target received no bonuses - generates quasi-random assignment in awards recipiency for schools that are close to their eligibility threshold. This quasi-random assignment allows us to generate credible estimates of the impact of the cash bonuses associated with this school-based incentive program on both test scores and a host of other outcomes. Most important for our purposes is the question of how these additional resources were used. Our approach also provides a large battery of over-identification tests, which allow us to evaluate the validity of the design.

While this research design has many advantages, it falls short of being a complete assessment of the program's effect on the level and distribution of measured achievement in California. Specifically, it cannot capture any effects of the program that occur uniformly to schools that received the awards ("treatment" schools) and schools that did not ("control" schools). For instance, if the mere presence of the awards program causes all schools to sabotage more important goals in favor of more narrowly tailoring curriculum to maximize test scores, our design will not capture such effects. Moreover, for a school that receives funds from the program, the counterfactual to which it is compared is an otherwise similar school within California that did not receive the funds. Also of interest, but outside the scope of the current paper, is the counterfactual school that was not subject to any accountability scheme.

Although such counterfactuals are of clear interest, our setting is not well suited for a direct evaluation of the incentive effects of California's accountability system. There are two reasons we believe the California rewards program offered only very weak incentives: rewards were allocated based on group performance (i.e., to all teachers and staff in a winning school) and the resulting

financial reward was only a one-time bonus. Individual level performance pay based on a clear measure of output is usually thought to provide the strongest incentives to workers. The group nature of the awards thus faces the free-rider problem, and this should mute the impact and incentive effect of the program for any individual teacher, for example. Moreover, CSPIA and SEPB were only funded for one year and GPAP for only two years.[4] Thus, schools and teachers did not have much opportunity to learn how to cost-effectively increase their odds of winning the rewards. In addition, the instability of the funding of the program both weakens their incentives and possibly sends a signal that they are not core elements of the California accountability scheme.

As a result our paper focuses not on the direct incentive effects of the program but on those of immediate policy relevance: how do schools spend additional resources from an accountability scheme? And does the schools' use of these resources increase student achievement?

We find that California's program had a significant impact on the financial resources allocated to some schools. The average value of the 2000 school year (SY) GPAP award was roughly $1400 per teacher and $1200 per teacher for the 2001 SY award.[5] At its peak, the financial resources flowing to districts with schools that qualified for awards was about 5 percent of per pupil and 6 percent of per teacher resources. We find no evidence, based on either financial or school census data, that these resources were used for instructional purposes. Thus, despite the increase in resources, we find little measurable improvement in standard metrics of achievement, such as exam performance, for those schools that received the award compared to those schools that did not.[6]

[4]California budget shortfalls led to the program cuts. The most sizable teacher bonuses of $5,000-$25,000 under the CSPIA was suspended after lawsuits were filed. Errors by Harcourt Educational Measurement led to scoring inflations and errors. Teachers and schools filed suit saying they were unfairly excluded from the awards program. (Modesto Bee, October 2001)

[5]When discussing school years, we adopt the convention in the literature of calling Fall YY01 to Spring YY02 the YY02 school year.

[6]One argument for accountability programs is its low cost relative to other types of education reforms such as class-size reduction. For instance, see Hoxby (2002).

Our findings suggest that untargeted awards and indiscriminate merit pay to teachers and staff do not guarantee future improvements in academic achievement. This is consistent with the mixed evidence from other studies of school-based teacher incentive pay. For example, while Clotfelter and Ladd (1996) and Ladd (1999) find Dallas' school-based program was associated with significant gains in student achievement, the positive effects found a year before the actual program was in effect suggest this was part of a pre-existing trend. Glewwe, Ilias, and Kremer (2003) find short-term but no long-term gains in achievement in a teacher incentive pay experiment in Kenya, where similar to California school teachers, all teachers in the winning school got the same reward. Evidence on the impact of individual teacher performance incentives is also mixed. While Lavy (2003) and Figlio and Kenny (2007) find individual teacher incentive pay is associated with gains in student achievement, Eberts, Hollenbeck and Stone (2002) find it is unrelated to student achievement.[7]

In what follows, we first discuss California's accountability system and the various awards programs, with particular focus on the determinants of awards eligibility. Along with the institutional background, we present a statistical portrayal of California schools by award receipt. We proceed in Section 3 with a discussion of our econometric framework for estimating the effect of the awards. Our findings are presented in Section 4. Finally, Section 5 offers a summary and concluding observations.

2 Background: California's Academic Performance Index and Governor's Performance Award Program

California's accountability system, which predates NCLB, was established by the Public Schools Accountability Act (PSAA) of 1999. The PSAA was motivated by assessments indicating that

[7] As Figlio and Kenney (2007) acknowledge, however, they cannot rule out pure selection effects using their cross-sectional design.

California students were not progressing at the rate necessary "to achieve a high quality education." Its goal is "to hold each of the state's public schools accountable for the academic progress and achievement of its pupils within the resources available to schools."[8]

To measure progress and rank public schools within the state, the PSAA created the "Academic Performance Index" (API). The API is intended to be a summary measure of school-wide performance on various standardized tests. The index ranges from 200 to 1000 and combines test scores from students in grades 2 to 11. The tests (or other indicators) used and weights accorded to each API component vary from year to year. For the first two years of the program – the only years that the budget allocated funds for performance awards – the API was based solely on the nationally norm-referenced Stanford 9 exam. In broad outline, the API is a (noisy) weighted average of several different exams measured in terms of national percentile ranks, although its precise calculation is somewhat unclear.[9] For middle and elementary schools, the API incorporated scores on reading, language arts, spelling, and math exams. For high schools, the API was based on reading, language arts, math, science and social studies exams.[10]

In the 2000 SY, two other awards programs were funded by the State – the Certificated Staff Performance Incentive Award (CSPIA) and the Schoolsite Employee Performance Bonus (SEPB).

[8]See California Education Code 52050-52050.5 for a statement of Legislative Intent.

[9]It is impossible to do justice to how the API is calculated but the following very abbreviated summary in Rogosa (2003) may be useful: "For completeness, here's a quick reminder of the calculations for the API metric. For a Stanford 9 test, transform the national percentile rank into quintiles: 1-19, 20-39, 40-59, 60-79, 80-99. The quintiles are assigned values 200, 500, 700, 875, 1000; an individual's API score on a single test is the value for the attained quintile. For any collection of students, the API component score for a single test (e.g. Reading) is the average, over the individuals, of these values (any missing test scores are imputed by the mean of the group). For the battery of tests, API scores in grades 2-8 are a combination of Math (.4), Reading (.3), Language (.15), and Spelling (.15). API scores in grades 9-11 are a combination of Math (.2), Reading (.2), Language (.2), Science (.2) and Social Science (.2)."

[10]In 2001, the API was based on both the Stanford 9 and the California Standards Test in English-Language Arts (CST ELA). Additional test components have been added since then. Documentation suggests that other performance indicators such as graduation and attendance rates have also been incorporated into the API calculation but it is unclear how this is done. By law, however, test results must constitute at least 60 percent of the API. For additional details, see API information guides for each year.

In contrast to the GPAP, which were awarded directly to schools, these awards targeted employees – certificated staff, i.e. teachers, in the first instance and both certificated and classified (i.e. paraprofessional, administrative and clerical) staff in the second. The SEPB did, however, grant half of its $350 million award to the schools for unrestricted uses.[11]

Like the GPAP, both awards were paid out based on API growth, although the $100 million CSPIA was allocated only to staff at schools that demonstrated the greatest growth over *twice* their GPAP target and had shown growth in the 2000 SY. The SEPB was paid to all schools (and staff in schools) that received the GPAP. Because only schools or the employees of schools that received the GPAP could have received these other two awards, our GPAP analysis below is sufficient to capture the combined effect of these award programs on achievement. We discuss the implications of these additional award programs for our analysis of resources in section 4.3.[12]

2.1 Award Eligibility – The Simple Case

One bit of complexity in California's accountability system is that performance awards are based on API "growth" scores – the year to year change in API – for a school as well as for each "numerically significant subgroup." Before describing what numerically significant subgroups are and how they affect eligibility, we first explain the award determination for a school without subgroups.

Award eligibility is based on a comparison of a school's growth score with its "target." In the simplest case, for schools without subgroups, the "API growth target" in a given year is five percent of the distance from the previous year's API to the statewide target of 800 or a specified minimum. In the 2000 SY, the minimum gain was set to one point; in the 2001 SY, it was raised to five points.

[11] The CDE provides very little information about the SEPB. Discussion of the sharing rules were found only in news reports, such as the one available here: http://www.svcn.com/archives/lgwt/04.04.01/education-0114.html

[12] Note, however, that because these programs were effectively suspended after the 2000 SY, our analysis of the 2001 SY, the second year of the awards program captures the effect of the GPAP alone. Results for the 2001 SY, which are available upon request are quite similar to those for the 2000 SY.

In other words, to receive an award based on 2001 SY performance, schools had to achieve the maximum of five percent of the distance to the statewide target of 800 or five points.

The 2000 and 2001 SY award decision rules can be expressed simply as

$$Target_{2000SY} = max(40 - .05 * baseAPI_{99}, 1) \quad (1)$$

$$Target_{2001SY} = max(40 - .05 * baseAPI_{00}, 5) \quad (2)$$

where $Target_t$ is the minimum gain score (or one year change in API) needed to qualify for an award in year t and $baseAPI_t$ is just the (adjusted) API from $t-1$.[13]

Figure 1 plots the 2000 and 2001 SY award targets and demonstrates several noteworthy features of California's awards program. First, although not made explicit in the official rules, gain scores are always rounded to the nearest integer and thus the awards eligibility thresholds are represented as a step function. Second, and perhaps most importantly, poor performing schools (i.e. schools with lower initial API scores) have to achieve larger test score gains to receive an award than do high achieving schools. Finally, the figure clarifies the effects of the minimum targets set in each year. In the 2000 SY, schools with base scores at or above 780 needed to gain only one point over their initial year score to qualify for an award. In the 2001 SY, award eligibility was contingent on a minimum gain score of 5 points. This change had the effect of uniformly raising the award threshold by 5 points for those at or above an API of 780 while increasing the target by the nearest integer value of $0.05 * baseAPI - 35$ for those with an API of 700 to 780.

[13]The California Department of Education adjusts base scores to make them "psychometrically" equivalent to the growth scores in the following year. In other words, in principle the base score in a given year, baseAPIt, can differ from the growth score in the previous year, APIt-1.

2.2 Award Eligibility with Numerically Significant Subgroups

The PSAA also mandates that "numerically significant" subgroups make "comparable achievement," defined as 80 percent of the school-wide growth target. Subgroups are defined by racial/ethnic categories (African American, American Indian, Asian, Filipino, Hispanic, Pacific Islander and Caucasian) and socioeconomic disadvantage.[14] A disadvantaged student must either qualify for free or reduced-priced meals or come from a family where the highest level of education is below high school completion. Subgroups with fewer than 30 tested students are not numerically significant. To achieve "numerical significance" a subgroup must have between 30 and 99 tested students and constitute at least 15 percent of total enrollment or have 100 or more tested students.

Table I documents the award eligibility calculation for two elementary schools with multiple subgroups in the 2001 SY). [15] The first column of Table I indicates the overall size of the school and the number of students tested in each subgroup.[16] Both schools have over 800 students enrolled, putting them above the 75th percentile of elementary school enrollments in the state. Both schools also have tested students in each of the state-defined subgroup categories but they differ in terms of which groups are sizable enough to be subject to performance targets. Neither school has tested numbers of American Indians, Filipinos, Asians, or Pacific Islanders above 30, exempting them from subgroup performance targets. African American students in Salida Union are also exempt since they number only 16. In contrast, since the tested numbers of Hispanics, whites, and socially

[14] While racial/ethnic subgroups are mutually exclusive, the socially disadvantaged category may contain students from the racial/ethnic subgroups.

[15] Data on academic performance for the 2000 - 2004 SYs as well as the monetary awards apportioned to schools under the GPAP for performance in the 2000 and 2001 SYs come directly from the California Department of Education (CDE). School characteristics come from the CDE's California Basic Educational Data System (CBEDS), an annual school-based census. From CBEDS, we collect data on student enrollment, the allocation of teachers across subjects, in addition to other demographic characteristics such as racial breakdown, parental education, and percent of students receiving free lunch. All of our data sources are described in more detail in the Data Appendix.

[16] Small schools, defined as having between 11 and 99 valid Stanford 9 test scores, as well as "very small schools" (fewer than 11 valid scores) were evaluated under a separate, "Alternative Accountability System".

disadvantaged students are greater than 100 in both schools, these groups must meet the subgroup performance targets. Tested African American students in Mission Elementary are also subject to subgroup rules as they number well over 30 and represent over 15 percent of tested students.

