

Incentive Effects of the No Child Left Behind Act in California Schools

Vivian Hwa
UC Berkeley

November 2006

Abstract

This paper evaluates the effect of the No Child Left Behind Act of 2001 (NCLB) on school performance and incentives. The NCLB establishes strict proficiency targets that schools must meet in order to avoid sanctions. Specifically, I argue that schools close to a target share similar characteristics, and random exogenous shocks will force some schools to miss and fail. Failing schools are expected to increase their efforts/costs as a response to impending punishments, creating a discrete gap in score growth as compared to other schools near the threshold that randomly passed. Using a regression discontinuity (RD) approach, my results agree with similar research that the incentive effects of failing are small; however, the interpretation is difficult without a more formal theoretical framework. To help interpret the RD results, this paper includes a dynamic optimization model that represents a school administrator's imputed investment in future scores given different states of failure. This simple behavioral model determines optimal levels of investment, or "costs", where the payoffs are linear in costs and the cost-to-proficiency gains production function is concave with an unknown elasticity. In this framework, the RD estimate has a precise interpretation: a function of the relative difference in investment between failing and passing schools and the elasticity of score production to investment. Under the specification where the largest potential incentive effects of the NCLB are expected, the joint solution to the RD and structural model implies that failed schools invested more but the low elasticity (around 0.04) did not allow for large gains in scores.

I would like to thank my advisors David Lee, David Card and Alexandre Mas for their guidance and excellent direction. Thanks also to seminar participants for their comments, especially to Jerome Adda and Enrico Moretti. Special thanks to Paul Chen for his valuable comments and edits.

1 Introduction

The American public is increasingly concerned about the state of public schools. The Program for International Student Assessment (PISA) released in 2000 showed that U.S. students were about average in reading but lagged in math when compared to other OECD countries. In the 2003 PISA, the United States had fallen behind countries such as Poland, Hungary and Spain by some measures of math proficiency and only achieving an average rank of 24 out of 29¹. While the debate continues on the validity of such assessments, it seems that increasing amounts of money and resources are being used for improving education without corresponding increases in graduation rates, SAT scores or national test scores². The U.S. has steadily increased per pupil expenditures around 3.5% a year from 1890 to 1990 (Hanushek and Rivkin, 1997) and the U.S. Department of Education has seen its budget grow over 50% in just the last 5 years³. Currently, one of the heavily debated issues in the education field involves looking at increases to educational funding and if it is the most effective way to raise American public school achievement and preparing a generation of high-skilled workers (Hanushek, 2003). There are sentiments, even within the U.S. House committees, that additional funding for certain educational programs is like “pumping gas into a flooded engine” (U.S. House Committee Report, 2004).

In response to this, Congress passed the No Child Left Behind Act of 2001 (NCLB) which represents the most significant departure in K-12 education policy of the past four

¹The PISA 2003 results are available online at <http://www.pisa.oecd.org>. A report on the U.S. results from PISA 2003 can be found online at <http://nces.ed.gov/surveys/pisa>.

²It is also commonly recognized that the demographic of American students has changed much over the last few decades with more students taking the SATs, increasing numbers of students from poor families, and an increase of English learners.

³From 2002 to 2005, the budget for the U.S. Department of Education increased from \$46.3 to \$71 billion dollars where 64-68% of that funding went to non-higher education programs. (Data from the U.S. Department of Education)

decades. Its main feature is the requirement of strict testing and reporting standards for all public schools and mandated improvement measures for schools that consistently perform poorly. Although schools historically assessed their performance on the basis of school-wide averages, now schools must report results for specific subgroups of students and ensure that each subgroup is achieving proficiency targets. The underlying premise of the NCLB is that holding schools accountable—by threatening the autonomy of principals, teachers and school boards—should have large positive effects on education production.

The existing literature that evaluates this premise suffers from two key limitations. First, many estimates of the effect of NCLB are potentially biased due to unobserved heterogeneity. Second, these estimates are difficult to interpret without an underlying structural model since the incentive effects could be strong enough to change behaviors but has no direct effect on school performance.

