Beyond Incentives: Do Schools use Accountability Rewards Productively? [1]

Marigee Bacolod

A.P.E.

John DiNardo

University of Michigan

Mireille Jacobson

RAND

April 1, 2011

[1]We wish to thank Tom Y. Chang, Julie Berry Cullen, Nora Gordon, Eric Hanushek, Justin McCrary, Doug Miller, Heather Royer, Diane Whitmore Schanzenbach, and participants at the 2006 SOLE meetings and the 2007 All-UC Labor Workshop for many helpful suggestions. We especially thank Tom Y. Chang for his assistance in putting together the data on award apportionments. Financial support was provided by the Haynes Foundation. All errors are our own.

Abstract

We use a regression discontinuity design to analyze an understudied aspect of school accountability systems - how schools use financial rewards. For two years California's accountability system financially rewarded schools based on a deterministic function of test scores. Qualifying schools received awards amounting to about 1% of statewide per pupil spending. Corroborating anecdotal evidence that awards were paid out as teacher bonuses, we find no evidence that winning schools purchased more instructional material, increased teacher hiring or changed the subject-specific composition of their teaching staff. Most importantly, we find no evidence that student achievement increased in winning schools.

1 Introduction

"Accountability mandates" – the explicit linking of a public school's resources and autonomy to student performance on standardized tests – have proliferated in the last 15 years. Accountability mandates can be crudely divided into those that enable school choice or district takeover for poor test performance and those that provide money and recognition for high performance. While many have studied the impact of sanctions for poor performance (e.g., see Chiang 2009; Chakrabarti 2007; Rouse et al. 2007), studies of financial award programs are less common. In this paper, we study an accountability reform in the California Public School system that rewarded schools and teachers within those schools that made adequate progress or attained a "passing grade" with cash bonuses.

Three programs rewarded high performing California schools and their teachers: the Governors Performance Award Program (GPAP), the School Site Employee Performance Bonus (SEPB) and the Certificated Staff Performance Incentive Award (CSPIA). Unlike the SEPB and the CSPIA, which were employee bonus programs, the GPAP, "was money to be used for school site purposes [e.g., purchasing computers]." However, the California Department of Education (CDE) found it was often "awarded to certificated staff [i.e., teachers] in the way of bonuses or stipends..." (Chladek, 2002, p. 4). Assuming, as the CDE suggests, that the $227 million GPAP funds were paid out as bonuses, distributed funds amounted to $1300 per teacher. Explicit bonuses paid out under the CSPIA ranged from $5,000 to $25,000 per teacher. The SEPB, which was shared by schools and their staff, paid on average $591 to each full-time equivalent (FTE) (see AP 2001). Thus, teachers at winning schools could have earned up to $27,000, although $1,900 was more typical.

We evaluate the effect of financial awards on questions of immediate policy relevance: what happens when schools receive payments through an accountablity system? How do schools spend these resources? Do the awards increase subsequent student achievement? As discussed more below, because of the relative size of the program, its adherence to clear assignment rules, its relative longevity and data availability, we focus specifically on the

GPAP. We discuss how the other programs affect the interpretation of our results.

To analyze these questions, we exploit the deterministic nature of the GPAP and SEPB. Schools (and teachers within schools) that met a pre-determined threshold for improvement in exam performance received financial awards. The discontinuity in the assignment rule - schools that barely missed the performance target received no reward - generates quasi-random assignment in awards recipiency for schools close to their eligibility threshold. This enables us to generate credible estimates of the impact of the program on school resource allocations and achievement. To validate this research design, we demonstrate that baseline (pre-program) school characteristics have the same distribution just above and below the awards threshold (Lee and Lemieux 2009). Failure to meet this key assumption would suggest that schools are able to precisely manipulate their test scores to secure an award and would cast doubt on a causal interpretation of our results.

The regression discontinuity design allows us to circumvent many difficult issues in evaluating accountability programs. Due to sampling variation alone, measures of aggregate student performance at smaller schools will be noisier than those at large schools with similar students, biasing naive before and after comparisons of schools facing accountability mandates (Kane and Staiger 2002). Chay, McEwan, and Urquiola (2005) document that accounting for mean reversion in test scores substantially reduces the estimated impact of a Chilean accountability-like program on achievement. In addition, as demonstrated by Figlio and Rouse (2006) in Florida, changes in exam performance may reflect changing student characteristics rather than actual improvements in achievement *per se*.

While our research design allows us to sidestep these issues, it cannot assess the program's full effect on achievement. Specifically, it cannot capture any effects that occur uniformly to schools that received awards and schools that did not. Schools might be uniformly affected if, for example, the awards program induces them all to narrowly tailor the curriculum to maximize test scores. To assess the incentive impacts of California's accountability system, we would need counterfactual schools not subject to the accountability scheme. For several reasons, however, incentives were likely weak. First, schools

and teachers were rewarded based on group performance, introducing a free-rider problem. Since performance pay based on a clear measure of individual output provides the strongest incentives to workers (Muralidharan and Sundararaman 2011), the group nature of the award likely muted the incentive effect of this program. Second, schools and teachers had little opportunity to learn how to increase their odds of winning awards because the duration of the program was short - funded for two years for the GPAP but only one for the CSPIA and SEPB - and the criteria for award receipt were revealed late and changed over time. Schools learned the award criteria in July 2000, several months after the first year of testing (Kerr 2000). And, as described below, the eligibility requirements changed across years. Finally, the instability of program funding may have sent a signal that awards were not core elements of California's accountability scheme.

Our design does allows us to study how schools spend additional resources from an accountability scheme and how this in turn affects 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 $1300 per teacher and $1100 per teacher for the 2001 SY award, where, as is convention in the literature, the 2000 SY captures fall of 1999 through spring of 2000. Consistent with reports that most funds were paid out as bonuses, we find no evidence that these resources were used for direct instructional purposes. Finally, 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.

Our findings suggest that untargeted awards do not guarantee improvements in academic achievement. Likewise, in contrast to work showing that the stigma threat of being labeled a failing school motivates low-performing schools to improve in Florida (Figlio and Rouse 2006) and North Carolina (Ladd and Glennie 2001), we find no evidence that publicly recognizing schools as passing (or failing) an accountability standard improves achievement in California. However, in our context, the direction of stigma and/or prestige effects are a priori ambiguous and could lead to improvements in outcomes for the "failers" relative to

the "winners." Finally, our findings are consistent with work showing that the much larger appropriations from the Federal Title I program had no positive effects, and possibly even adverse effects, on student achievement in New York (van der Klaauw 2008).

In what follows, we discuss California's accountability system, 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. In Section 3 we present our econometric framework for estimating the effect of the awards and in Section 4 we present our findings. Finally, in Section 5 we offer a summary and concluding observations.

2 Background: California's Academic Performance Index and Financial Award Programs

California's accountability system, which predates the No Child Left Behind Act (NCLB), was established by the Public Schools Accountability Act (PSAA) of 1999. The PSAA aims to hold "public schools accountable for the academic progress and achievement of its pupils within the resources available to schools" (see California Education Code 52050-52050.5). To measure achievement, the PSAA created the Academic Performance Index (API), calculated annually for all public schools. The index ranges from 200 to 1000 and combines test scores from students in grades 2 to 11. For the first two years after the PSAA's passage – the only years that the budget allocated funds for financials awards – the API was based solely on the nationally norm-referenced Stanford 9 exam. For middle and elementary schools, the API incorporated reading, language arts, spelling, and math exam scores. Test components were similar for high schools, except science and social studies replaced spelling. Test components have been added over time. As described in more detail in Bacolod et al. (2009) and Rogosa (2003), the API coarsens exams scores at many levels, imputes missing values, and weighs subject components differently based on grade and year.

Several programs financially rewarded schools based on API growth: the $227 million Governor's Performance Award Program (GPAP), the $350 million School Site Employee

Bonus Program (SEPB), and the $100 million Certificated Staff Incentive Award Program (CSPIA). While the SEPB was a 1-year program, the GPAP and CSPIA were to be ongoing. All were paid out based on 1999 to 2000 SY API growth but, due to budget cuts and lawsuits over the CSPIA, only the GPAP survived another year, albeit at a reduced budget of $144.3 million (Hausler 2001; Folmar 2001). The GPAP was intended for school site purposes, such as instructional materials and equipment, but anecdotal evidence and our own results sugggest that many schools used the funds as teacher bonuses. The CSPIA and the SEPB targeted employees explicitly – certificated staff (teachers), in the first instance and both certificated and classified staff (paraprofessional, administrative and clerical) staff in the second. In principle, the SEPB granted half of the award to schools for unrestricted uses. However, we could find no official information on the SEPB sharing rules or payments.