To gauge award eligibility, columns (2) and (3) show growth and base year API scores, respectively. Column (4) takes the difference between the two and column (5) calculates the score needed to qualify for an award. A school is award eligible if the gain score (column 4) equals or exceeds the target (column 5) for the school as a whole and for each numerically significant subgroup.

Based on school performance alone, both schools would have qualified for an award. This can be seen in the first row (labeled "Overall") in each panel. Specifically, Mission Elementary had a growth year score of 692 and a base year score of 682. Thus, it achieved growth of 10 API points, 4 points above the school-wide target of 6. Salida Union Elementary had a growth year score of 696, gaining 32 points on its base year score of 664 and 25 points on its target of 7. However, only Salida Union met both the school target and all of its subgroup performance requirements. Two (out of four) of Mission Elementary's numerically significant subgroups failed to meet their performance target of 5 API points (80 percent of the school target). The API for Hispanic students actually fell nine points and that for socially disadvantaged students (a category that may contain students from the racial/ethnic subgroups) stayed the same.

Since a school cannot qualify for an award unless all performance targets are met, for the purposes of the regression discontinuity design employed subsequently, we characterize each school by the *minimum* of the difference between the gain scores and targets for the school overall and each of its numerically significant subgroups. Thus, Mission Elementary, despite a passing performance overall, is assigned the Hispanic award gap of -14 points, its highest barrier to award eligibility. In contrast, all subgroups at Salida Union Elementary had gain scores that exceed their performance

targets. Socially disadvantaged students had the smallest "improvement" in scores, 18 points, and the minimum difference in gain score and performance target. Thus, we characterize Salida Union Elementary by an "award gap" of 12 API points, which qualified the school for an award.

2.3 GPAP Award Allocations

Table II describes GPAP allocations and performance overall and by award receipt status. The overall means, in the first column of the table, are based on all elementary, middle and high schools that met the testing participation requirements for the program (95 percent in elementary and middle schools and 90 percent in high schools) and had valid API scores for both base and growth years in the 2000, 2001 or both school years. The first column gives the overall means. The next four columns separate the data into schools that never won an award over the sample period, schools that won an award only for the 2000 SY, schools that won an award only for the 2001 SY, and schools that won an award for both school years.

As shown in the first row, about 23 percent of the sample is composed of schools that never won an award. About 31 percent of schools won awards for the 2000 SY alone. Due at least in part to the state's raising the award eligibility threshold, the share winning awards for 2001 SY performance alone is only 14.7 percent or about half of the corresponding 2000 SY figure. In contrast, almost 32 percent of schools won awards in both the 2000 and 2001 school years.[17] The average per pupil payment was about $63 across both years, a few dollars less than the average $69 per pupil payment indicated by the state.[18] To put this in perspective, K-12 public school expenditures in California in the 2000 SY year were roughly $6000 per student (Carroll et al. 2005). Thus, in principle, these

[17]Note that these figures are not strictly correct since the observations in the table are at the school-year not the school level. But, since there are roughly the same number of schools with valid testing data in both years, this is a good approximation. Expressed by school, 19.2 percent never won an award, 34.4 percent won for the 2000 SY alone, 14.5 percent won for the 2001 SY alone, and 31.9 percent won for both years.

[18]This small discrepancy between our and the official figures, however, is likely due to differences in the quality of enrollment data.

awards raised per pupil spending by just over 1 percent. More importantly, perhaps, awardees have considerable discretion (requiring only local school board approval) in how they use these funds. To the extent that these resources were paid to teachers, as the CDE has suggested, they amount to bonuses of almost $1400 per teacher. Moreover, as we will show in section 4.3, additional resources flowed to districts and thus presumably schools that qualified for awards. Thus, the official figures likely understate the true awards that winning schools received.

The fourth row of Table II shows school enrollments. The enrollment figures speak to a fundamental problem with relying on mean test scores to measure school performance (Kane and Staiger 2001; Chay et al. 2005). Specifically, due to large error variances, test scores from any given year provide a poor measure of school rankings. This problem is particularly acute for small schools, since, all else equal, their mean score will have higher sampling variation. The implication of this higher sampling variation is that small schools are more likely to have a particularly lucky year and win a performance award. Consistent with this, Table II shows that schools that won awards in both years are smaller ($p < 0.001$) and schools that never won an award larger ($p < 0.001$) than the average school. Furthermore elementary schools, which are the smallest type of school in our sample, with an average enrollment of 636 pupils, are underrepresented in the category of schools that never won awards and overrepresented among those winning awards in both years. While they make up 70 percent of schools in our sample, elementary schools represent only 49 percent of schools that never won an award and about 85 percent of those winning in both years. At the other extreme, high schools, which are by far the largest type of school in our sample, represent about 13 percent of the sample but almost 30 percent of schools that never won an award and only 3 percent of schools that won awards in both years.

The eighth and ninth rows of Table II explore the implications of the state's subgroup rules.

The typical school has only one subgroup; 15 percent of schools have no subgroups. The most common subgroups are socially disadvantaged, followed by Hispanic and white. Not surprisingly, since schools with subgroups face additional eligibility criteria, schools that never won an award have more subgroups and those that won awards in both years have fewer subgroups than the average school. The next row shows the share of schools in each category that lost an award due to subgroup rules. Put differently, this row shows the share of schools that would have won an award absent the subgroup rules. Overall, about 18 percent of schools would have won an award based on school performance alone but were ineligible because of subgroup rules. About 45 percent of schools that never won an award would have won without the subgroup rules. This average over the two award years masks differences across the two years attributable to the evolving awards eligibility criteria. Whereas 53 percent of schools in the never group would have won awards based on the school criteria alone for the 2000 SY, only about 38 percent would have won for 2001 SY performance were it not for the subgroup rules. In other words, raising the standards for schools had the effect of minimizing the bite of the subgroup rules.[19].

The next set of nine rows in Table II show the average API score across award years overall and separately for all numerically significant subgroups. The mean API score in the sample is 652. White, Asian and Filipino subgroups perform well above average. Socially disadvantaged subgroups, which are numerically significant in 9082 of our 10720 school-year observations, have an API of 581, almost two thirds of a standard deviation below the overall average. Hispanics are the next most common subgroup. They are numerically significant in 8462 school-years and have an API of 574, also well below the overall mean. Not surprisingly, since those with lower API scores

[19]This effect can also be seen by comparing the share of 2000 SY awards winners that lost awards in the 2001 SY with the share of 2001 SY awards winners that lost awards in the 2000 SY due to subgroup rules. While 20 percent of 2001 SY awards winner would have won awards in 2000 based on the school performance alone, only about 15 percent of 2000 SY winners would have done so in 2001

need to make larger gains in order to qualify for an award, mean API scores overall and within each subgroup are lowest for the set of schools that never won awards. Below the API scores are nine rows showing the average gain scores, the basis for award eligibility, across years for schools overall as well as for each subgroup. The average gain score is about 26 points but is less than 10 for schools that never won an award and just over 40 for schools that won in both years.

These summary statistics illustrate the differences in characteristics of schools who won awards versus those that did not, and highlight the need for a valid empirical design in evaluating the impact of award receipt on school behavior, ceteris paribus. Consequently, in the work that follows we will first demonstrate that there was in fact a treatment generated by the award program. We then try to determine whether award receipt had any effect on either the level of API scores (overall and for each numerically significant subgroup) or on school resources.

3 Econometric Framework

To identify the causal impact of California's awards program on outcomes such as API scores or resource allocations, we employ a regression discontinuity (RD) design where we essentially compare the behavior of schools that just barely won an award to those that just barely missed winning an award.[20] Our design is similar to the original RD approach used by Thistlewaite and Campbell (1960) to estimate the impact of a test-based scholarship award program on future academic outcomes except our unit of analysis is the school. Although individual student data would be appealing, particularly for estimating the impact of the awards program on achievement, these data are difficult, if not impossible, to obtain. Moreover, since the PSAA is ultimately based on average school performance, school-level data is sufficient for characterizing the program.

[20] More precisely, as shown in Lee (2005), we will estimate a weighted average of the population treatment effects, where the weights are positively related to each observation's distance to their awards target. Thus, schools closest to their awards threshold will contribute the most and those farthest away the least to the estimated treatment effect.

Suppose that the relationship between schools' average performance and resources and their receipt of an award is given by the constant treatment effects model:

$$Y_i = \alpha + \beta T_i + \gamma X_i + \epsilon_i \quad (3)$$

$$T_i = \mathbf{1}(D_i \geq 0) \quad (4)$$

where Y_i is school i's achievement score or measure of resources; α is a constant; and T_i is an indicator equal to 1 if school i received an award. In addition, let D_i equal the school's distance to its award eligibility target $((API_{it} - API_{it-1}) - Target_i)$, so that zero corresponds to having just met the target.

The primary challenge to estimating the effect of the award program β is that awards are not randomly allocated across schools. A simple comparison of schools that receive the treatment to those that do not will be biased because treated schools differ greatly from untreated schools for reasons other than the treatment. As our discussion above and Table II show, for example, schools with more minorities or subgroups are significantly less likely to receive an award.

To overcome this problem, we exploit the rules assigning treatment. Schools for whom $D_i \geq 0$ win the award; schools for whom $D_i < 0$ do not. The sharp discontinuity in the rules that translate test scores into award eligibility generates quasi-random assignment of award receipt near the eligibility threshold. To see how this works, consider the average outcome for treated schools with a specific value for their score ($D_i = \Delta > 0$):

$\mathrm{E}[\mathrm{Y}_i|T = 1, D_i = \Delta] = \alpha + \beta + \gamma E[X|D = \Delta] + E[\epsilon|D_i = \Delta]$.

Likewise consider a school that does not receive the treatment $D_i = -\Delta < 0$ whose score is Δ below the threshold:

$\mathrm{E}[\mathrm{Y}_i|T=0, D_i=-\Delta] = \alpha + \gamma E[X|D=-\Delta] + E[\epsilon|D_i=-\Delta]$.

A naive comparison – the average outcome for treated schools above the threshold by an amount Δ to those untreated schools who are below the threshold by an amount Δ to the average outcome for the untreated is merely:

$$\begin{aligned} E[Y_i|T=1, D_i=\Delta] - E[Y_i|T=0, D_i=-\Delta] &= \beta + \gamma \underbrace{E[X|D=\Delta] - \gamma E[X|D=-\Delta]}_{\text{Difference in Observables}} \\ &\quad + \underbrace{E[\epsilon|D_i=\Delta] - E[\epsilon|D_i=-\Delta]}_{\text{Difference in Unobservables}} \end{aligned}$$

Now consider choosing Δ to be small so that we are comparing schools just above the threshold to those just below. In the limit, as $\Delta \longrightarrow 0$ the above expression reduces to:

$$\begin{aligned} \lim_{\Delta \longrightarrow 0} E[Y_i|T=1, D_i=\Delta] - E[Y_i|T=0, D_i=-\Delta] &= \beta + \gamma E[X|D=0] - \gamma E[X|D=-0] \\ &\quad + E[\epsilon|D_i=0] - E[\epsilon|D_i=-0] \\ &= \beta \end{aligned}$$

That is, as we approach the threshold from the left and the right, both the unobservable and observable differences of treated schools become smaller and smaller. This research design provides a large number of *testable* restrictions, which are similar to those available in a randomized controlled trial (RCT), and add to the credibility of the design. In particular, in an RCT a basic specification test is to compare the average baseline characteristics of treated group to the control group. In a proper RCT these will be the same on average. Likewise, in this research design, schools just to the left and just to the right of the threshold should look similar.

As a practical matter, it is not necessary to limit the comparison to just the few schools to

the left and the right of the threshold. One can recast the problem as estimation of the following relationship:

$$Y_i = \alpha + \beta T_i + g(D_i) + \epsilon_i$$

where $g(\cdot)$ is a unknown continuous function. Although unknown it can be approximated sufficiently well by polynomials in D and its full interactions with the awards indicator T. In practice below, we determined that the fifth-order polynomial was the most parsimonious specification implied by the underlying data.[21]

A minor issue is the fact that we have only considered the causal impact of the binary treatment. In fact, there are different levels of the award, or varying treatment intensities. This can be accommodated easily by recasting the problem as a simple instrumental variables estimator for the following equation system:

$$\begin{aligned} Y_i &= \alpha' + \theta A_i + g'(D_i) + \epsilon_i' \\ A_i &= \alpha + \psi T_i + h(D_i) + \nu_i \\ Y_i &= \alpha + \beta T_i + g(D_i) + \epsilon_i \\ &= \alpha + \theta\psi T_i + g(D_i) + \epsilon_i \end{aligned}$$

In this set up, A_i is the endogenous regressor, the second equation is the "first stage" equation, and the third equation is the "reduced form" equation for the outcome. Our estimate of θ is merely the indirect least squares estimate $\frac{\widehat{\beta}}{\widehat{\psi}}$ – the ratio of the discontinuity in the outcome equation to

[21] To illustrate this, our graphs below superimpose the predicted values from the polynomial on top of the means for each discrete point of support. We explain these graphs and other specification tests below.

the ratio of the discontinuity in the award equation. If we wish to consider the case where the treatment effect is random, then provided that monotonicity holds (the effect of higher growth scores is uniformly to increase the probability of receiving treatment) then this parameter identifies the local average treatment effect or the effect of the program on those schools induced to win the award by their score (Hahn, Todd, and van der Klauuw 2001). Finally, just as in an RCT, provided that the X's are balanced (a restriction we test), we can also include exogenous covariates X for variance reduction purposes, although they are not required for consistent estimation of the parameter of interest.