This paper investigates the consequences of educational accountability in the following three ways. First, it provides an estimate of the effectiveness of NCLB on increasing scores using a regression discontinuity design, thereby overcoming some of the endogeneity issues. Second, it develops a dynamic model of a school administrator’s decision process. This enables me to interpret my RD estimate—or even other NCLB estimates—in terms of its effects on the actual education production process and the incentives of school administrators. Third, it contributes to the burgeoning literature⁴ on the effects of incentives schemes on desired outcome and highlights the importance of modeling the causal link between incentives and outcomes when evaluating the effects these systems of rewards or punishment.

⁴Studies on the effectiveness of incentives include many papers on retirement savings schemes, teacher incentives, sales/management incentives, etc.. For example, there is a study by Friedman and Kelman (2006) where hospitals in England were rewarded if they could lower their average emergency room waiting time under a certain target. Preliminary results showed increases in performance right after the announcement and throughout the duration of the program, indicating that the hospitals were completely internalizing the incentives and making early changes to take advantage of the impending rewards.

My main finding is that NCLB has positive incentive effects but weak performance effects because of the limitations in the education production of schools. According to my most favorable RD estimates, NCLB had a negligible effect on school performance, as measured by proficiency rates in English and math. However, according to the structural model, NCLB had a positive effect on the incentives of school administrators, who exerted greater effort in response to pressure from the NCLB. The reason this effort did not translate into significant increases in proficiency rates is due to the low elasticity of test scores with respect to effort. Therefore, we should not expect strong results from educational accountability reforms without significant improvements in the education production technology.

There is a limited literature that estimates the effects of NCLB on student performance. According to the U.S. Department of Education, the main achievement for NCLB in California between 2003 and 2005 was that fourth-grade proficiency increased by eight percentage points in English and five percentage points in math. This type of simple analysis relies on defining progress as average changes in scores over time. Unfortunately, it is unclear if all of those gains can be attributed to the program or perhaps declining enrollment or other demographic shifts during the same time period. Since this is a state average, there could be large gains for some subgroup of students while growth has stagnated for other groups.

When schools have such high levels of unobserved heterogeneity, there may be various reasons unrelated to the NCLB that can explain why schools fail or experience high growth rates. For example, low performing schools may experience high rates of improvement due to mean reversion, which could overestimate the effects of the NCLB program (Chay et al., 2005). Another example would be that comparisons between schools far below and above the targets could be inaccurate due to large differences in school size or racial composition.

In this paper, I work through these issues by focusing only on schools near the target, so

there is less of a chance that schools are completely different in every dimension but somehow obtained similar proficiency rates. Continuing with the idea that “effectiveness” should be measured as increases in proficiency rates, the identification strategy used in this paper tries to isolate the schools for which the NCLB incentives are expected to be the strongest and most binding, namely schools near the cutoffs. The data is constructed to include just the schools where NCLB failure depends only on the single dimension of test proficiency. This study will use a regression discontinuity (RD) approach which focuses on schools that “just” failed in one year and schools that “just” passed proficiency targets. The logic is that very similar schools near the target would find that random shocks would push some schools over the target and others under. This resembles random assignment, thus the gap in score growth is an estimate of the program treatment effect. I look at a each school’s worst indicator for the year—the lowest recorded proficiency rate relative to a target—and try to find if those proficiency rates have different levels of growth in the following year after being labeled failing or passing. By including several years of data and using a non-linear specification, the estimated effects seem to be small. Doing RD estimation on educational data is not a new idea. Previous research has shown that with minimal assumptions, a similar methodology can be implemented for data from several different accountability programs, but this paper differs from past studies as it incorporates the incentive structure into a model such that the estimates can be interpreted within a framework.

Small estimated effects from the literature have not been conclusive enough to rule out the incentive effects of NCLB. When an RD result is small or negative, there are often competing explanations other than the “no effect” conclusion. Schools may not have been aware of how the system was supposed to work, or a small RD estimate could imply that the punishments might have been too remote for a large incentive effect or perhaps schools

did not have the capability or resources to improve.

To help understand some of the confounding explanations, I create a basic framework for interpreting the results using a structural model. I express a school administrator’s decision to invest in increasing their proficiency levels as a stochastic dynamic optimization problem. To start, I will construct a simplified “2-strikes you’re out” accountability system. This approach highlights the conditions under which one should expect differences in behavior near critical thresholds. The main result is that the regression discontinuity estimate can be defined as a simple function of the relative difference in investment between failing and passing schools and the elasticity of proficiency rates to investment. Since actual investment is not measurable, the model is based on an imputed cost which can be interpreted as the proportion of the rent an administrator receives from not having two strikes. Furthermore, the model is flexible enough to accommodate several different estimators from the data with small adjustments; i.e. using changes in graduation rates or test participation instead of proficiency rates.