While the GPAP and SEPB were based on the same growth target, the CSPIA was available only to staff at schools that demonstrated growth over *twice* their 2000 SY GPAP target, performed in the bottom five deciles on the 1999 SY API and demonstrated some growth in the 1999 SY. Moreover, meeting these criteria did not guarantee CSPIA receipt. Rather, local education agencies had to apply on behalf of schools. The state ranked eligible schools by their growth and paid out awards until all CSPA funds were spent. Local school districts distributed funds in negotiation with the teachers' union (http://www.cde.ca.gov/TA/ac/pa/cspiprogram.asp). Because the CSPIA was not paid to all schools meeting the target and is not based on a clear set of observable rules, this program is not well-suited to an RD. Consequently, and because it was small in scope (affecting about 260 schools compared to the thousands receiving GPAP and SEPB awards), we focus instead on the GPAP (and SEPB) API target. Our RD estimator for the 2000 SY captures the combined effect of these two awards programs. Because the SEPB was suspended after the 2000 SY, our analysis of the 2001 SY captures the effect of the GPAP alone.

[Insert Figure 1 here]

We study outcomes several years out to account for the lag between award announcements and payouts (see the timeline in Figure 1) and the potential for announcements to

affect outcomes. The first GPAP award (and the only SEPB and CSPIA awards) was based on testing in the spring of 2000. Awards were announced in the fall of 2000 but the first payment was not made until January 2001, the middle of the 2001 SY. The second and final payment was made in March 2002, the following school year. In between, the state paid out the SEPB and CSPIA awards, with the former apportionment occurring in March 2001 and the latter in October 2001. State budget problems increased the time to apportionment in the second GPAP year. Awards were announced late in 2001 but not apportioned until July and October 2002. We next detail the process for determining award eligibility, which is central to our RD design.

2.1 Award Eligibility – The Simple Case without Subgroups

California's accountability system is based on API "growth" scores – the year to year change in API – relative to a target for a school as well as for each "numerically significant subgroup." For schools without numerically significant subgroups, the API growth target is 5% of the gap between the previous year's API and a statewide goal of 800 or a specified minimum. In the 2000 SY, the minimum gain was one point; in the 2001 SY, it was raised to five points. Operationally, this can be expressed as:

$$\begin{aligned} Target_{2000SY} &= max(.05 * (800 - baseAPI_{99}), 1) \\ Target_{2001SY} &= max(.05 * (800 - baseAPI_{00}), 5) \end{aligned} \tag{1}$$

where $Target_t$ is the minimum change in the base year API score needed to qualify for an award in year t and $baseAPI_{t-1}$ is just the (adjusted) API from $t-1$.

[Insert Figure 2 here]

Figure 2 plots the 2000 and 2001 SY award targets and underscores several key issues. First, although not made explicit in the rules, the California Department of Education rounds gain scores to the nearest integer, such that eligibility thresholds are represented by a step function. Second, schools with lower initial API scores have to achieve larger API gains to receive an award than high achieving schools. Finally, raising the minimum

target from one to five points between the 2000 and 2001 SYs had the effect of increasing the award threshold by four points for schools at or above an API of 780 while increasing it by the nearest integer value of $0.05 * baseAPI - 35$ for those with an API of 700 to 780.

2.2 Award Eligibility with Numerically Significant Subgroups

To receive an award, each school's numerically significant subgroup must make "comparable achievement," defined as 80% of the school's growth target. Subgroups are defined by race/ethnicity (African American, American Indian, Asian, Filipino, Hispanic, Pacific Islander and Caucasian) or socioeconomic disadvantage (eligible for free or reduced-priced meals or from a family where the highest education level is below high school completion). Racial/ethnic subgroups are mutually exclusive; the socially disadvantaged subgroup may contain students from other subgroups. To achieve "numerical significance" a subgroup must have (a) 30 to 99 tested students and constitute at least 15% of total school enrollment or (b) have 100 or more tested students, irrespective of enrollment share.

To make this calculation concrete, Table I of the Online Appendix documents the 2001 SY award eligibility calculation for two schools, Salida Union Elementary and Mission Elementary. Both tested about 450 students and have students in each of the state-defined subgroups. They differ in which subgroups are sizeable enough to face performance targets. Neither school tests even 30 American Indians, Filipinos, Asians, or Pacific Islanders, exempting these groups from performance targets. The African American subgroup (16 tested students) in Salida Union are also exempt. Since the tested number of Hispanics, whites, and socially disadvantaged students are each greater than 100 in both schools, each faces subgroup performance targets. African Americans in Mission Elementary also face subgroup rules since they number more than 30 and over 15% of tested students.

Based on school performance alone, both schools qualified for awards. Mission Elementary had growth of 10 API points, exceeding its school-wide target of six points. Salida Union had a 32 point gain, exceeding its seven point target. However, only Salida Union met both the school and all subgroup performance targets. Two (out of four) of Mission

Elementary's numerically significant subgroups–Hispanics and the socially disadvantaged–failed to meet their performance target of 5 API points (80% of the school target).

To capture award eligibility, we characterize each school by an "award gap," the *minimum* of the difference between the gain score and performance target for the school and each of its numerically significant subgroups. We use the minimum since a school is award-eligible only if all performance targets are met. Thus, Mission Elementary receives its -14 point Hispanic gap, its highest barrier to awards, and Salida Union is characterized by the +12 point gap for the socially disadvantaged subgroup, its smallest gain score. Schools with negative award gaps, like Mission Elementary, are ineligible for awards; those with gaps that are greater than or equal to zero, like Salida Union, are eligible.

2.3 GPAP Award Allocations

Table I describes GPAP allocations, school performance and some school characteristics overall and by award receipt status for the 2000 and 2001 SYs. Student enrollments and school characteristics come from the California Basic Educational Data System (CBEDS), an annual school-based census. GPAP award apportionment data are from the California Department of Education (CDE). SEPB apportionment data were not available. Data construction is detailed in our Online Data Appendix.

[Insert Table I here]

Column (1) provides means for all elementary, middle and high schools that met the testing participation requirements for the program (95% in elementary and middle schools and 90% in high schools) and had valid API scores for both base and growth years in the 2000, 2001 or both school years. The next four columns separate the data into schools that a) never won an award, b) won an award for the 2000 SY, c) won an award for the 2001 SY, d) won an award for both years. Describing the data in this way reveals several important features of the award program.

Schools that never won awards account for 23% of the sample (row (1)). About 31% won awards for the 2000 SY alone while only 14.7% won for the 2001 SY alone, due to the

higher minimum gain score in that year. About 32% won awards in both SYs. Calculated over schools receiving any awards, per pupil payments averaged $63 across both years (row (2)), which, due to differences in enrollment data, is a few dollars less than the state-reported average of $69 per pupil. To put this in perspective, public K-12 expenditures in California were roughly $6000 per student in the 2000 SY year (Carroll et al. 2005). Thus, awards increased per pupil spending by just over 1%. Importantly, awardees had considerable discretion in using these funds. To the extent it was paid to teachers, the GPAP amounts to bonuses of almost $1300 per teacher. The SEPB, which was based on the same eligibility threshold, kicked in another $591 to each FTE and the same amount to their schools. Moreover, as we will show in section 4.3, additional resources may have flowed to districts with schools that qualified for awards.

Table I, row (4) shows enrollments and reflects a fundamental problem with using mean test scores to measure school performance (Kane and Staiger 2002; Chay et al. 2005). All else equal, smaller schools have mean scores with higher sampling variation and thus are more likely to have a lucky year. Consistent with this, schools winning awards in both years are smaller ($p < 0.001$) and schools never winning are larger ($p < 0.001$) than the average school. Elementary schools, which account for 70% of the sample and are the smallest schools, are underrepresented (49%) among schools that never winning and overrepresented (85%) among those winning awards in both years. At the other extreme, high schools, which are the largest schools, represent 13% of the sample but 30% of schools that never won an award and only 3% of schools that won awards in both years.