3.1 Estimation Issues

In order to implement the RD, we need to first verify that the Governor's Performance Award Program rewarded API growth in both according to the established rules. To cut down on the number of figures, we only show results for the 2000 SY. Results for the 2001 SY are similar and are available upon request.

Figure 2 shows the share of schools receiving an award payment for their 2000 SY Performance at each distance from the eligibility threshold. Average award payments expressed per pupil and per teacher for schools at each distance from the eligibility threshold are shown in Figures 3a and 3b, respectively. Across figures, the solid lines represent a parametric estimate of the conditional probability of an awards payment (Fig. 2) or of the per pupil (Fig. 3a) or per teacher (Fig. 3b) payment amount to schools at each distance, D, from the eligibility threshold. Each school's distance to its eligibility threshold is just $((API_t - API_{t-1}) - Target_t)$ or the difference between its gain score and growth target as defined above. Consistent estimation of the treatment effect requires that our polynomial be sufficiently flexible to capture the true underlying continuous function.

Operationally, our parametric estimates are just the least squares fitted values from the following equation:

$$A = \delta T + P'\alpha_0 + TP'\alpha_1 + X'\beta + \varepsilon \tag{5}$$

where A is either the probability of awards recipiency or the per pupil award payment, $T \equiv 1(D \geq 0)$ is an indicator for whether a school crossed the eligibility threshold, $P' = (D, D^2, D^3, D^4, D^5)$ is a fifth order polynomial of the distance, D, to the awards threshold and TP' is the interaction of our eligibility indicator with this fifth order polynomial. We include the interactions so as to allow the polynomial fit to differ on either side of the eligibility threshold.[22]

One potential problem, more apparent than real, is that our polynomial provides an inadequate parameterization of the true underlying continuous function. While in principle we could use nonparametric local linear regression or other techniques, as Lee and Card (2008) observe, such techniques confront the difficulty that the underlying data (changes in test scores) are not actually continuous but rather discrete. In this case, the "true" nonparametric estimator is just the set of mass points in the running variable.

Consequently, as we explain in more detail below, we assess the adequacy of our parametric representation by comparing our model to the fully saturated model that includes a separate indicator for every specific value of the running variable:

$$A = d\sum Z_d\gamma_d + X'\beta + \mu \tag{6}$$

[22]For variance reduction purposes, we also include X, a set of control variables that include: a school's total enrollment, the number of numerically significant subgroups in the school, the percent of tested students by race/ethnicity (white, black, Hispanic, Filipino, Asian, Pacific Islander, or American Indian), the share of students qualifying for free or reduced price meals, and dummies for whether the school is an elementary or high school (with middle school the omitted category). All covariates correspond to the academic year of the growth year score, i.e. t not $(t-1)$. Standard errors are adjusted to allow for an arbitrary correlation in errors at the level of D, the distance to the awards threshold, as suggested by Card and Lee (2006) to account for potential misspecification.

where Z_d is a dummy variable that equals 1 if the school's distance is equal to d, and 0 otherwise, γ_d are fixed effects for each distance to (and including) the awards eligibility threshold, and X is the set of covariates defined above. The coefficients γ_d, which are just regression adjusted average award payment probabilities or per pupil award payments, are represented as open circles in our figures below. Plotting these coefficients along with our parametric fit allows for a simple "eyeball" test of the extent to which our estimates are a spurious consequence of "noise" in the data. Below, we also describe and present the results of more formal tests.

Figure 2 shows that, as per the award rules, there is a marked discontinuity in the probability of receiving an award at the eligibility threshold. Schools that failed to meet their target, and thus are to the left of the eligibility threshold, are not observed receiving an award payment.[23] At the threshold, where a school's API equals its target, the probability of receiving an award jumps to almost one. The estimated discontinuity is 0.93 with a t-stat of almost 80. Both the regression adjusted averages and the polynomial fit past zero are also strictly below one because a small fraction of schools, about 8 percent in the 2000 SY, made their API target but did not receive an awards payment. According to the California Department of Education these schools may have been disqualified because of "data irregularities," too high a share (over 15 percent) of parents obtaining exam waivers for their children, or student population changes that invalidated the school's API.

Figures 3a and 3b are analogous to Figure 2 except that the dependent variable is the per pupil and per teacher award payments rather than a simple dichotomous measure of recipiency. We show the payments in per teacher terms in light of evidence suggesting that the awards were

[23] In actuality, in the 2000 SY, 5 schools or 0.3 percent of schools that according to our data failed to meet their target received awards payment from the State averaging $60.5 per pupil. In 2001, none of the "failing" schools received GPAP payments.

paid out as cash bonuses to teachers. As expected, schools to the left of the eligibility threshold did not receive award apportionments and thus have per pupil award payments of \$0. At the discontinuity the per pupil award payments jump sharply. The estimated discontinuity based on 2000 SY performance is about \$62 per pupil with a t-stat of 80. Expressed as a per teacher award, the estimated discontinuity is about \$1300 with a t-stat of over 70.

A visual comparison of the estimates from our parametric models and the regression adjusted averages of award recipiency suggests that the fifth-order polynomial fits are reasonable. We confirm this using a more formal test. Following Lee and Card (2008), we calculate a goodness of fit statistic, $G \equiv \frac{(RSS_r - RSS_{ur})/(J-K)}{RSS_{ur}/(N-J)}$, where RSS_r is the residual sum of squares from the restricted (polynomial-fitted) model, RSS_{ur} is the residual sum of squares from the fully flexible or unrestricted model, J is the number of parameters in the unrestricted model, K the number of parameters in the restricted model and N the number of observations. Under normality G is distributed $F(J-K, N-K)$. With this F-statistic we can test the null hypothesis that the fit of the polynomial model is as good or has as much explanatory power as the fully flexible model. Across all our awards recipiency models (any award, award per pupil and award per teacher), G is less than one (0.773, 0.764, and 0.790, respectively). In all cases, the F-statistics indicate that we cannot reject the null that the restricted and unrestricted models have similar goodness of fits.

The framework used above to establish the discontinuity in the GPAP, is the same estimating equation we employ to determine the causal impact of the program. More specifically, we will use (4) to estimate the treatment effect on Y (instead of A), and we will use (5) to test the sensitivity of this estimate to our functional form assumptions. The ratio of the RD estimate of our outcome of interest, say math teachers or spending per pupil in year $t+1$, to the RD estimate of award recipiency or per pupil award apportionments in t, will form our causal estimate of the treatment

on the treated.

One concern in interpreting this ratio as a causal estimate, however, is that other predetermined school characteristics may be changing at the same time. As discussed above, identification of δ requires that X is continuous at $d = 0$. For example, if schools near the discontinuity encourage certain types of students to transfer to other schools, we may see changes in the share of students by race or socioeconomic status. Since these factors independently affect outcomes, they will, at least in principle, confound our estimates of the treatment effect of the awards program.

The fact that the state uses API *changes*, which should be considerably noisier and more difficult to manipulate than levels, to allocate awards should lend considerable credibility to our research design. But, while we cannot prove that all other predetermined characteristics are balanced, we can check to see whether observable characteristics are smooth through the discontinuity. To do this, we have plotted regression adjusted averages of and estimated polynomial fits to the 2000 SY values of all covariates described above as well as some other available characteristics, against the distance to the threshold for schools.

Appendix Table I reports the estimated discontinuities at the 2000 SY awards threshold for each of these predetermined characteristics. In 12 out of 14 cases, the estimated discontinuity is not statistically distinguishable from zero at even the 10 percent level. Thus, by in large, the estimated discontinuities reported in Appendix Table I provide support for the idea that schools close to the eligibility threshold are similar on predetermined characteristics.

Two cases merit some additional discussion. We estimate a small discontinuity in the percent of students qualifying for free or reduced price meals and the percent of tested students that are Asian American. Neither are significantly different from zero at conventional levels (i.e. 5 percent), but the p-values for each are only about 0.07. For free or reduced price meals, the point estimate

implies a 4.7 percentage point or roughly 10 percent drop in the share of students qualifying in schools that just received awards relative to those schools that just missed receiving an awards. For the share of test-takers that are Asian American, the implied effect is a 1.8 percentage point or an almost 23 percent increase. Importantly, we find no evidence of a discontinuity in the share of students eligible for free or reduced price meals or in the share of test takers that are Asian American in the 2001 SY (available upon request). The fact that those characteristics for which we cannot strictly rule out a discontinuity differ across the two award years, gives us some hope that they are generated by random variation. As is well understood, with a large number of comparisons a small fraction of these are expected to be "significant" by sheer random variation.

To the extent that these discontinuities are real, however, they suggest that our estimates may be slightly biased towards finding an impact of the awards program on academic achievement. For example, students qualifying for the school meals program, who are automatically characterized as socially disadvantaged, are more likely to perform poorly (see the API scores for socially disadvantaged subgroups in Table II). Similarly, African American students perform below average as a subgroup. Asian Americans, by contrast, perform well above the state average. Thus, at the discontinuity, a drop in the share of students qualifying for free or reduced price meals and a bump up in the share of Asian Americans might lead us to overstate any change in test scores (or other positive outcomes) associated with the 2000 SY awards program. Fortunately, our estimates also suggest that to the extent that such a bias exists, it will be small as long as higher scores do not induce some schools who – in the absence of their higher score – would have received an award, to in fact be denied an award. This so-called "monotonicity" condition appears to be reasonable in our context.[24]

[24]For a discussion, see Lee (2005.)

4 Results

Having established that the data support the validity of our research design, we next discuss our estimates of the impact of the award program on achievement and resources. As above to minimize redundant plots, we only provide figures for 2000 SY awards program. In all tables, however, we present estimates of the magnitude of the discontinuity, its standard error, and the F-test of the correspondence between our polynomial fit and the fully flexible model for both award years.

4.1 Evidence on Achievement

We first consider the impact of the awards program on the level of achievement. If schools that win awards are able to spend these resources in ways that positively impact achievement, then we should see a break in API scores at the discontinuity. In other words, schools that just barely won the awards should have higher scores in subsequent years than their counterparts that just barely missed winning an award.

Figures 4a, 4b, and 4c graphically represent our RD estimates of the impact of the 2000 SY awards program on achievement levels in 2001, 2002, and 2003. Figures 5a, 5b, and 5c present corresponding plots for the socially disadvantaged subgroup, the most common of all the subgroups. Because the first apportionment for the 2000 SY awards program was made in January 2001, the middle of the 2001 academic year, and the second and final payment in March 2002, the following school year, we do not anticipate finding any impact on achievement in 2001 (as measured by test scores in May 2001). [25] Thus, it is reassuring that, as shown in Figure 4a, 2001 school-wide API scores are smooth across the awards eligibility threshold. The same basic pattern holds for the 2001 API scores for socially disadvantaged subgroups.

[25] See http://www.cde.ca.gov/ta/sr/gp/history.asp for details on the history of awards apportionments. Note that there were other awards programs for the 2000 SY but these were paid to teachers and not schools.

To the extent that the additional resources were put towards instruction, as the CDE intended, we might expect a bump up in achievement in the 2002 or 2003 school years. But, Figures 4b and 5b also indicate that scores were smooth across the award threshold. Moreover, the close correspondence between the polynomial fits to the API (the solid lines on either side of the awards target) and the regression adjusted average API scores at each distance from the threshold (the open circles) suggest that our estimates of the discontinuity in API scores (or lack thereof) is not an artifact of our modeling choices. This is confirmed by the F-statistics reported in Table III. The estimates from 2002 are considerably noisier. But, the graphical analysis and the estimates provided in Table III broadly confirm a finding of no effect of the awards program on API scores. Similarly, as shown in Panel B of Table III, we find no evidence of an impact of the 2001 SY awards program on schoolwide or subgroup achievement. In no case are the estimates of the impact of the 2001 SY award program on API scores significantly different from zero.