Under the specification where the largest potential incentive effects of the NCLB are expected, the joint solution to the RD and structural model indicates that failed schools invested more but the low elasticity (around 0.04) did not allow for large gains in scores. Within the framework that I develop, the magnitudes of the estimates in the RD imply that the elasticity of the education production function is likely to be small, which an RD design alone would have not been able to address. If the estimate of elasticity is taken at face value, it implies that investigating avenues to increase the effectiveness of effort (or costs) on raising scores could be extremely useful—where the incentive schemes already in place can have even larger effects once elasticities are raised.

The remainder of this paper is organized as follows. A brief review of the current

literature is followed by some background information on the No Child Left Behind Act. Next, I will discuss the data and empirical strategy. After discussing the results from the empirical section, part six will explain the structural model. Part seven shows how the empirical estimate and structural model are solved together. I will end with further extensions and conclusions.

2 Literature

Existing literature regarding NCLB seeks to isolate the effect of NCLB on test scores or other academic indicators using similar RD methodology. A working paper by Figlio and Rouse (2005) looked at Florida's implementation of NCLB and focused on voucher threats and stigma effects. They separated their data into 4 groups: schools that faced the threat of vouchers, those that received an "F" grade, those that faced both problems, and those that faced neither. With school and student level data from a sub-sample of districts, they find that voucher threats were not as salient as stigma effects and score growth was higher for poor performing students in failing schools. Also, the relatively low gains on nationally normed tests indicated that after implementing the grading system, schools may have shifted attention away from non-high-stakes grades and extra effort was only found in the lowest performing schools. However, unobserved heterogeneity in schools also explained more of the score variation than any incentive effects from the system. This result agrees with a study by Kane and Staiger (2002) which shows that sampling variation and random noise accounted for almost 40% of the variation in year to year test score gains.

A different paper by Kane and Staiger (2002) looked at how being held responsible for a subgroup of students did not lead to higher gains for that group. Focusing on Texas

and California, the authors separated schools by percentage Latino and African-American students. Using state-defined thresholds for “subgroup significance”, they find no evidence that schools held responsible for their Latino or African-American subgroups experienced different proficiency gains from schools not held responsible for those groups. However, they point out that this could be because teachers find it hard to tailor their focus to students from certain subgroups, especially if classes are very diverse.

Directly related to this paper is an unpublished study by Forbes, Gordon, Rosaen, and Schwartz (2006) that uses a similar approach using only 2 years of California school data. Using the full sample and controls for select subgroups⁵, they find no significant effects on test score growth for schools that failed the previous year in either English or math.

While all of these studies rely on the assumption that punishments or threats should lead to changes that increase in school test performance, none of them directly model why one should expect such behavior on the part of school administrators.

3 Background on the NCLB Act

The NCLB marked the first time the federal government had threatened to remove education funding based on compliance of its rules. Before, the policies only offered general goals and ideals for states along with initiating Head Start and Title 1 funding for disadvantaged students. With the new law, four main “pillars” were established concerning accountability, flexibility, research-based education and parent options. The combination of the four objectives solidified the role of standardized testing of students and forcing schools receiving Title 1 money to comply at the risk of losing their funding. Over 90% of the school systems in

⁵The authors controlled for school size, percent African-American, Hispanic, Low-SES, white and English language learner.

the US receive some sort of Title 1 funds; and in California, that amount has been around \$1.7 billion in recent years.

The NCLB's main objective is to raise all students and subgroups (gender, race, socioeconomic status, learning disability, English as a second language) into state defined levels of proficiency in math and English Language Arts (referred to as "English"). Being labeled a passing school in California requires overcoming a list of criteria where failing any part will force you to fail. The key concept is Adequate Yearly Progress (AYP) and depending on the diversity and size of your school, the number of criteria needed to pass this can range from 2 to over 35.

The criteria are as follows:

- Tested at least 95% of all students and 95% of each subgroup?
- All students and each subgroup met required percent proficiency in math?⁶
- All students and each subgroup met required percent proficiency in English?
- Met API growth target or API minimum ?⁷
- If high school, met target graduate rate?