The bottom of Panel A analyzes subgroup rules. To improve legibility and because they are rarely "numerically significant," we omit American Indian, Pacific Islander and Filipino subgroups from the table (but not the analysis). The typical school has only one subgroup; 15% have no subgroups. The most common subgroups are socially disadvantaged, Hispanic and white. Since subgroups face additional eligibility criteria, schools that never won awards have more subgroups and those that won in both years have fewer subgroups than the average school. The last row of Panel A shows that 18% of schools would have won

an award based on school performance alone but were ineligible because of subgroup rules. Likewise 45% of schools that never won awards would have won without the subgroup rules. This average masks the effect of raising the minimum API growth targets between the 2000 and 2001 SY. Whereas 53% of schools in the never group would have won awards based on school criteria in the 2000 SY, only 38% would have in the 2001 SY. In other words, raising the minimum growth targets reduced the bite of subgroup rules.

Panel B shows API scores and Panel C shows gain scores averaged across award years and for each numerically significant subgroup. The mean API is 652. White and Asian subgroups perform well above average. Socially disadvantaged subgroups have an API almost two thirds of a standard deviation below the average. Hispanic subgroups are also well below the mean. The sharp difference in school characteristics across categories highlights the importance of a strong empirical research design in the work that follows.

3 Econometric Framework

The main challenge to estimating the ex-post effect of financial awards on achievement or resource allocations is that awards are not randomly assigned. For example, schools with more subgroups are less likely to receive awards. To circumvent this issue, we use a regression discontinuity (RD) design that compares schools that just barely won an award to those that just barely "lost" an award. Let D_i equal the school's award gap or minimum distance between its gain scores and award eligibility targets $((API_{it} - API_{it-1}) - Target_i)$, so that zero corresponds to having just met the target. Schools with $D_i \geq 0$ win the financial award; schools with $D_i < 0$ do not. The discontinuity in the rules translating test scores into award eligibility generates quasi-random assignment in award receipt near the eligibility threshold. As we approach the threshold from the left and the right, both the unobservable and observable differences across schools shrink.

As a practical matter, we need not limit the comparison to the few schools just to the

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

$$Y_i = \alpha + \beta T_i + g(D_i) + \epsilon_i \quad (2)$$

where Y_i measures school i's achievement or resources; α is a constant; T_i is an indicator equal to 1 if school i received an award; and $g(\cdot)$ is a unknown continuous function. Although unknown, it can be approximated by polynomials in D and full interactions with the awards indicator T. Based on visual inspection of the API data, a comparison of the F-tests across API models, and the fact that odd order polynomials tend to have better efficiency and do not suffer from boundary bias problems like even order polynomials (Fan and Gijbels1996), we chose a fifth-order polynomial. We test the sensitivity of our estimates to alternate specifications of the control function and to the use of nonparametric local linear regression techniques. As shown in Lee (2008), the RD estimates can be interpreted as a weighted average of the population treatment effects, where the weights are positively related to each observation's distance to their award target. Thus, schools closest to their target contribute the most and those farthest away the least to the estimated treatment effect.

To accommodate the different levels of the award, or varying treatment intensities, we can further recast the problem as an instrumental variables estimator where award receipt, A_i is the endogenous regressor and the "first stage" equation is given by:

$$A_i = \alpha + \psi T_i + h(D_i) + \nu_i \quad (3)$$

In this set up, the impact of awards on achievement or resource allocations is merely the indirect least squares estimate, $\frac{\widehat{\beta}}{\widehat{\psi}}$, the ratio of the discontinuity in the outcome equation to the discontinuity in the awards equation. Where the treatment effect is random, 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 Klaauw 2001), provided that monotonicity holds (i.e. that growth scores uniformly increase the probability of receiving treatment). Finally, just as in an RCT, we can include exogenous covariates X for variance reduction purposes, provided they are balanced (a restriction we test).

Our design is similar to the original RD approach used by Thistlewaite and Campbell (1960) to estimate the impact of a test-based scholarship program on future academic outcomes, except our unit of analysis is the school. Although individual student data are appealing, particularly for estimating the impact of the awards program on achievement, these data are difficult, if not impossible, to obtain. However, since the PSAA is based on average school performance, school-level data is sufficient for characterizing the program.

3.1 Validity of the RD

To implement the RD, we begin by verifying that API growth was awarded according to the rules. First we plot the regression-adjusted average share of schools receiving GPAP payments at each distance, D, from the eligibility threshold in the 2000 and 2001 SYs. Recall $D = ((API_t - API_{t-1}) - Target_t)$ or the difference between its gain score and growth target. In addition, we plot parametric estimates of the conditional probability of award payments to schools at each distance. Operationally, our parametric estimates are just the least squares fitted values from the following equation:

$$A = \delta T + P'\alpha_0 + TP'\alpha_1 + X'\beta + \varepsilon \tag{4}$$

where A is the probability of awards recipiency, $T \equiv 1(D \geq 0)$ 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 to allow the polynomial fit to differ on either side of the eligibility threshold.

For variance reduction, we include X, a set of controls that include: a school's enrollment, number of numerically significant subgroups, percent of tested students by race/ethnicity (white, black, Hispanic, Filipino, Asian, Pacific Islander, or American Indian), share qualifying for free or reduced price meals, and dummies for school type (elementary or high school with middle school the omitted). All covariates correspond to the academic year of the growth year score, i.e. t not $(t-1)$. To account for potential misspecification, standard

errors are adjusted to allow for an arbitrary correlation in errors at the level of D, the distance to the award threshold (Lee and Card, 2008).

Since our underlying data (test score *changes*) are discrete, the "true" nonparametric estimator is just the set of mass points in the running variable (Lee and Card, 2008). To formally assess the adequacy of our parametric representation, we compare our model to the fully saturated model that includes a separate indicator for every specific value of the running variable, D:

$$A = \sum_d Z_d \gamma_d + X'\beta + \mu \tag{5}$$

where Z_d is a dummy variable equals 1 if the school's distance is 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. 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 and RSS_{ur} are the residual sum of squares from the restricted (polynomial-fitted) and the unrestricted (fully flexible) models, respectively; J and K are the number of parameters in the respective models; and N is the number of observations. Under normality G is distributed $F(J-K, N-K)$. With this F-statistic, we test the null hypothesis that the polynomial model has as much explanatory power as the fully flexible model. To complement the formal tests, we conduct a "visual analysis" that plots the coefficients γ_d, the regression adjusted outcomes, and the parametric fit to gauge whether our estimates might be spurious.

[Insert Figures 3a and 3b here]

Figures 3a and 3b show the share of schools receiving awards at each distance from the eligibility threshold for the 2000 and 2001 SY, respectively. The open circles are the regression-adjusted average shares; the solid lines are the parametric fits. In both figures we see a marked discontinuity in the probability of receiving an award at the eligibility threshold. Schools to the left of the threshold, which failed to meet their targets, did not receive award payments. In actuality, in the 2000 SY, 5 schools (0.3%) that by our data did not meet their targets, received awards averaging $60.5 per pupil. At the threshold, where a school's API equals its target, the probability of receiving an award jumps to almost one

in both years. The estimated discontinuity is 0.93 with a t-stat of almost 80 for the 2000 SY and 0.84 with a t-stat of 55 for the 2001 SY. Both the regression adjusted averages and the polynomial fits past zero are strictly below one because a small share of schools, about 8% in the 2000 SY and 11% in the 2001 SY, made their API target but did not receive a payment. According to the CDE these schools may have been disqualified because of "data irregularities," over 15% of parents requesting exam waivers, or student population changes that invalidated the API. Because we do not observe what causes award rule violations, we include these cases in our work. Discarding them is problematic if the causes of the violations are correlated for unobservable reasons with our outcomes of interest. Consequently, our analysis is a "fuzzy" RD case, similar in spirit to an "intent-to-treat" design. However, results are similar if we exclude violations (available upon request).

[Insert Figures 4a, 4b here]

Figures 4a and 4b show average per pupil award payments for schools at each distance from the eligibility threshold based on 2000 and 2001 SY performance, respectively. As expected, schools to the left of the eligibility threshold have per pupil award payments of \$0. At the discontinuity award payments jump sharply. The estimated discontinuity is \$62 per pupil with a t-stat of 80 and \$50 per pupil with a t-stat of 50 based on 2000 and 2001 SY performance, respectively. Expressed per teacher, the estimated discontinuities are about \$1300 with a t-stat of over 70 and \$1083 with a t-stat of 46 based on 2000 and 2001 SY performance, respectively. Expressing payments per teacher is useful in light of evidence that the awards were paid to them as cash bonuses.