We also examined alternative measures of academic achievement to test for an impact of the awards program. Specifically, we use measures of the percent of students in a school that tested proficient in English and language arts and in mathematics. These data, which are first available to us in 2001, are based on the California Standards Tests (CSTs) for grades 2-8 and the California High School Exit Examination (CAHSEE) for secondary school students. While these scores have been incorporated into the API over time – making them imperfect complements to the API – they have the advantage of being reported in a very transparent form.[26] Yet, for neither the 2000 nor 2001 SY awards program can we detect any evidence of improvements in either the percent of students testing proficient in ELA or math for schools that just qualified relative to schools that just missed qualifying for an award. Analyses using API or proficiency gain scores, noisier measures

[26]The CST in English and language arts was added to the API measure in 2002 SY. And, the CST in math and the CAHSEE were added to the API in 2003 SY.

of achievement, yield similar conclusions.

Finally, we also focus on the API scores of the subgroup that determined awards eligibility within the school. The idea is that since this subgroup is the worst performing group in the school, award resources in the following periods may be targeted to them. However, similar to previous analyses, we find no significant impact of the awards on these subgroup-specific outcomes.

4.2 Evidence on School Resources

One reason for these null findings, other than the simple possibility that resources do not translate directly or easily into academic improvements, may be that GPAP funds were not used for instruction. There are several ways this could happen. First, it is possible that districts, which have fiscal authority over schools, or the state, which provides much of the funds to districts, offset the awards paid out by the GPAP through corresponding reductions in other funds.[27] Under this scenario, award schools might have no additional financial resources to invest in achievement. Second, even if neither districts nor the state undo the monetary transfer required by the GPAP, schools are not constrained to spend these funds on instruction. In fact, they have almost complete discretion over the use of GPAP funds, needing only the approval of the local school board. Thus, to the extent that awards are spent on non-instructional staff, capital outlays, and so on, we may not expect to see any improvements in academic achievement.

Unfortunately, fiscal data on revenues and expenditures are not available at the school level. Rather, they are reported at the district level. This limits our ability to determine whether an individual school receives its award money from the district and, if it does, how it gets spent. We

[27] Baicker and Jacobson (2006) provide evidence that counties engage in this type of budgetary offsetting by reducing allocations to police agencies that receive financial rewards from state or federal government for drug enforcement activities. Similarly, Gordon (2004) finds that increases in federal spending on low income school districts are offset by reductions in local spending.

can, however, observe school-level inputs such as the number of teachers per pupil overall, the share allocated to each subject (math, English, science, physical education, and special education) as well as the number of instructional computers per pupil and internet-connected classrooms per 100 students.

We have estimated the impact of both the 2000 and 2001 SY awards program on each of these inputs for the 2001-2003 school years. Table IV presents a subset of these estimates. Panel A presents our estimates of the impact of the 2000 SY award program on teachers per pupil, the share devoted to math instruction, the share devoted to English instruction, computers per pupil and internet connections per 100 students in the 2001 academic year. Panel B presents estimates of the 2001 SY award program on the same category of outcomes but for the 2003 academic year. We consider this year because the 2001 SY award disbursements were not made until July and October 2002.

We find little evidence of any changes in these inputs. Given that the award payments were a one shot deal and hiring requires a long term fiscal commitment, it may not be surprising that we find little impact of either the 2000 or 2001 SY award program on the number of teachers per student. On the other hand, we might expect to see a change in the allocation of teachers across subjects, if, for example, increased funds could be used to encourage some instructors to switch from their normal subject to one that is more valuable in an accountability system. But, our estimates of the impact of the award programs on teachers per pupil and the share of teachers in math or English are neither statistically nor economically significant.

A more likely use of the award funds might have been for instructional resources. Indeed,the California Department of Education anticipated that these funds should be used "for the purchase of

computers, instructional materials, or playground improvements."[28] But, like the CDE, we find no evidence that award payments were used to increase the number of computers or internet connected classrooms within a school. If anything, the results for the 2001 SY award program suggest that computers per pupil increased less among those schools that just qualified for awards than those that just missed qualifying. But, this interpretation of the point estimate for computers per pupil (-.020 with a standard error of .006) should be viewed with considerable skepticism. As our goodness of fit statistic and the corresponding p-value (.0002) suggest, the polynomial fit for this model is poor relative to the fully flexible model. Thus, the estimated reduction in computers per pupil for awardees relative to nonawardees is likely driven by functional form. The more important lesson from Table IV is that we find little evidence of increases in computers per pupil or any other measure of resource allocations among schools that received the GPAP.

4.3 Evidence on Fiscal Outcomes

We next turn to the fiscal data. Although we cannot observe changes in an individual school's revenue or expenditures, we can determine whether the state offsets or alternatively disproportionately increases funds to districts that just barely qualified for the GPAP relative to those that just missed qualifying. And, if we find no evidence of offsetting, we should be able to trace out where (district-wide) any additional funds get allocated.

In order to characterize where each district lies relative to the school-level award eligibility threshold, we sort all of its schools by their distance from this threshold. We then assign to each district the *maximum* of the school-level "award gap." Recall that because of the subgroup rules we characterize each school's "award gap" as the minimum difference between the gain scores and

[28] See page 4 of the statements from Patrick J. Chladek at the November 2002 National Conference on Teacher Compensation and Evaluation, http://www.wcer.wisc.edu/CPRE/conference/nov02/chladek.pdf.

targets for the school overall and each of its numerically significant subgroups. Thus, each district "award gap" is the across-school maximum of the within-school minimums. This definition of the district "award gap" then picks up whether any school in the district received treatment. It further characterizes the district according to the distance to the award eligibility threshold of the school that performed best relative to this target. If any treated school in a district was far from the award eligibility threshold, then the district as a whole will be characterized as far from the cutoff. If no schools were treated but at least one was close to its target, then the district as a whole will be characterized as just barely missing award eligibility.

Figure 6 is the district level analogue to Figure 3a. It shows the mean district-level apportionment per pupil in 2001 by proximity to the awards threshold (open circles) as well as the polynomial fits to these data.[29] The figure (and the results reported in the first column of both panels in Table V) provide confirmation that we can still uncover the treatment at the district level. Because any given district may contain schools that won awards of varying per pupil amounts as well as schools that did not win awards, the estimated discontinuity in apportionments is below that for the school-level. For both 2000 and 2001, districts to the left of the eligibility threshold did not receive awards apportionments and thus have per pupil award payments of (roughly) $0. At the discontinuity the per pupil award payments jump to about $42 per pupil (with a t-statistic of about 12) for 2000 and about $28 per pupil (with a t-statistic of 9) pupil for 2001 test performance.

We next turn to district revenues and expenditure data. According to the School Fiscal Services Division of the CDE, GPAP apportionments are classified as unrestricted revenues. The jump in per pupil revenue at the awards threshold is, in fact, much larger than the $42 apportionment

[29]To save degrees of freedom, we estimate polynomials with equal slopes on either side of the polynomial. We do not include covariates as these tend to decrease rather than increase the precision of the district-level estimates. Our F-tests suggest that these fits are still quite close to the fully flexible model.

shown in Figure 6. The RD estimate (reported in Table V) suggests a jump of $104 per pupil (with a t-statistic of 21) at the discontinuity. Why should this be so much larger than the apportionment estimate? Other funds are also included in the unrestricted revenue category.[30] But to the extent that our RD provides quasi-random assignment, these funds should not differ systematically at the awards eligibility threshold except through the (direct and indirect) effects of the award program. Thus, the RD estimate in Table V implies that for every per pupil dollar a district was supposed to get through the 2000 awards program, the district, in fact, received closer to $2.50 per pupil (with a t-statistic of over 4). Some of these additional funds may be attributable to the $350 million Schoolsite Employee Performance Bonus program. Bonuses, which were only available for 2000 performance, were allocated based on the number of FTEs in schools winning the GPAP and were divided equally between a school and its employees. We were not able to locate any documentation of disbursements under SEPB, however. Any additional funds not attributable to this program may simply be evidence of crowd-in. In other words, the state may be further padding the budget of districts with award winning schools. We find no evidence, based on this revenue category, that such activity continued in 2002 or 2003.

For the 2001 awards program, we do not detect any increase in unrestricted revenues until 2003. This makes good sense since apportionments for the 2001 awards program were not made until the 2002-03 fiscal year. More importantly, as shown in Table V, the RD estimate suggests that the unrestricted revenues only increase by about $20 per pupil (with a t-statistic of 3) at the discontinuity. This implies that for every per pupil dollar a district was supposed to get through the 2001 awards program, they, in fact, received only about $0.72 per pupil (with a t-statistic of almost 3). We cannot reject that this point estimate is significantly different from one. Thus, the

[30]Specific sources of revenue within this category are not available from district-level fiscal data.

results for the 2001 program, when the only monetary awards at stake were from the GPAP, are also inconsistent with crowd-out.

To see if the apparent increase in unrestricted per pupil revenues associated with the 2000 award program and the decrease associated with the 2001 award program were in fact real, we next consider total revenues per pupil. This will help us get around any reporting problems as well as allow us to estimate the net effect of the awards program on revenues. For the 2000 award program (reported in Table V), we estimate a jump in total per pupil revenues in 2001 of about $340. This estimate is only statistically distinguishable from zero at the 10 percent level. Taken literally, it suggests that for every $1 per pupil increase in funds due to the 2000 awards program a district actually received closer to $8 per pupil in revenues. This point estimate is only statistically distinguishable from zero at about the 11 percent level. Moreover, some of the apparent crowd-in may be driven by the additional awards programs that were in place in the 2000 school year. For example, $100 million was paid to teachers in the form of bonuses as part of the 2000 Certificated Staff Performance Incentive Awards Program. Because the eligibility target for this program was twice that for the GPAP, only schools that qualified for the GPAP could have received it. It is important to point out, however, that these awards were not classified as unrestricted revenue (but rather had a category of its own). Thus, this cannot explain the more than dollar for dollar increase in unrestricted per pupil revenues.

The estimate for the 2001 award program, while quite imprecise, implies an increase in total per pupil revenue in 2003 of about $123 per pupil. In other words, contrary to what we might conclude based on the unrestricted revenue alone, this estimate suggests that total per pupil revenues in 2003 may have also increased more than dollar for dollar as a result of the 2001 award program.

Total 2001 expenditures per pupil also appear to jump by about $343 dollars in response to the

2000 awards program. Similarly, 2003 expenditures per pupil increase by about $200 in response to the 2001 awards. This estimate, however, is not statistically significantly different from zero at any reasonable level of significance. Together, however, these estimates rule out the possibility of significant district-level crowd-out, and raise the possibility of significant crowd-in, as a result of California's Governors Performance Award Program.

Since anecdotal evidence suggests the awards ended up mainly as teacher bonuses, and to further illustrate that the flow of resources due to this program was quite substantial, Figures 7, 8, and 9 plot the equivalent RD estimates in per teacher terms relative to the 2000 award threshold. Figure 7 shows the mean district-level award revenue per teacher in 2001 by proximity to the awards threshold (open circles) as well as the polynomial fits to these data. Similar to Figure 6, this provides confirmation that we can still uncover the treatment at the district level. Figures 8 and 9 also rule out district-level crowd-out and instead raise the possibility of crowd-in.

5 Summary and Conclusion

In this study, we focus on a relatively understudied feature of accountability systems - the productivity of financial rewards for schools making "adequate" progress on state achievement exams. We analyze the case of California, where for the 2000 and 2001 school years, schools that met or exceeded their accountability targets were eligible to receive monetary awards through the Governor's Performance Award Program (GPAP). In addition, for 2000, teachers and staff in winning schools were also eligible for the Certificated Staff Performance Incentive Award (CSPIA) and the Schoolsite Employee Performance Bonus (SEPB). Because these awards were allocated based on a discontinuous function of school (and subgroup) test scores, we employ a regression-discontinuity design to evaluate how they affected achievement and resources. This design allows us to take

advantage of the possibility (verified in our data) that schools close to the eligibility threshold are similar but for award receipt. In this framework, award receipt close to the eligibility threshold is "as good as randomly assigned," much like in an actual randomized controlled trial.

We find that the programs did have a significant impact on the financial resources allocated to some schools and its staff. In 2000, the average value of the GPAP award was 60 dollars per student and in 2001 about 50 dollars per pupil. More importantly, the financial rewards from the GPAP appear to have been supplemented by payments from other award programs (CSPIA and SEPB in 2000) and state budgetary discretion (in 2000 and 2001). At its peak in 2000, districts with schools that qualified for GPAP awards received budgetary increases totaling about 5 percent of per pupil spending. Since anecdotal evidence suggests the GPAP awards were mainly distributed as teacher bonuses, teachers at winning schools could have earned up to $27,000, although we calculate $2,000 per teacher was probably more typical.

Despite the increase, we find little measurable improvement in standard metrics of achievement, such as exam performance, for those schools that received the award compared to those schools that did not receive the award. This is perhaps not surprising, as Project STAR, which increased resources by about 50%, yielded improvements in exam performance of less than a quarter of a standard deviation (Schanzenbach 2006). Moreover, because the resources from California's program are more akin to a random shock than a guaranteed income stream, schools may not have been able to incorporate them into projects that determine educational achievement. However, we also find no increase in "capital expenditures," such as computers or internet connections, which should be more responsive to a one-time shock.

It is also possible that the instability in the funding of the program weakened its incentives and failed to act as a strong signal of reward for teacher and school administrator effort. Because the

awards ended up being distributed mostly as bonuses to teachers (and possibly support staff as well) in an effective school regardless of their individual contribution, the group-based incentives in the GPAP can also have the free rider problem. The free rider problem could give rise to teachers no longer exerting as much effort after award receipt, in the period when we evaluate the schools.[31]

Our estimates show that accountability "on the cheap" had no significant impact across schools that won awards versus those that didn't. On the other hand, we cannot assess if the program would have a significant impact if implemented in conjunction with other reforms, such as reduced class sizes or raising teacher salaries.[32] This also leaves the question of whether the competition for the awards itself raised student achievement across all schools in California. However, a comparison of 1996-2000 4th and 8th grade Math NAEP scores shows California declined or performed the same as the rest of the country during this period.[33]

Finally, our findings also suggest that California's subgroup rules have had the (unintended) consequence of making diverse schools as well as schools that serve disadvantaged populations, more likely to fail their accountability targets. Because meeting these targets are tied to financial rewards, subgroup rules have had the unintended consequence of putting these schools at greater risk of receiving relatively fewer resources.

References

Baicker, Katherine and Mireille Jacobson, 2007. "Finders Keeper: Forfeiture Laws, Policing Incentives and Local Budgets," *Journal of Public Economics*, 91(11-12): 2113-2136.

Carroll, Stephen J., Cathy Krop, Jeremy Arkes, et al., 2005. "California's K-12 Public Schools:

[31]This is in contrast to recent work on the impact of paying bonuses to teachers based on schoolwide performance, for instance Lavy(2002).

[32]California implemented class size reduction beginning in 1996-97 at a cost of up to $850 per student. While class size reduction is several times more costly than the GPAP, there is no evidence that the program had a significant impact on student achievement. However, evaluating the program is fraught with empirical difficulties. See CSR Research Consortium (2002).

[33]Source: NCES, US Dept of Education.

how are they doing?" RAND Education Report.

Carnoy, Martin and Susanna Loeb, 2002. "Does External Accountability Affect Student Outcomes? A Cross-State Analysis," *Education Evaluation and Policy Analysis*, 24(4): 305-331.

Chay, Kenneth Y., Patrick J. McEwan, Patrick and Urquiola, Miguel, 2005. "The Central Role of Noise in Evaluating Interventions That Use Test Scores to Rank Schools," *The American Economic Review*, 95(4): 1237-1258.

Clotfelter, Charles T. and Helen F. Ladd, 1996. "Recognizing and rewarding success in public schools," in Helen F. Ladd, ed. Holding Schools Accountable: Performance-Based Reform in Education. Brookings Institution, Washington, D.C.

CSR Research Consortium, 2002. "What Have We Learned About Class Size Reduction in California?" George W. Bohrnstedt and Brian M. Stecher, eds. September 2002.

Eberts, Randall, Kevin Hollenbeck, and Joe Stone, 2002. "Teacher performance incentives and student outcomes," *Journal of Human Resources*, 37(4): 913-927.

Figlio, David N. and Lawrence W. Kenny, 2007. "Individual teacher incentives and student performance," *Journal of Public Economics*, 91(5-6): 901-914.

Figlio, David N. and Cecilia E. Rouse, 2006. "Do Accountability and Voucher Threats Improve Low-Performing Schools," *Journal of Public Economics*, 90(1-2): 239-255.

Glewwe, Paul, Nauman Ilias, and Michael Kremer, 2003. "Teacher Incentives" NBER Working Paper 9671.

Gordon, Nora, 2004. "Do Federal Grants Boost School Spending: Evidence from Title I," *Journal of Public Economics*, 88(9-10): 1771-1792.

Hahn, Jinyong, Petra Todd and Wilbert Van der Klaauw, 2001. "Identification and Estimation of Treatment Effects with a Regression–Discontinuity Design" *Econometrica*, 69(1), January, pp. 201-209

Hanushek, Eric A., 2003. "The Failure of Input Based Schooling Policies," *The Economic Journal*, 113 (February): F64–F98.

Hanushek, Eric A. and Margaret E. Raymond, 2005. "Does School Accountability Lead to Improved Student Performance," *Journal of Policy Analysis and Management* , 24(2): 297-327.

Hanushek, Eric A. and Margaret E. Raymond, 2003. "Improving Educational Quality: How Best to Evaluate Our Schools," in Education in the 21st Century: Meeting the Challenges of a Changing World.

Hoxby, Caroline, 2002. "The Cost of Accountability." NBER Working Paper 8855.

Jacob, Brian 2006. "Test-Based Accountability and Student Achievement: An Investigation of Differential Performance on NAEP and State Assessments." NBER Working Paper 12817.

Jacob, Brian and Levitt, Steven 2003. "Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating." *Quarterly Journal of Economics*, 118(3): 843-877.

Jacob, Brian and Levitt, Steven 2003. "Catching Cheating Teachers: The Results of an Unusual

Experiment in Implementing Theory." *Brookings-Wharton Papers on Urban Affairs*, pp. 185-209.

Kane, Thomas J. and Douglas O. Staiger, 2003. "The Promise and Pitfalls of Using Imprecise School Accountability Measures," *Journal of Economic Perspectives*, Fall, pp 91-114.

Kane, Thomas J. and Douglas O. Staiger, 2003. "Unintended Consequences of Racial Subgroup Rules," in Paul Peterson and Martin West, eds., *No Child Left Behind? The Politics and Practice of School Accountability* (Washington, DC: Brookings Institution Press, 2003), pp 152-176.

Ladd, Helen F. 1999. "The Dallas school accountability and incentive program: an evaluation of its impacts on student outcomes," *Economics of Education Review*, 18(1): 1-16.

Lavy, Victor 2002. "Evaluating the Effect of Teacher's Group Performance Incentives on Pupil Achievement," *Journal of Political Economy*, 110(6): 1286-1317.

Lavy, Victor 2003. "Paying for Performance: The Effect of Individual Financial Incentives on Teachers' Productivity and Students' Scholastic Outcomes," Mimeo.

Lee, David S. and David Card, 2008. "Regression Discontinuity Inference With Specification Error," *Journal of Econometrics*, 142(2): 655-674.

Lee, David S., 2008. "Randomized Experiments from Non-random Selection in U.S. House Elections," *Journal of Econometrics*, 142(2): 675-697.

Rogosa, David, 2003. "Four-peat: Data Analysis Results from Uncharacteristic Continuity in California Student Testing Programs," Unpublished Paper, September 2003.

Schanzenbach, Diane, 2006. "What Have Researchers Learned from Project STAR?," *Brookings Papers on Education Policy*, pp 206-228.

Thistlethwaite, D., and D. Campbell, 1960. "Regression-Discontinuity Analysis: An alternative to the ex post facto experiment." *Journal of Educational Psychology*, 51: 309-17.

A Academic Performance Index (API) Database

Established as part of California's Public Schools Accountability Act of 1999 (PSAA), the API is calculated by the California Department of Education annually for all public schools. These data can be downloaded from http://api.cde.ca.gov/datafiles.asp. Small schools, defined as having between 11 and 99 valid test scores, as well as "very small schools" (fewer than 11 valid scores) are evaluated under a separate "Alternative Accountability System." Growth targets are not calculated by the Department of Education for these schools.

The API ranges from 200 to 1000 and reflects a school's performance on various standardized tests. Students in grades 2 to 11 are tested. The tests and indicators used and weights accorded to each to calculate the API vary from year to year. Beginning with the 2004 API base, the weights are applied at the individual student test level; prior to this, it was applied at the school level. The State Board of Education (SBE) "recognizes that the question of the appropriate test weights is a policy issue rather than a technical issue"; its members adopted test weights they believe "reflect the curriculum priorities in California public education." In practice, each school's content area weights are determined based on the test weights established by the SBE and also on the number of valid test scores in each content area and grade level at a school. Because of this, API calculations result in content area weights that may be slightly different for each school.

In 1999 and 2000, the API was based primarily on the Stanford 9, a nationally-normed test that is administered annually to California public school students in grades 2 through 11 as part of the Standardized Testing and Reporting (STAR) program. In 2001, the API was based on (1) the Stanford 9 and (2) the California Standards Test in English-Language Arts (CST ELA). In addition to (1) and (2), the 2002 API was also based on (3) the California Standards Test in Mathematics for grades 2-11, (4) the California Standards Test in History/Social Science for grades 10-11, and (5) the California High School Exit Exam (CAHSEE) for high schools. Beginning in 2003, instead of the Stanford 9, norm referenced assessment was based on the California Achievement Test, 6th Edition (CAT-6). Two new tests were also added: (6) the California Standards Tests in Science for grades 9-11 and (7) the California Alternate Performance Assessment (CAPA) for students with disabilities. With every addition/change in test components, the SBE changed the weights in calculating the API and also added a scale calibration factor to minimize the impact of shifts in tests and ensure that "the statewide average API does not fluctuate solely as the result of adding new API components."[34]

Although the components of the API have changed over time, the performance calculation has not. In each year, a school's "API growth target" is five percent of the distance from the previous year's API to the statewide target of 800 or a minimum of one point growth between 1999-2000 or 5 points thereafter, as described more formally in the text.

Schools receive API scores as a whole as well as for each "numerically significant ethnic and

[34]For additional details, see API information guides for each year at http://www.cde.ca.gov/ta/ac/ap/

socio-economically disadvantaged subgroup in the school." These subgroups and what constitutes numerical significance are described more fully in the text. The subgroup target is calculated by first multiplying the school-wide target by 0.8 and then rounding the product to the nearest whole number.

B Awards Apportionment Data

Under state PSAA requirements, if a school meets or exceeds its growth target, it may be eligible to receive monetary or non-monetary awards through 3 programs: (1) the Governor's Performance Award Program (GPAP) [35]; (2) the Certificated Staff Performance Incentive Act; and (3) the Schoolsite Employee Performance Bonus (authorized for 1999-2000 SY only). Currently, no funding is appropriated in the budget for monetary awards.

Governor's Performance Awards (GPAPs) were paid for performance in the 1999-2000 and 2000-2001 school years. Awards were suspended thereafter because of budgetary problems. Existing site governance teams or the school-wide council decide how the funds from the GPAs are used, which then gets ratified by the local school board. The data on the apportionments received by each school for performance in the 1999-2000 and 2000-2001 school years can be downloaded from http://www.cde.ca.gov/ta/sr/gp/ap/apport00a.asp and http://www.cde.ca.gov/ta/sr/gp/ap/apport99a.asp.

The GPA program was funded up to $150 per pupil, but there is substantial variation in the rewards paid out in our data. There are 4 observations that suggest payouts over $150 per pupil. Excluding those, the average reward was $66 per pupil (standard deviation $8.25) and ranged from $21 per pupil to $111 per pupil for 7433 schools over the two school years.

The Certificated Staff Performance Awards (CSP) were paid only in 1999-2000 to schools that made "substantial improvement in their API," which meant at least twice their API target. The distribution of awards was decided by the local district in negotiation with the teachers' union. The payment setup was such that teachers from schools with the highest growth received the largest bonuses. In particular:

- 1000 certificated staff in schools with largest growth got $25,000 each;
- 3750 certificated staff get $10,000 each;
- 7500 certificated staff get $5,000 each.

Since the eligibility was the same for the GPAP, the impact of the CSP as well as the Schoolsite Employee Performance Bonus are all captured in our current estimates.

C School Characteristics

Data on school characteristics come from the California Department of Education's California Basic Educational Data System (CBEDS), an annual school-based census. From CBEDS, we collect

[35] To qualify, elementary and middle schools (high schools) must have a 95% (90%) test participation rate, in addition to the school and all its subgroups meeting or exceeding its API growth target.

data on student enrollment, the allocation of teachers across subjects, and a host of demographic characteristics such as racial breakdown, parental education, and percent of students receiving free lunch. These data can be downloaded from http://www.cde.ca.gov/ds/sd/cb/filesethsch.asp.

D Fiscal Data

Fiscal data on revenues and expenditures are available not at the school level but at the school district level. We also obtained this data from the California Department of Education. They can be downloaded from http://www.cde.ca.gov/ds/fd/fd/. From these unaudited fiscal data, we create variables such as district level total spending per pupil, per pupil expenditures on all staff's salaries and benefits, teacher salaries, all classified staff salaries, all certificated staff salaries, books and instructional materials expenditures, and other line items.