A visual comparison of our parametric estimates and the regression adjusted averages of award recipiency suggests that the fifth-order polynomial fits are reasonable. We confirm this with the F-statistic described above, which tests the null that the polynomial model has as much explanatory power as the fully flexible model. Across our award recipiency models (any award and award per pupil), G is less than one (0.773 and 0.764 for the 2000 SY program and 0.742 and 0.638 for the 2001 SY program). In no case can we reject the null that the restricted and unrestricted models have similar goodness of fits.

The RD provides many *testable* restrictions, similar to those in a randomized controlled trial (RCT). Specifically, schools just to the left (control) and right (treatment) of the eligibility threshold should have baseline characteristics that are the same on average. Said differently, the conditional expectation of predetermined characteristics with respect to the award gap, $E[X|d]$, should be continuous through the threshold, $d = 0$. This might not occur if some schools sort themselves to one (presumably the winning) side of the eligibility threshold by, for example, encouraging certain types of students to transfer schools. If this occurred, we might see differences in in the share of students by race or socioeconomic status. Since these factors independently affect outcomes, they could, in principle, confound our estimates of the treatment effect of awards.

The fact that the state allocates awards based on API *changes*, which are noisier and more difficult to manipulate than levels, lends credibility to our research design. Late finalization of the award rules and the change in rules across award years adds to this credibility. As a check, however, we test for explicit manipulation of the awards gap following the approach set out in McCrary (2008). Based on results from a "first step histogram" with a bin size of one and a second step smoother, we cannot reject the null hypothesis of continuity in the density of awards gap at zero, the threshold for an award, in either of the awards years (see Online Appendix Table II). We have also performed these tests separately by school size. In no case do we find evidence of manipulation of awards receipt.

We have also tested for smoothness in the observable, predetermined characteristics of schools. Online Appendix Figures 1a and 1b plot regression adjusted averages and polynomial fits of total enrollment and the number of numerically significant subgroups against the distance to the 2000 SY eligibility threshold. Both are smooth through the discontinuity. Similarly, 2000 SY award receipt rates, shown in Online Appendix Figure 2, do not change discontinuously at the 2001 SY award threshold. In other words, schools just qualifying and failing to qualify for 2001 SY awards are similar in terms of past awards recipiency. Panels A and B of Online Appendix Table III report estimated discontinuities at the 2000 SY award threshold for 14 predetermined characteristics. In 12 cases, the

discontinuity is not statistically distinguishable from zero at even the 10% level, implying that schools close to the eligibility threshold are similar on predetermined characteristics

Two cases merit 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 the 5% level but the p-values are only 0.07. These critical values are unadjusted for the well-known "multiple comparisons" problem, leading to overfrequent rejections of the null hypothesis of no effect when the null is in fact correct (Savin 1980). For free or reduced price meals, the point estimate implies a 4.7 percentage point or 10% drop in the share of students qualifying in schools that just received awards relative to those schools that just missed receiving them. For the share of test-takers that are Asian American, the implied effect is a 1.8 percentage point or a 23% increase. Importantly, we find no evidence of discontinuities in either outcome in the 2001 SY (available upon request). Moreover plots of these characteristics in Online Appendix Figures 2c and 2d provide no compelling evidence of discontinuities. Together, the plots and the 2001 findings suggest that the 2000 SY disconcontinuities may be a result of random variation and the large number of comparisons made.

To the extent these discontinuities are real, our estimates may be biased towards finding a positive impact of awards on achievement. Students qualifying for school meals are more likely to perform poorly (see socially disadvantaged subgroup API scores in Table I). Asian Americans perform well above the state average. A drop in the share of students qualifying for meal assistance and a bump up in the share of Asian Americans at the discontinuity, could lead us to overstate increases in test scores (or other positive outcomes). Fortunately, if such a bias exists, it should be small so long as higher scores do not induce some schools that (in the absence of their higher score) would have received an award, to be denied an award. This "monotonicity" condition is reasonable in our context and supports the idea that the conditions for identification are satisfied. Finally, because we find no effect of awards on outcomes, we are not concerned about overstating the effects of the program.

4 Results

We employ the same framework used to establish the discontinuity in the GPAP and smoothness in covariates, to estimate the causal impact of the award program. Using (4), we estimate the treatment effect on Y (instead of A). We use the goodness of fit statistic to test the sensitivity of this estimate to our functional form assumptions. We also estimate models using third or seventh order global polynomials and using a local linear approach with a range of bandwidths (available upon request). To minimize redundant plots, we provide figures for 2000 SY awards program. All tables present estimates 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. In appendix tables, we show local linear estimates using a bandwidth chosen according to the rule of thumb (ROT) procedure described in Lee and Lemieux (2009) for a rectangular kernel and using a quartic specification to estimate the curvature and standard error of the regression.

4.1 Evidence on Achievement

We first consider the impact of the awards program on achievement. If schools that win awards can spend these resources in ways that positively impact achievement, then we should see a jump in API scores at the discontinuity. In other words, schools that just barely won awards should have higher scores in subsequent years than their counterparts that just barely missed winning an award. Because it is unclear when such returns accrue, we study achievement for several years out.

[Insert Figures 5a, 5b, 5c, 5d here]

Figures 5a, 5b, 5c and 5d graphically represent our RD estimates of the impact of the 2000 SY awards program on the API in the 2001, 2002, 2003 and 2004 SYs, respectively. Because the first 2000 SY award apportionment was made in January 2001, in 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). Unsurprisingly, Figure 5a shows that 2001 API scores are

smooth across the awards eligibility threshold.

[Insert Table II here]

To the extent that additional resources were put towards instruction, as the CDE intended, we might expect achievement gains in the 2002 SY or later. But, Figures 5b - 5d show that 2002-2004 SY API scores were also smooth across the award eligibility threshold. The close correspondence between the polynomial fits to the API (solid lines), the regression adjusted average API scores at each distance to the award threshold (open circles), and the F-statistic reported in Table II indicate that the estimates are not artifacts of our modeling choices. The estimates in Table II Panel A confirm that the 2000 award program had no effect on API scores and suggest, if anything, it hindered achievement. Panel B indicates that the 2001 SY awards program did not affect achievement either. Across both award years, the estimates are neither statistically nor economically significant. The standard deviation of the API is 108 in the 2001 SY and declines monotonically to 82 in the 2004 SY. Thus, even our largest estimate implies an API increase of less than 4% of a standard deviation. Online Appendix Table IV, which is analogous to Table II but based on local linear regressions using a bandwidth chosen according to the ROT procedure described in Lee and Lemieux (2009) for a rectangular kernel, confirms that these conclusions are not driven by our estimation approach. The same patterns hold for subgroup API scores, including the scores of the subgroup that determined a school's award eligibility.

We have also examined other measures of achievement – the share of students that test proficient in English and language arts (ELA) 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 were incorporated into the API over time (in the 2002-2004 SYs)– making them imperfect complements – they have the advantage of being reported in a transparent form. Yet, for neither the 2000 nor 2001 SY award programs can we detect improvements in either ELA or math proficiency rates in schools that just qualified relative to schools that just missed qualifying for an award. Analyses using API or proficiency gain

scores, noisier measures of achievement, yield similar conclusions.

Finally, we analyzed heterogeneity in achievement gains by school characteristics. As shown in Online Appendix Table IV, supplementing our main specification with an interaction (and all relevant main effects) between our treatment dummy and (1) an indicator for whether the school had a numerically significant disadvantaged subgroup in Panel A, (2) indicators of school type in Panel B, or (3) indicators of tertiles of the enrollment distribution in Panel C, we find no evidence of heterogeneous achievement effects. While in some cases we cannot reject differential effects of the treatment by sub-sample (i.e. the interactions are significant), in no case is the full effect of the award program on achievement – whether for schools with socially disadvantaged subgroups, different grade levels, or different enrollments – distinguishable from zero.

4.2 Evidence on School Resources

One reason for these null findings, other than the possibility that resources do not translate easily into academic achievement, may be that GPAP funds were not used for instruction. This could happen, for example, if districts, which have fiscal authority over schools, or the state, which provides much of the funds to districts, offset the awards through reductions in other funds (see Baicker and Jacobson (2006) and Gordon (2004) for evidence of budgetary offseting in local public agencies). Under either scenario, "winning" schools might not have additional resources to invest in achievement. Alternatively, since schools had considerable discretion in using GPAP funds, needing only local school board approval in principle, awards may not have been used in ways that improve academic achievement.