E Data for Analyses

We created two datasets for the analyses in our paper. The first, which is at the school level, is used to analyze the impact of the award program on school outcomes such as achievement and school-level resources. It was created by merging the API, Award Apportionments, and CBEDS databases together using the unique school identifiers. Our analysis sample includes all schools with valid scores and for whom we could determine eligibility to receive an award. These include elementary and middle schools with at least a 95% test participation rate and high schools with at least a 90% test participation rate.

Since we were also interested in examining the impact of the program on fiscal outcomes, we also created a second district level dataset. To merge API and award eligibility information to the district-level fiscal data, we first collapsed our school-level dataset to the district level. For the purposes of the regression discontinuity design employed in the paper, we characterize each district by the *maximum* of the school-level "awards gap" or eligibility threshold. As described in the text, due to the subgroup rules we characterize each school's "awards gap" as the minimum difference between the gain scores and targets for the school overall and each of its numerically significant subgroups. Thus, each district's "award gap" is the across-school maximum of the within-school minimums. The district "award gap" as defined then picka up whether any school in the district received treatment. Finally, we merged this collapsed data to the fiscal data using unique district identifiers.

Table I
Awards Calculation for Two Elementary Schools in the 2001 School Year[a]

Mission Elementary, ID No.: 7616486084941

	Students Tested	API_t	API_{t-1}	*Gain Score*	*Award Target*	*Award Gap*
Overall	450	692	682	10	6	4
Black	71	610	589	21	5	16
Amer Indian	4					
Asian	14		Not	Numerically	Significant	
Filipino	9					
Hispanic	122	644	653	-9	5	-14
Pacific Islander	4					
White	225	735	720	15	5	10
Disadvantaged	211	632	632	0	5	-5
Min Awards Gap						-14

Salida Union Elementary, ID No.: 50712666113823

	Students Tested	API_t	API_{t-1}	*Gain Score*	*Award Target*	*Award Gap*
Overall	453	696	664	32	7	25
Black	16					
Amer Indian	1					
Asian	2		Not	Numerically	Significant	
Filipino	7					
Hispanic	218	627	576	51	6	45
Pacific Islander	5					
White	201	763	745	18	6	12
Disadvantaged	212	625	593	32	6	26
Min Awards Gap						12

[a]Notes:

1. Only students in grades 2-11 are tested. Parents can obtain waivers for their children exempting them from the exam.
2. A numerically significant subgroup is any of the groups listed above with (i) at least 30 students with valid Stanford 9 scores and at least 15 percent of school's tested enrollment or (ii) at least 100 students with valid Stanford 9 scores (regardless of percent of tested enrollment).
3. The disadvantaged subgroup is not mutually exclusive, i.e. it may contain students counted in other subgroups. A student is classified as socio-economically disadvantaged if (1) he or she qualifies for free or reduced price meals or (2) neither parents has received a high school diploma.
4. In 2000, the awards target = max (40 - .05 $*$ API_{t-1}, X) , where $X = 5$ for schools and 4 for subgroups. This calculation is rounded up and is always based on school's API_{t-1} even for subgroups.
5. The minimum awards gap is the minimum of the awards gaps for a school overall and each of its subgroups. The value must be non-negative for a school to receive an award. School B received about $50 per pupil or a total of $42847 for performance in the 2001 SY.

Table II
Sample Characteristics by Award Receipt Status[a]

Panel A			*Basic Statistics*		
	All	*No Awards*	*Award for 2000*	*Award for 2001*	*Award Both Years*
Percent by Category	100	22.8	30.7	14.7	31.8
Award Per Pupil ($)	63.1	–	66.5	58.9	62.4
	(7.76)	–	(2.67)	(9.80)	(8.21)
Total Award ($)	48554	–	53742	52566	45019
	(29487)	–	(33739)	(37973)	(23890)
School Enrollment	856	1068	824	884	720
	(606)	(845)	(541)	(618)	(366)
Elementary	69.7	49.3	70.5	65.7	85.3
Middle	17.1	21.5	17.4	21.7	11.5
High School	13.2	29.2	12.1	12.6	3.2
# of Subgroups	1.17	1.34	1.19	1.18	1.04
	(0.74)	(0.84)	(0.71)	(0.72)	(0.67)
"Lost" Award	17.9	45.1	15.1	20.3	–
Panel B			*API Scores*		
	All	*No Awards*	*Award for 2000*	*Award for 2001*	*Award Both Years*
School	652	632	669	636	658
	(110)	(108)	(105)	(112)	(111)
African Americans	550	527	568	533	572
	(88)	(81)	(88)	(93)	(83)
American Indians	579	530	653	667	583
	(98)	(76)	(113)	(63)	(86)
Asians	749	708	782	737	761
	(121)	(124)	(111)	(116)	(120)
Filipino	743	709	758	742	770
	(70)	(67)	(65)	(63)	(65)
Hispanics	574	551	589	558	582
	(86)	(84)	(84)	(85)	(85)
Pacific Islanders	585	475	696	–	–
	(120)	(30)	(19)	–	–
Whites	746	725	755	733	759
	(75)	(77)	(71)	(76)	(71)
Socially disadvantaged	581	555	595	570	592
	(86)	(84)	(84)	(88)	(84)
Panel C			*API Gain Scores, (API_t - API_{t-1})*		
	All	*No Awards*	*Award for 2000*	*Award for 20001*	*Award Both Years*
School	26	9.4	24	24	42
	(30)	(27)	(32)	(26)	(23)
African Americans	27	9.4	28	30	51
	(38)	(29)	(43)	(35)	(27)
American Indians	24	10	11	63	45
	(39)	(31)	(29)	(55)	(34)
Asians	22	9.3	21	22	37
	(28)	(24)	(28)	(27)	(25)
Filipino	20	7.1	22	18	37
	(28)	(23)	(32)	(25)	(23)
Hispanics	31	13	28	30	49
	(35)	(32)	(38)	(32)	(26)
Pacific Islanders	15	0	29	–	–
	(57)	(29)	(78)	–	–
Whites	22	10	21	20	37
	(31)	(30)	(33)	(29)	(24)
Socially disadvantaged	30	11	27	30	49
	(37)	(34)	(41)	(32)	(28)

[a]Notes:

1. Standard deviations are given in parenthesis.
2. Zeros are not counted in the award payment calculations.
3. American Indians are "numerically significant" in only 29 school-years and Pacific Islanders in only 8 school-years.

Table III
Impact of the Awards Program on API Scores[a]

Panel A: 2000 SY Awards Program

	School Overall			Socially Disadvantaged Subgroups		
	2001 API Score	*2002 API Score*	*2003 API Score*	*2001 API Score*	*2002 API Score*	*2003 API Score*
Mean	689	697	721	595	615	654
Treatment	-3.56	-3.38	-5.58	.690	-.536	-1.98
	(5.75)	(6.24)	(4.99)	(7.00)	(6.85)	(5.37)
F-statistic	1.02	.949	.828	.979	.861	.778
p-value	.416	.679	.960	.567	.916	.989

Panel B: 2001 Awards Program

	School Overall			Socially Disadvantaged Subgroups		
	2001 API Score	*2002 API Score*	*2003 API Score*	*2001 API Score*	*2002 API Score*	*2003 API Score*
Mean		699	723		616	656
Treatment		.502	1.97		-2.24	1.95
		(3.79)	(3.92)		(4.74)	(4.78)
F-statistic		.974	1.04		.937	1.00
p-value		.591	.350		.725	.486

[a]Notes:

1. Standard errors are in parenthesis and are clustered at the level of a school's distance to the awards threshold.
2. The p-value corresponds to the F-test of the explanatory power of the 5th order polynomial fits relative to the fully flexible model.
3. Differences in the mean API scores for a given year across award program samples occur because some schools evaluated for an award in the 2000 SY are disqualified in 2001 SY and vice versa. Disqualifications are due to data irregularities, failure to meet required participation rates, and so on. See text for further details.

Table IV
Impact of the Award Program on School Resource Allocations[a]

Panel A	*2001 Allocations Relative to 2000 Award*				
	FTE Per Pupil	*Share FTE Math*	*Share FTE English*	*Computers Per Pupil*	*Internet Connections Per 100 Pupils*
Mean	.048	.040	.065	.152	3.32
Treatment	.0004	-.0003	.003	-.005	-.006
	(.003)	(.002)	(.003)	(.006)	(.306)
F-statistic	1.12	.762	.884	1.63	1.07
p-value	.121	.997	.875	.105	.252
Panel B	*2003 Allocations Relative to 2001 Award*				
	FTE Per Pupil	*Share FTE Math*	*Share FTE English*	*Computers Per Pupil*	*Internet Connectionsl Per 100 Pupils*
Mean	.049	.042	.064	.191	4.63
Treatment	-.0006	.0002	-.003	-.020	-.490
	(.0004)	(.002)	(.003)	(.006)	(.290)
F-statistic	2.22	1.10	.961	1.12	1.08
p-value	.0000	.167	.642	.0002	.205

[a]Notes:

1. The first row gives the mean of the dependent variables.
2. FTE are full time equivalent teachers.
3. Standard errors are given in parenthesis.

Table V
Impact of the Award Program on District Resources[a]

Panel A	*2001 Per Pupil Allocations Relative to the 2000 Award Gap*			
	Award Apportionment	*Unrestricted Revenue*	*Total Revenue*	*Total Expenditures*
Treatment	41.8	104	343	348
	(3.36)	(20.6)	(206)	(186)
F-statistic	.471	.721	.877	.716
p-value	.999	.991	.829	.992

Panel B	*2003 Per Pupil Allocations Relative to the 2001 Award Gap*			
	Award Apportionment	*Unrestricted Revenue*	*Total Revenue*	*Total Expenditures*
Treatment	28	20.2	123	202
	(3.10)	(6.92)	(318)	(302)
F-statistic	.615	.786	1.12	1.08
p-value	.999	.964	.174	.260

[a]Notes:

1. Standard errors are given in parenthesis.

Appendix Table I
Predetermined Characteristics Relative to the 2000 SY Awards Threshold[a]

Panel A — *2000 SY Characteristics of California Schools*

	Enrollment	*Share Free Meals*	*Share Elementary*	*Share High Schools*	*Number of Subgroups*	*Disadvantaged Subgroups*
Mean	842	.471	.706	.124	1.08	.715
Treatment	26	-.047	-.033	.008	.039	-.031
	(54)	(.025)	(.060)	(.043)	(.059)	(.047)
p-value	.631	.066	.578	.853	.510	.506

Panel B — *Share of Tested Students by Race/Ethnicity*

	Hispanic	*Black*	*Asian*	*American Indian*	*Filipino*	*White*
Mean	.380	.081	.080	.010	.024	.418
Treatment	-.015	-.018	.018	-.004	.008	.005
	(.024)	(.011)	(.010)	(.003)	(.005)	(.022)
p-value	.543	.125	.068	.232	.116	.835

[a]Notes:

1. The first row gives the mean of the dependent variable.
2. Standard errors are given in parenthesis and are clustered at the level of a school's distance to the awards threshold.