[Insert Figures 6a, 6b, 6c, 6d here]

Unfortunately, revenue and expenditure data are reported only at the district level, which limits our ability to determine whether an individual school receives its award money and, if it does, how it gets spent. Nonetheless, we have some school-level inputs such as the number of teachers overall and by subject as well as the number of instructional computers and internet-connected classrooms. Table III panels A and B present 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 and 2002 academic years, respectively. Panels C and D present estimates of the 2001 SY award program on the same category of outcomes but for the 2003 and 2004 academic years. We consider these years because the 2001 SY award disbursements were not made until July and October 2002 (see Figure 1). Figures 6a-6d graphically represent our RD estimates of the 2000 SY award program on computers per pupil in each of the 2001- 2004 SYs.

[Insert Table III here]

We find little evidence that either award program affected these inputs. Given that the GPAP was short-lived and hiring requires a long term fiscal commitment, it is not surprising that the number of teachers was unchanged. Schools may have anticipated winning an award in the second year but this expectation should have been low given that the standards for winning an award were raised between the two award years (see Figure 2), California's budget outlook put award funding at constant risk, and most importantly, the use of API growth rather than levels to determine award eligibility, made it difficult for schools to anticipate their eligibility or consider the awards a reliable source of funding. On the other hand, schools might have used the additional funds to encourage some instructors to switch from their normal subject to one that is more valuable in an accountability system, such as math or English. But, our estimates of the impact of the award programs on the share of teachers by subject are neither statistically nor economically significant.

[Insert Table III here]

That said, we may not have power to detect increases in FTE. To see this, note that approximately $60 per pupil was allocated to winning schools from the GPAP. These schools were also entitled to half of the SEPB award or about $49 per pupil. With average enrollment of 800 in winning schools, this amounts to $87,200, enough to hire 1 to 2 additional teachers or 0.00125-0.0025 FTE per pupil at about $45,000 per year. With a standard error of 0.003 in 2002, increased hiring would not be detectable even if if all funds were

put to this use. Although more difficult to quantify, providing bonuses or using the funds for teacher training to encourage more existing FTE to teach "high-stakes" subjects would have required significantly fewer resources and be more easily detected.

The CDE encouraged using awards "for the purchase of computers, instructional materials, or playground improvements" (Chladek 2002). Although we should have ample power to detect purchases with the 2000 awards, we find no evidence that awards increased the number of computers or internet-connected classrooms in a school. Results from the 2001 SY award program suggest, if anything, that computers per pupil increased less among schools that just qualified than those that just missed qualifying for awards. This interpretation should be viewed with caution, however, as our goodness of fit statistic and its p-value (.0002) suggest that the estimated reduction in computers per pupil (-.020 with a standard error of .006) is driven by functional form. The more important lesson from Table III is that we find little evidence of increased resource allocations among schools receiving GPAP awards. These conclusions are supported by our local linear regression estimates in Online Appendix Table V.

4.3 Evidence on Fiscal Outcomes

Since GPAP funds did not increase instructional resources, we test for fiscal crowd-out. Crowd-out might occur if the state offsets awards or, alternatively, disproportionately increases funds to schools that just barely qualified for the GPAP relative to those that just missed qualifying. One difficulty in studying this issue is that revenue and expenditure data are available only at the district level. As such, we view this analysis as suggestive.

To characterize districts based on school-level award eligibility thresholds, we sort all schools in a district by their distance from the award threshold. We assign to each district the *maximum* of the school-level "award gaps." Thus, each district's "award gap" is determined by its best-performing school. If any treated school in a district is far from the award eligibility threshold, then the district will be characterized as far from the cutoff. If no schools are treated but at least one is close to its target, then the district as a whole will

be characterized as just barely missing award eligibility. Alternatively, we can characterize districts by the *minimum* of the school-level award gaps. In this case, treatment implies all schools in a district won awards. Since districts composed of only winning schools are small (with a mean of 3 and median of 1 school per district) and the control groups of districts with at least one losing school comparatively large (with a mean of 11 and median of 6 schools per district), we opt for the *maximum* definition as our main approach. As discussed below, however, our results are not sensitive to this or other alternative definitions. Furthermore, Online Appendix Table VIII presents a set of alternative estimates that account for variation in district size. Importantly, these estimates are similar to the ones reported below.

Before moving to our results, we first demonstrate in Panels C and D of Appendix Table III that the observable characteristics of districts are generally smooth through the discontinuity. Districts just above the target do have a smaller share of students receiving free or reduced price meals and smaller shares of schools with disadvantaged and American-Indian subgroups. But in 11 of 14 cases, we cannot reject that districts close to the eligibility threshold are similar on predetermined characteristics. This finding supports our classification of districts based on the school that performs best relative to its award target.

[Insert Figure 7 here]

Figure 7, the district analogue to Figure 4a, shows the mean district-level apportionment per pupil in 2001 by proximity to the awards threshold and the polynomial fits to the data. To save degrees of freedom, we estimate polynomials with equal slopes on either side of the threshold. We exclude covariates as these tend to decrease rather than increase the precision of the district estimates. The figure (and column 1 in both panels of Table IV) establish district-level treatment. Because districts may have schools that won awards of varying amounts and schools that did not win awards, the discontinuity is below the school-level estimate. For both 2000 and 2001, districts to the left of the eligibility threshold (21% of all districts in 2000 and 30% in 2001) did not qualify for awards and thus have per pupil award payments of approximately \$0. At the discontinuity award payments jump to about

$42 per pupil for 2000 and about $28 per pupil for 2001 test performance. To relate this to the school-level estimates, one can multiply these estimates by the share of schools in treated districts that received apportionments. For instance, in 2000 this is 42*(5.11/7.73) which is $64, close to the $62 found at the school-level. While the district estimates are quite noisy, the goodness of fit statistics suggest the parametric models are reasonable. Local linear regression estimates in Appendix Table VII are also quite similar.

[Insert Table IV here]

GPAP apportionments are classified by the CDE's Fiscal Services Division as unrestricted revenues. We find that per pupil unrestricted revenue increases by more than the $42 apportionment shown in Figure 7. The RD estimate (reported in Table IV and shown in Figure 8) indicates a $103 per pupil jump at the discontinuity. Thus, "winning" districts received closer to $2.50 per pupil for every dollar they were supposed to get through the 2000 awards program. The local linear regression estimate is similar at $2.00 per pupil and confirms significant crowd-in. To the extent that our RD provides quasi-random assignment, other funds included in unrestricted revenues should not differ systematically at the award threshold except through the (direct and indirect) effects of the program. Unfortunately specific sources of revenue within this category are not available from district-level fiscal data. Some of the additional funds may be attributable to the SEPB program, which was also paid based on 2000 performance. SEPB disbursements were not released by the CDE but half of the $350 million from this award were to be shared with schools. If used as intended, they would have increased the funds flowing to schools by about 75%.

[Insert Figures 8, 9, 10 here]

We also study the impact of awards on total revenues per pupil. For the 2000 award program, we estimate a jump in per pupil revenues in 2001 of about $340 (reported in Table IV and Figure 8). Estimates based on expenditure data are quite similar. Per pupil expenditures in 2001 jump by $343 dollars in response to the 2000 awards program (see Table IV and Figure 9). Together with the evidence on unrestricted revenue, these estimates suggest that total per pupil revenues increased more than dollar for dollar as a result of the

2000 award program. Because awards for 2000 SY performance were paid out over the 2001 and 2002 SYs, we have studied the impact of the program on revenues and expenditures in the 2002 and 2003 SYs. In neither year, do we find any convincing evidence of a jump in revenues or expenditures at the award gap.

The 2001 award program offers a cleaner test of crowd-in since the SEPB was suspended after 2000. Consistent with the allocation of 2001 awards in the 2002-03 fiscal year, we do not detect an increase in unrestricted revenues until 2003. Importantly, as shown in Table IV, the RD estimate indicates that unrestricted revenues increase by only $20 per pupil at the discontinuity. This implies that districts received only $0.72 for every every dollar they were supposed to receive through the 2001 awards program. However, we cannot reject that this point estimate is equal to one. Thus, the results for the 2001 program, when the only monetary awards at stake were from the GPAP, are also *inconsistent* with crowd-out.

The magnitude of the estimated discontinuities in total revenues and total expenditures per pupil in 2003 are both consistent with crowd-in. As a result of the 2001 award program, revenues per pupil increase by $123 per pupil in the 2003 SY. Similarly, per pupil expenditures in 2003 increase by about $200 in response to the 2001 awards. Neither estimate, however, is statistically distinguishable from zero. Local linear regression estimates for both the total revenue and expenditure categories are also quite noisy and in one case has the perverse sign. While we view the district level analysis with considerable caution, the estimates in columns (3) and(4) of Table IV are clearly inconsistent with district-level crowd-out.

While district-level estimates cannot capture the flow of resources to individual schools, they do suggest that significant crowd-out is unlikely and that resources may have flowed more than dollar-for-dollar to winning districts. This is likely the result of additional funds paid out by the SEPB but might also have occurred if schools or districts leveraged their success to raise outside funds. While California mandates fiscal equalization (equality in per pupil spending), some public schools get around this by setting up private foundations and other instruments (see Sonstelie, Brunner, and Ardon, 2000 and Betts, Rueben,

and Danenberg, 2000). To the extent that awards increase fundraising potential, winning schools/districts could have captured additional fiscal resources.

5 Summary and Conclusion

We analyze a relatively understudied feature of accountability systems - a financial award program for schools making "adequate" progress on state achievement exams. We focus on California, where for the 2000 and 2001 school years, schools that met or exceeded their accountability targets were eligible for monetary awards through the Governor's Performance Award Program (GPAP). In the 2000 SY, teachers and staff in winning schools were also eligible for the School Site Employee Performance Bonus (SEPB) and, in limited cases, the Certificated Staff Performance Incentive Award (CSPIA). Because the GPAP and SEPB were allocated based on a deterministic, discontinuous function of school (and subgroup) exam performance, we employ a regression-discontinuity design to evaluate how awards affected achievement and resource allocations. This design allows us to take advantage of the fact (verified in our data) that schools close to the eligibility threshold are similar but for award receipt and thus that award receipt close to the eligibility threshold is "as good as randomly assigned," much like in an actual randomized controlled trial.

We find that awards significantly increased the financial resources allocated to some schools and their staff. The average per pupil GPAP award was $60 and $50 based on performance in the 2000 and 2001 SYs, respectively. Schools qualifying for the 2000 SY GPAP also received SEPB (and in some cases the CSPIA) funds. Moreover, districts may have supplemented these funds. Based on the 2000 awards program, districts with schools that qualified for GPAP awards received increases of 5% of per pupil spending. Since anecdotal evidence suggests that GPAP awards were distributed as teacher bonuses, this amounts to an additional $1,900 per teacher in the first year of the program.

Despite the increase, we find no measurable improvement in standard metrics of achievement for those schools that received the award compared to those schools that did not. This may not be 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 additional resources were more akin to a random shock than a guaranteed income stream, schools may have had difficulty translating them into 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.

Our estimates show that financial awards had no impact on achievement in schools that won awards versus those that did not. However, we cannot assess if the program would have an impact if implemented in conjunction with other reforms, such as reduced class sizes or raising teacher salaries. California instituted class size reductions 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 improved student achievement. However, evaluating the program is fraught with empirical difficulties (see CSR Research Consortium, 2002). Our work also leaves open the question of whether the competition for awards itself raised student achievement across all schools in California. However, the instability of award funding land the group-based nature of the programs, which introduces a free rider problem among teachers, likely muted the incentive effect of the program.

In contrast to our findings on achievement, California's program, in particular its subgroup rules, have put diverse schools and schools that serve disadvantaged populations at greater risk of failure. Furthermore, because the accountability targets were tied to financial awards, the program may have had the unintended consequence of diminishing the relative resources available to these schools.

References

Associated Press, 2001. "School reward program at-a-glance," April 11, 2001.

Bacolod, M., DiNardo, J. and M. Jacobson, 2009. "Beyond Incetives: Do Schools Use Accountability Rewards Productively," NBER Working Paper 14775.

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

Betts, Julian, Kim Rueben, Anne Danenberg, 2000. "Equal Resources, Equal Outcomes? The Distribution of School Resources and Student Achievement in California." Public

Policy Institute of California Report.

Carroll, Stephen J., Cathy Krop, Jeremy Arkes, et al., 2005. "California's K-12 Public Schools: how are they doing?" RAND Education Report.

Chakrabarti, Rajashri. 2007. "Vouchers, Public School Response, and the Role of Incentives: Evidence from Florida,"Federal Reserve Bank of New York Staff Reports, 306.

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.

Chiang, Hanley, 2009. "How Accountability Pressure on Failing Schools Affects Student Achievement,"*Journal of Public Economics*, 93(9-10): 1045-1057.

Chladek, Patrick J, 2002. Presentation at the November 2002 National Conference on Teacher Compensation and Evaluation, http://www.wcer.wisc.edu/CPRE/conference/nov02/chladek.pdf.

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

Fan, Jianqing and Irene Gijbels, 1996. *Local Poynomial Modelling and Its Applications*. Chapman and Hall, London.

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.

Folmar, Kate, 2001. "Lawsuit Put Teacher Bonuses on Hold,"*San Jose Mercury News*, August 9, 2001.

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

Hausler, Alexa, 2001. "Governor Pledges to Spare Cities' Car Tax Dollars When Cutting Budget." *Associated Press State and Local News Wire*, December 19, 2001.

Kane, Thomas J. and Douglas O. Staiger, 2002. "The Promise and Pitfalls of Using Imprecise School Accountability Measures," *Journal of Economic Perspectives*, Fall, pp 91-114.

Kerr, Jennifer, 2000. "$667 Million in Rewards for Schools, Teachers Approved by Board." *Associated Press State and Local News Wire*, July 12, 2000.

Ladd, Helen and Elizabeth Glennie, 2001. "A Replication of Jay Greene's Voucher Effect Study Using North Carolina Data," in Martin Carnoy ed*Do School Vouchers Improve Student Performance?*, Washington DC: Economic Policy Institute.

Lee, David S. and David Card, 2008. "Regression Discontinuity Inference With Specifica-

tion 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.

Lee, David S. and Thomas Lemieux, 2009. "Regression Discontinuity Designs in Economics," NBER Working Paper 14723.

McCrary, Justin, 2008. "Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test," *Journal of Econometrics*, 142(2): 698-714.

Muralidharan, Karthik and Venkatesh Sundararaman. 2011. "Teacher Performance Pay: Experimental Evidence from India," *Journal of Political Economy*, Forthcoming.

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

Rouse, C., J. Hannaway D. Goldhaber and D. Figlio, 2007. "Feeling the Florida heat? How lowperforming schools respond to voucher and accountability pressure. ,"NBER Working Paper 13681.

Savin, N.E. 1980. "The Bonferroni and the Scheffé Multiple Comparison Procedures," *The Review of Economic Studies*, 47(1): 255-273.

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

Sonstelie, Jon, Eric Brunner, and Kenneth Ardon, 2000. "For Better or For Worse? School Finance Reform in California." Public Policy Institute of California Report.

Thistlewaite, D. and D.E. Campbell, 1960. "Regression discontinuity analysis: An alternative to the ex-post-factor experiment," *Journal of Educational Psychology*, 51: 309-317.

van der Klaauw, Wilbert, 2008. "Breaking the link between poverty and low student achievement: An evaluation of Title I" *Journal of Econometrics*, 142(2): 731-756.

Table I
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)
Asians	749	708	782	737	761
	(121)	(124)	(111)	(116)	(120)
Hispanics	574	551	589	558	582
	(86)	(84)	(84)	(85)	(85)
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)
Asians	22	9.3	21	22	37
	(28)	(24)	(28)	(27)	(25)
Hispanics	31	13	28	30	49
	(35)	(32)	(38)	(32)	(26)
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. Means are based on data from the 2000 and 2001 SY. Standard deviations are given in parenthesis.
2. Zeros are not counted in the award payment calculations in this table.
3. To improve legibility, we omit American Indians, Pacific Islanders and Filipino's; these subgroups are rarely "numerically significant."