Figure 1. API Growth Required to Qualify for Governor's Performance Award as a Function of School's Base API Score: 1999 and 2000 Rules

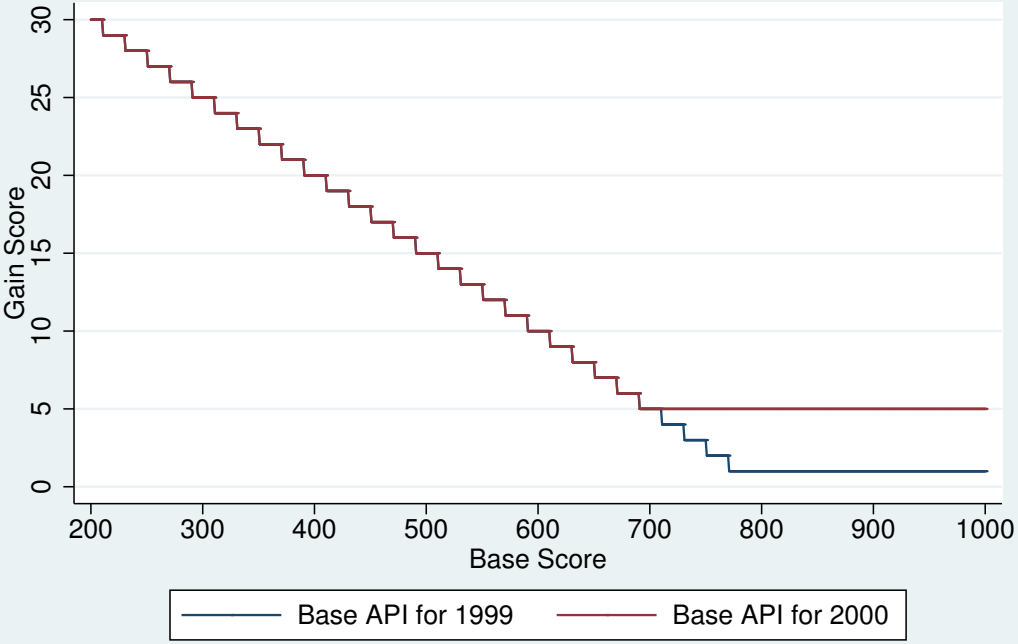


Figure 2. Share of Schools Receiving an Award for 2000 SY Performance Relative to the Distance to the Awards Eligibility Threshold

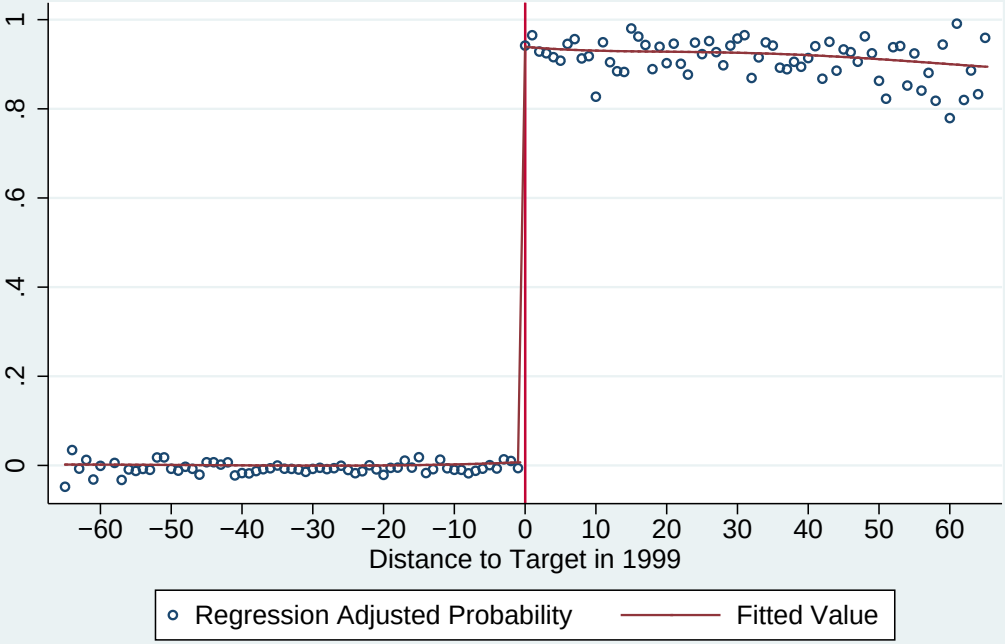


Figure 3a. Per Pupil Award Payment for 2000 SY Performance Relative to the Distance to the Awards Eligibility Threshold

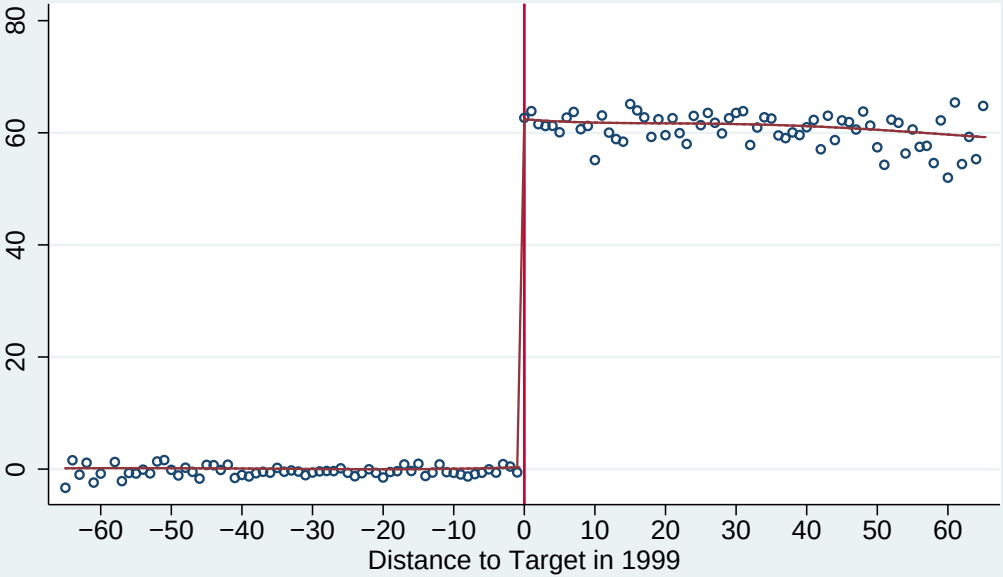
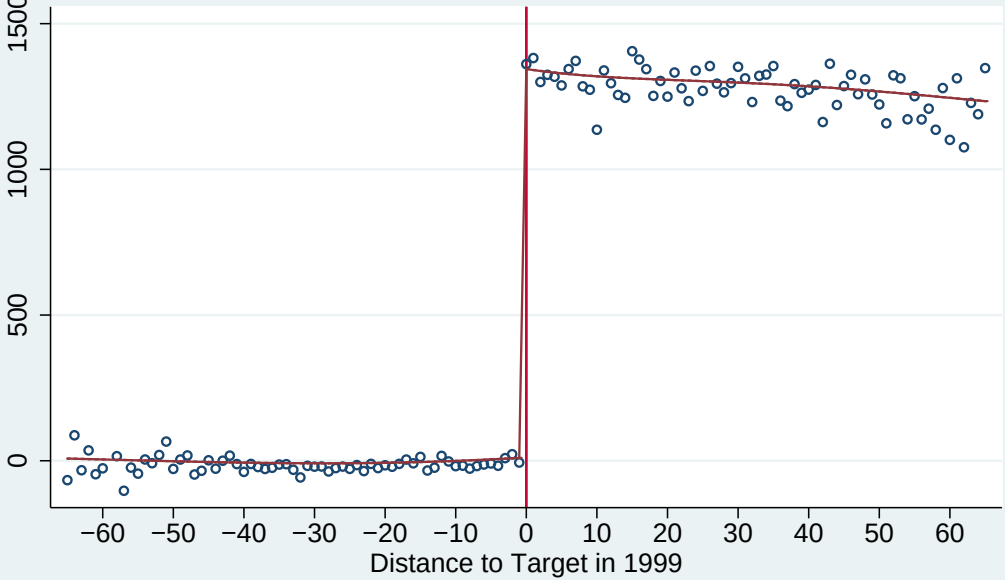


Figure 3b. Per Teacher Award Payment for 2000 SY Performance Relative to the Distance to the Awards Eligibility Threshold



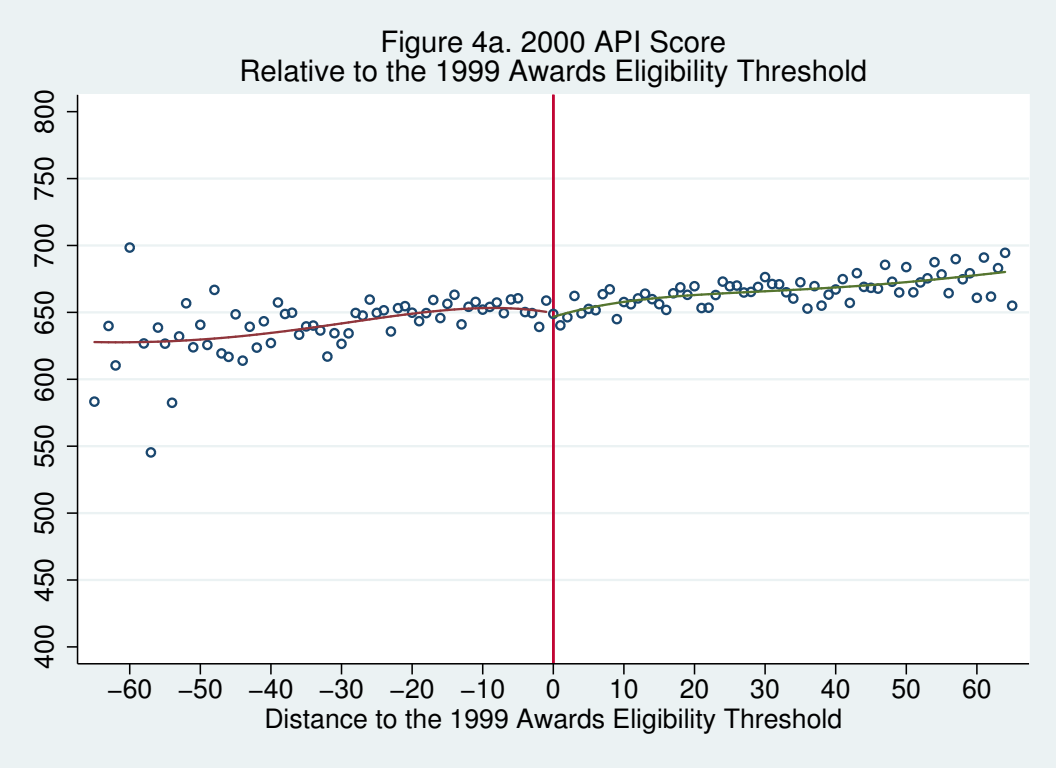


Figure 4a. 2000 API Score
Relative to the 1999 Awards Eligibility Threshold
800
750
700
650
600
550
500
450
400
−60
−50
−40
−30
−20
−10
0
10
20
30
40
50
60
Distance to the 1999 Awards Eligibility Threshold

Figure 4b. 2001 API Score
Relative to the 1999 Awards Eligibility Threshold

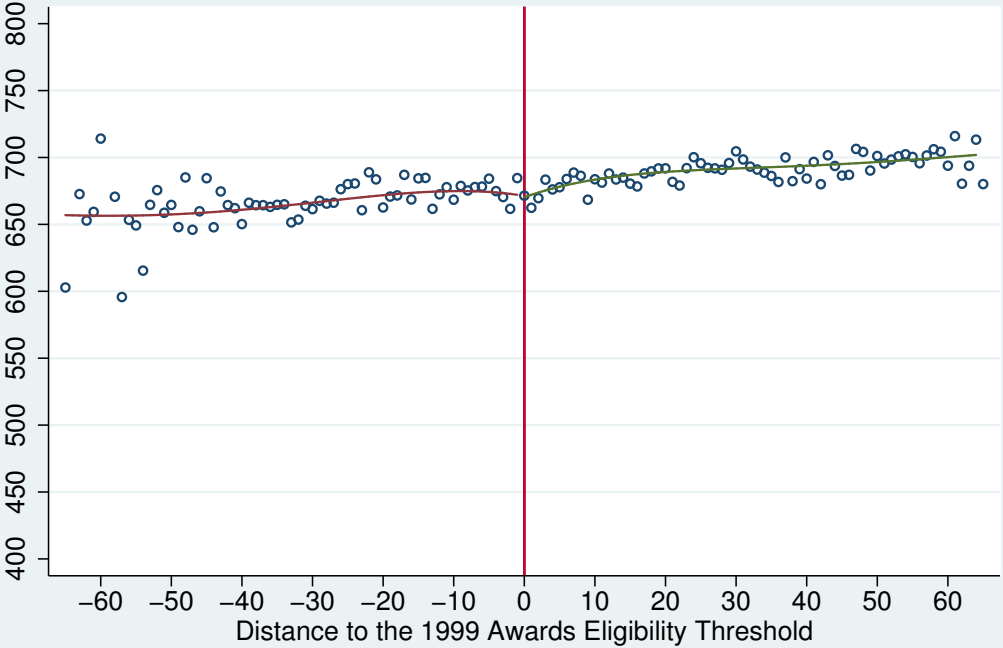


Figure 4c. 2002 API Score
Relative to the 1999 Awards Eligibility Threshold

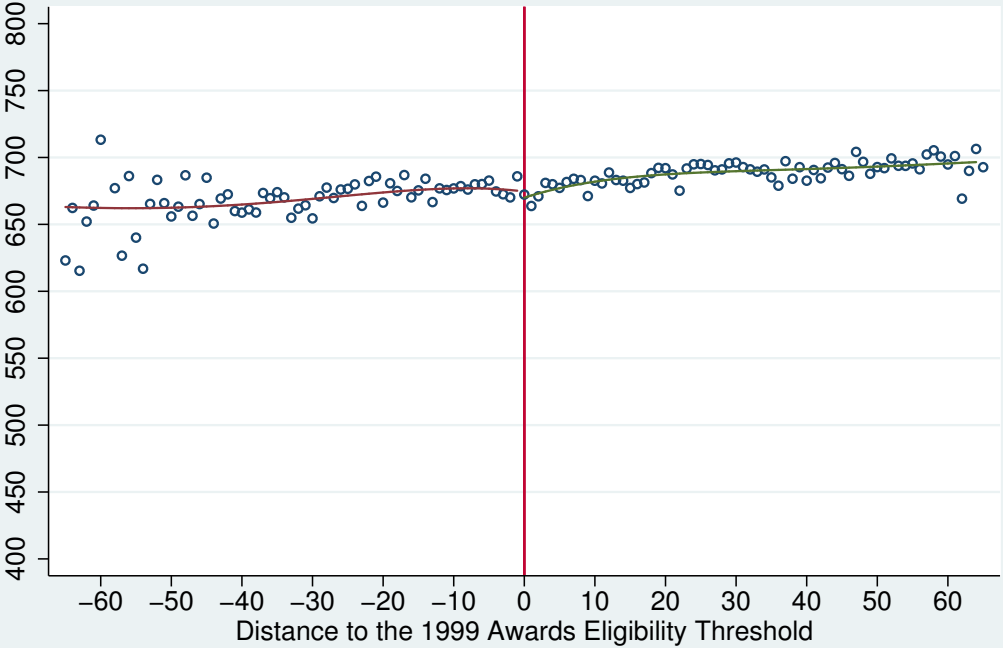


Figure 5a. 2000 API Score for Disadvantaged Subgroups
Relative to the 1999 Awards Eligibility Threshold