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

	Panel A: 2000 SY Awards Program			
	2001 API Score	*2002 API Score*	*2003 API Score*	*2004 API Score*
Mean	689	697	721	708
Treatment	-3.56	-3.38	-5.58	-4.16
	(5.75)	(6.24)	(4.99)	(4.40)
F-statistic	1.02	.949	.828	.957
p-value	.416	.679	.960	.651
	Panel B: 2001 SY Awards Program			
	2001 API Score	*2002 API Score*	*2003 API Score*	*2004 API Score*
Mean		699	723	706
Treatment		.502	1.97	3.35
		(3.79)	(3.92)	(4.52)
F-statistic		.974	1.04	1.01
p-value		.591	.350	.454

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

Panel A: 2001 Allocations Relative to 2000 Award

	Log FTE Per Pupil	*Share FTE Math*	*Share FTE English*	*Computers Per Pupil*	*Internet Connections Per 100 Pupils*
Mean	.048	.041	.068	.149	3.21
Treatment	.009	-.00003	.003	-.005	-.006
	(.007)	(.002)	(.003)	(.006)	(.306)
F-statistic	1.11	.745	.884	1.63	1.07
p-value	.152	.997	.875	.0000	.252

Panel B: 2002 Allocations Relative to 2000 Award

	Log FTE Per Pupil	*Share FTE Math*	*Share FTE English*	*Computers Per Pupil*	*Internet Connections Per 100 Pupils*
Mean	.049	.044	.068	.170	3.94
Treatment	.003	-.001	.001	.003	.074
	(.008)	(.001)	(.002)	(.009)	(.261)
F-statistic	1.05	.756	.920	1.23	.924
p-value	.317	.995	.780	.017	.766

Panle C: 2003 Allocations Relative to 2001 Award

	Log FTE Per Pupil	*Share FTE Math*	*Share FTE English*	*Computers Per Pupil*	*Internet Connections Per 100 Pupils*
Mean	.049	.045	.068	.189	4.56
Treatment	-.010	.0002	-.003	-.020	-.490
	(.008)	(.002)	(.003)	(.006)	(.291)
F-statistic	1.49	1.10	.961	1.40	1.08
p-value	.000	.167	.642	.0002	.205

Panel D: 2004 Allocations Relative to 2001 Award

	FTE Per Pupil	*Share FTE Math*	*Share FTE English*	*Computers Per Pupil*	*Internet Connections Per 100 Pupils*
Mean	.048	.045	.065	.198	4.85
Treatment	-.009	.003	.0001	-.010	-.308
	(.010)	(.008)	(.002)	(.006)	(.252)
F-statistic	1.42	1.04	1.03	1.31	1.02
p-value	.000	.351	.374	.002	.408

[a]Notes:

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

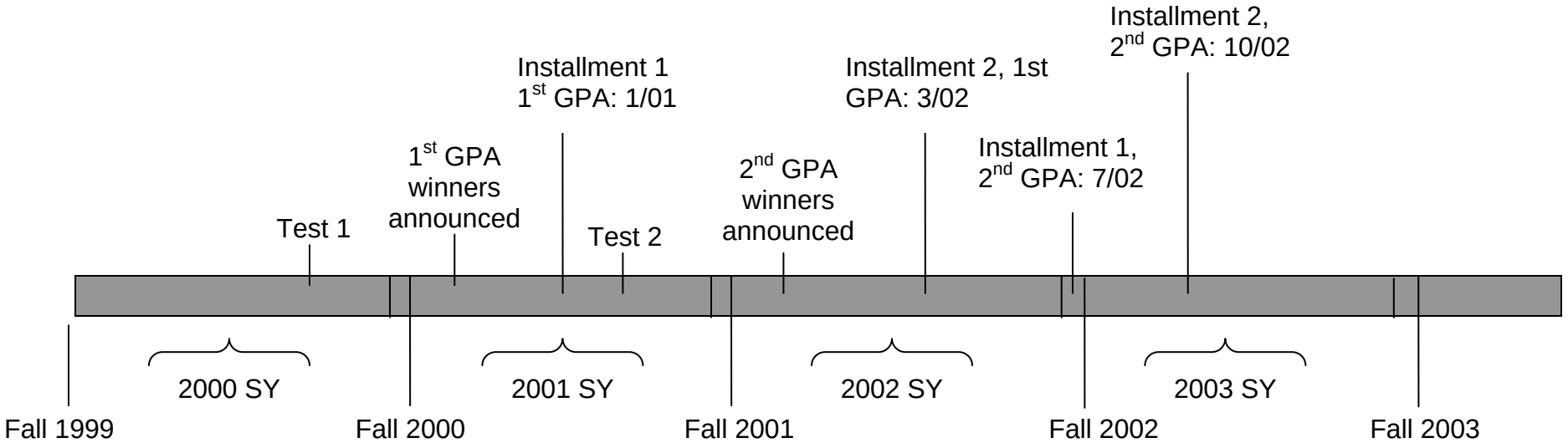
Table IV
Impact of the Award Program on District Resources[a]

Panel A	*2001 Per Pupil Allocations Relative to the 2000 Award Gap*			
	Award Apportionment	*Unrestricted Revenue*	*Total Revenue*	*Total Expenditures*
Treatment	41.8	103	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.

Figure 1. Timeline of Governor's Performance Award Program Announcements and Payouts



Notes: The Certificated Staff Performance Incentive Awards (CSPIA) and the Schoolsite Employee Performance Bonus (SEPB) were only in effect for the 2000 school year. They were each paid out in one installment – October 2001 for the CSPIA and March 2001 for the SEPB.

Source: History of Apportionments – Governor's Performance Awards, California Department of Education.
Previously available at http://www.cde.ca.gov/ta/sr/gp/history.asp Accessed on 12/17/2005. Available in hard copy from the authors.

Figure 2. API Growth Required to Qualify for Governor's Performance Award as a Function of School's Base API Score: 2000 and 2001 SY

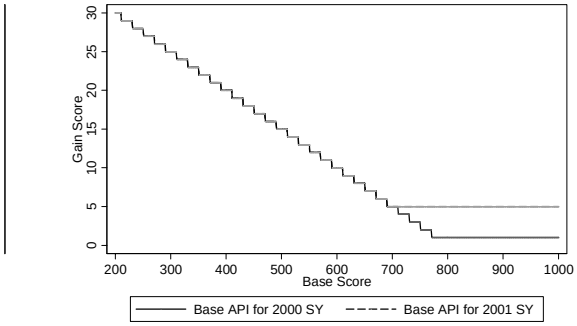


Figure 3a. Share of Schools with an Award for 2000 SY Performance Relative to the Distance to the Awards Eligibility Threshold

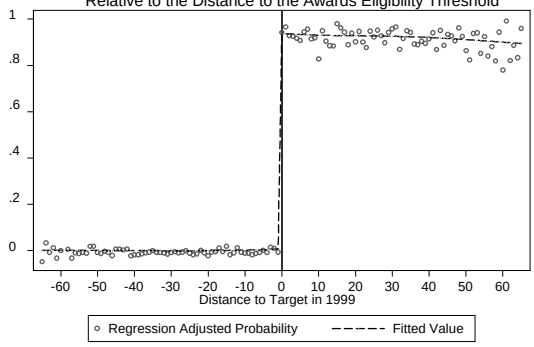


Figure 3b. Share of Schools with an Award for 2001 SY Performance Relative to the Distance to the Awards Eligibility Threshold

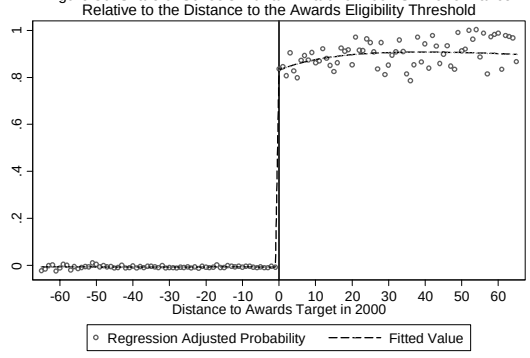


Figure 4a. Per Pupil Award Payment for 2000 SY Performance Relative to the Distance to the Awards Eligibility Threshold

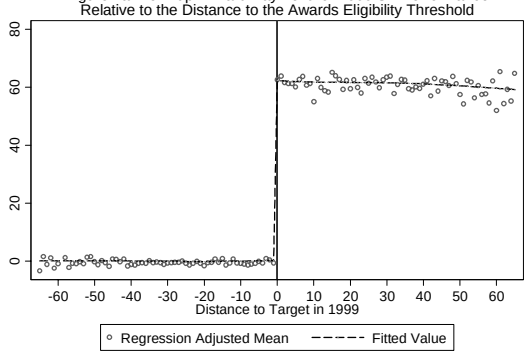


Figure 4b. Per Pupil Award Payment for 2001 SY Performance Relative to the Distance to the Awards Eligibility Threshold