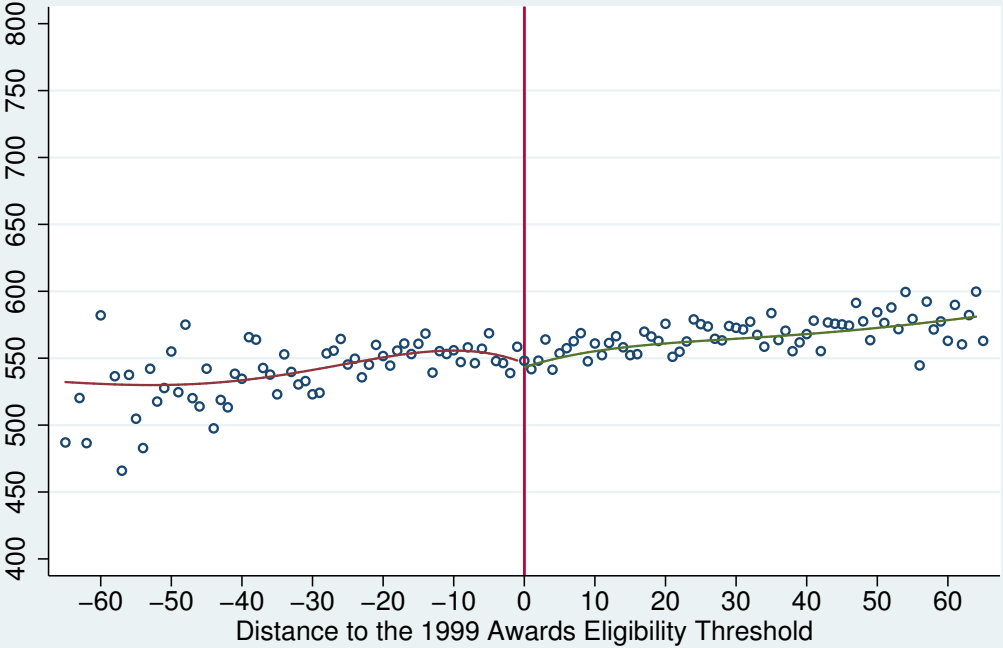


Figure 5b. 2001 API Score for Disadvantaged Subgroups
Relative to the 1999 Awards Eligibility Threshold

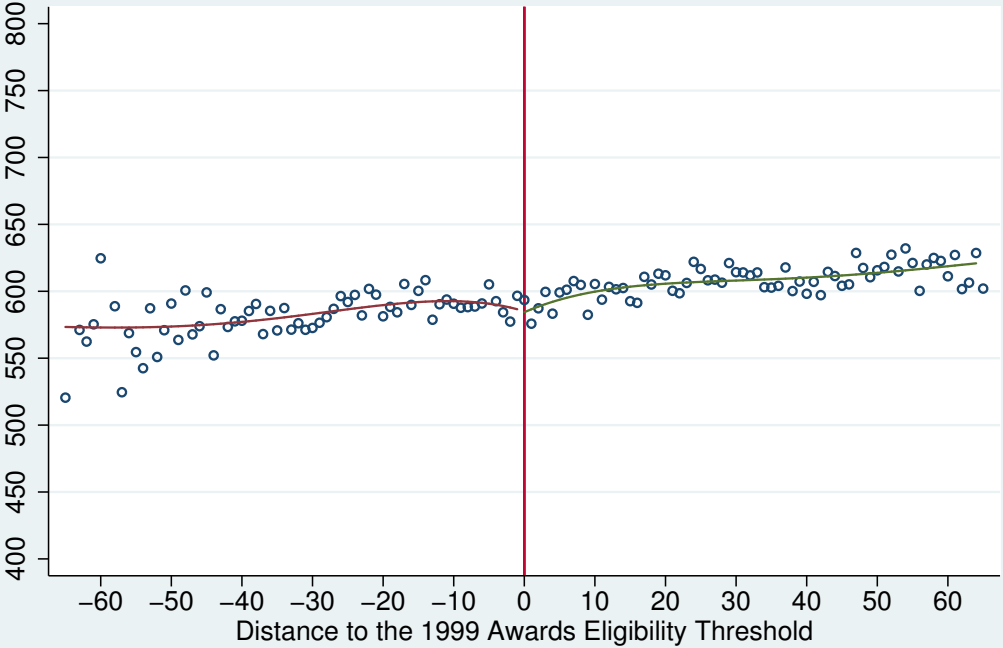


Figure 5c. 2002 API Score for Disadvantaged Subgroups
Relative to the 1999 Awards Eligibility Threshold

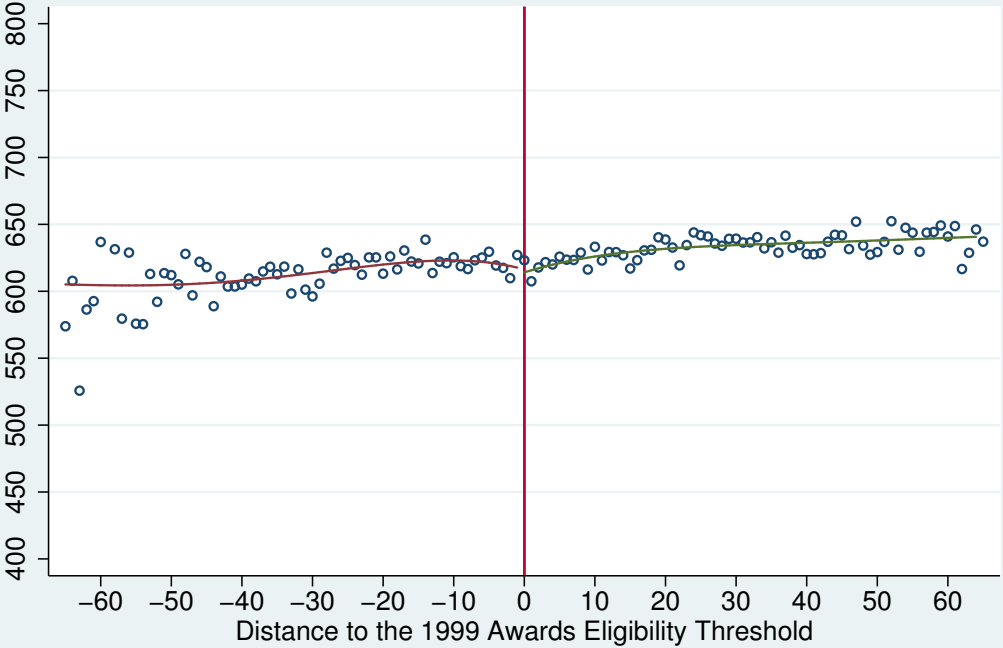


Figure 6. District−Level Per Pupil Award in 2000 SY
Relative to 2000 SY Eligibility Threshold

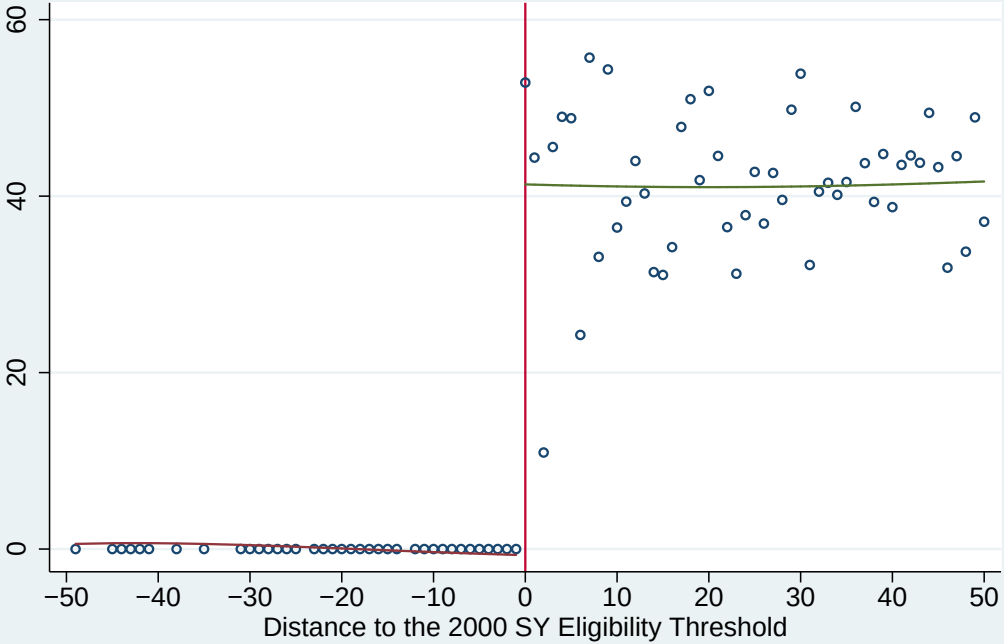


Figure 7. District Awards Category Revenue Per Teacher in 2001
Relative to 2000 SY Eligibility Threshold

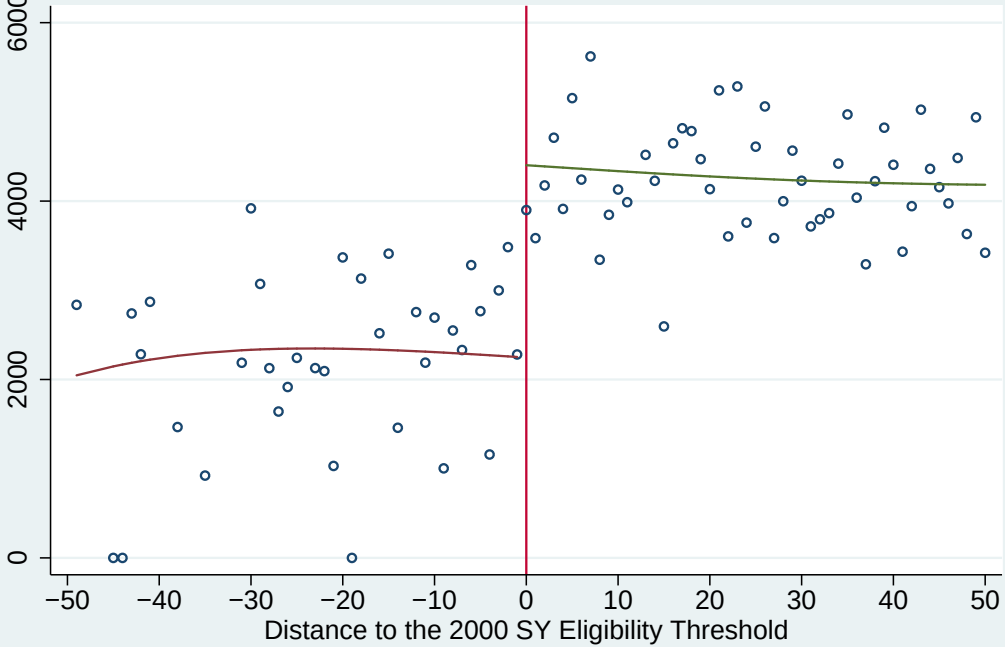


Figure 8. Total Per Teacher Revenue in 2001
Relative to 2000 SY Eligibility Threshold

Figure 9. Total Per Teacher Expenditures in 2001
Relative to 2000 SY Eligibility Threshold