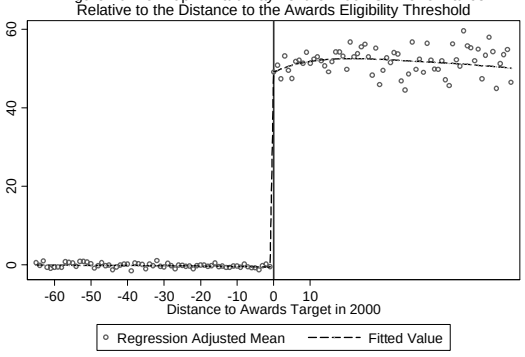


Figure 5a. 2001 SY API Score
Relative to the 2000 SY Eligibility Threshold

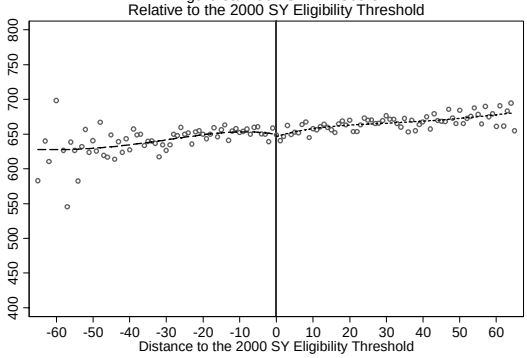


Figure 5b. 2002 SY API Score
Relative to the 2000 SY Eligibility Threshold

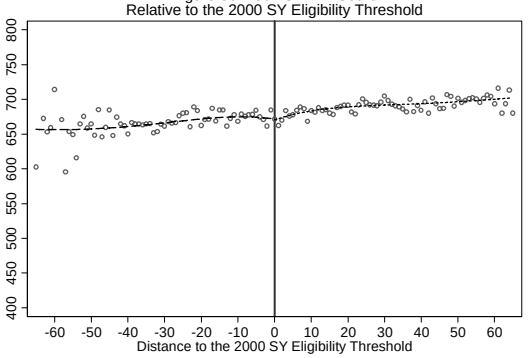


Figure 5c. 2003 SY API Score
Relative to the 2000 SY Eligibility Threshold

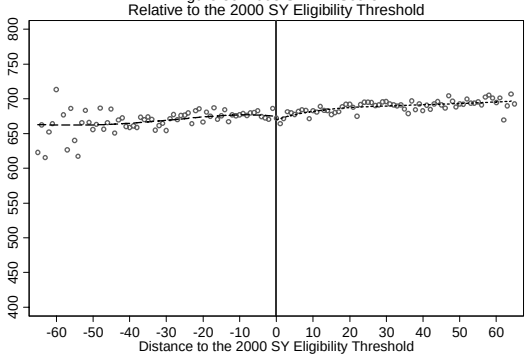


Figure 5d. 2004 SY API Score
Relative to the 2000 SY Eligibility Threshold

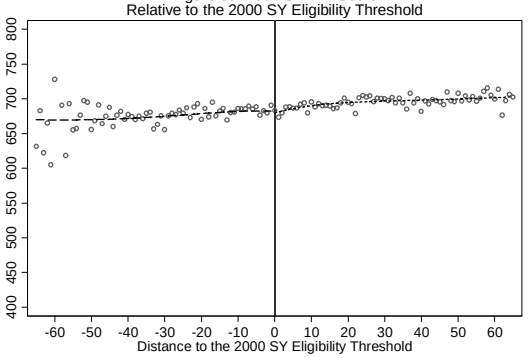


Figure 6a. Computers Per Pupil in the 2001 SY
Relative to the 2000 SY Eligibility Threshold

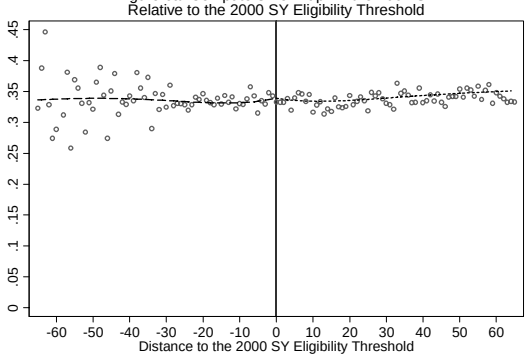


Figure 6b. Computers Per Pupil in the 2002 SY
Relative to the 2000 SY Eligibility Threshold

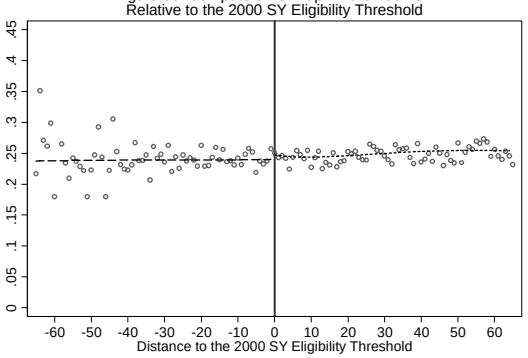


Figure 6c. Computers Per Pupil in the 2003 SY
Relative to the 2000 SY Eligibility Threshold

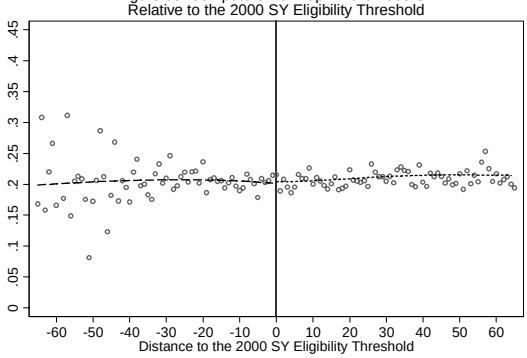


Figure 6d. Computers Per Pupil in the 2004 SY
Relative to the 2000 SY Eligibility Threshold

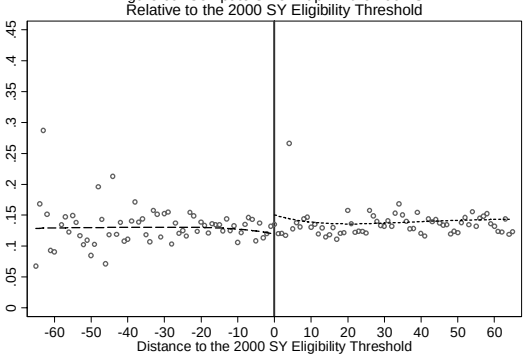


Figure 7. District-Level Per Pupil Award in the 2000 SY
Relative to 2000 SY Eligibility Threshold

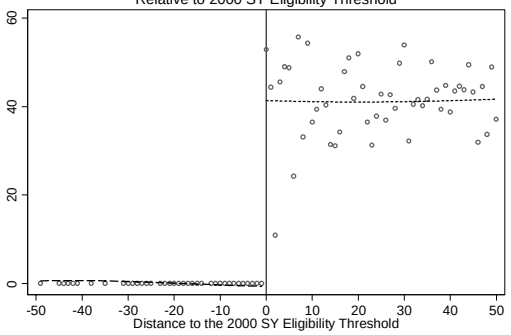


Figure 8. District Awards Category Revenue Per Pupil in 2001
Relative to 2000 SY Eligibility Threshold

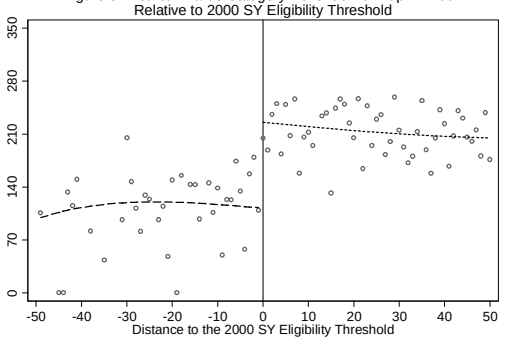


Figure 9. Total Per Pupil Revenue in 2001
Relative to 2000 SY Eligibility Threshold

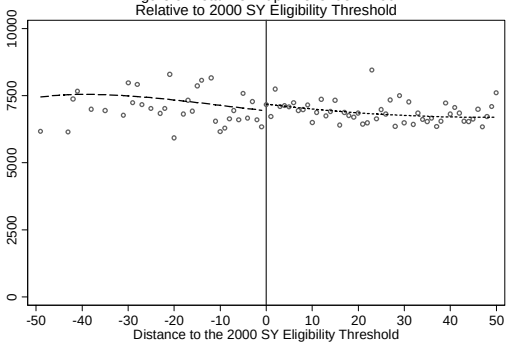


Figure 10. Total Per Pupil Expenditures in 2001
Relative to 2000 SY Eligibility Threshold