Answering no to any of the points would result in failure of AYP. The required proficiency rates are increased using a schedule that requires around a 10 percentage point increase in proficiency rates over the next 7 years in order to reach 100% proficiency by 2014. (see Figure 1)

⁶There are "safe harbor" systems in NCLB where schools that demonstrate adequate growth can still pass this criteria even if proficiency levels in ELA or math are under the target.

⁷API is the Annual Performance Index, which is a weighted average of all test scores for a school.

Since the NCLB started in 2001, the nation will see the most severe sanctions coming into effect in the year 2006. (see Table 1) NCLB is not a rewards system—it establishes pressure by reducing the autonomy of principals and schools boards when the school is performing poorly. Sample punishments include forcing school choice or school restructuring—such as reopening the school as a charter or replacing staff. If a school fails AYP two consecutive years⁸, they are put in Program Improvement (PI) status. Thus the last chance for a school to avoid PI is to pass after their first “strike”.

Currently in California, the most recent tests show that proficiency rates in English and math have risen to 24% and 27% respectively. However, while minority students and English learners experience high growth rates, they are still lagging behind and unable to close the achievement gaps. In 2006, around 2250 schools were in some stage of PI, and 80% of those schools progressed up the PI scale to harsher consequences.

4 Data

The California Department of Education provides standardized test scores from the years 1998 to the present from its Star Testing and Reporting Program. The STAR program consists of several tests given to students in the spring from grades 2-11 in various subjects including English and math. One of the tests is the California Standards Test (CST) which is administered to all California public school students except those with an exception form parents or with special needs due to learning disabilities. Learning disabilities are divided into categories where some receive special accommodations but took the regular test (or took it

⁸NCLB failure occurs when a school fails the same criteria/subject area two years consecutively. That is, if a school fails math proficiency in one year and ELA the next, they are not in PI status. However, if the Asian students failed ELA one year and then the Hispanic students failed ELA the next year, the school is on PI status.

at a lower grade level), while others were opted out completely and took an alternative exam. English learners are all required to take the regular test. For math, a general grade-level math test was given to all students while students in grades 8-11 had to take additional math tests including an Algebra I test and whatever subjects were appropriate. Test scores are scaled such that a score of 350 is proficient for any grade level and any subject. Proficiency rates for schools are given as the fraction of tested students who attained a score of 350 or above.

The data is given at the school level with data separated by grade, test subject and student type. The types are “all students”, Learning English Proficiency (LEP), male, female, economically disadvantaged, non-economically disadvantaged, LEP <12 months, LEP >12 months, special education and non-special education. For the years 2002 and later, the subgroups included racial/ethnic designations as well as parental education factors, Title I participation and free-lunch program participation. No scores are available for individual students, and grade/type cells that contain less than 10 valid test scores are considered missing.

The data for AYP provides the enrollment, number tested, valid scores, percent proficient, API scores and growth, and the targets met for both math and English⁹. Along with test data, graduation rates are reported if applicable. The Department of Education also provides demographic enrollment information, school characteristics, and teacher demographics at the school level and some financial data on the district level.

Targets are reported for elementary schools, middle schools and high schools in both

⁹A high percentage, 82%-92%, of schools meet their English or math targets. In 2005, the targets were raised for all schools and both subjects, which led to a 7-10 point difference in percentage proficiency rates. The data shows that on average, 83% of schools managed to pass both. After including all the other criteria for passing AYP, the fraction of schools passing all criteria went from 52% in 2002 to 65% in 2006. Of the schools that pass in any given year, over 90% pass in the next (this includes schools with very high levels of proficiency). Of the schools that fail, about 40% pass in the next year.

English and math proficiency rates, testing rates, graduation rates and API scores. Another variable is created to represent the difference between the actual results and the targets. A running variable is created for every school each year and it represents the criteria for which a school is farthest from its target. Since a school is considered failing if it fails any criteria, tracking the worst performing variable is sufficient, i.e. if the worst criteria for a school is over the target, then the school passes. Conversely, if the worst criteria is far below the target, then the school must fail. (see Figure 2)

The data set is composed of all observations over 2002 and 2005 that include growth (i.e. scores from 2003-2006) and demographic information. There are approximately 9000 schools per year and each school has graduation rates, API scores, test participation rates and test results for the entire school and by each subgroup. The full data set contains over 200,000 observations.

Since this study measures the effectiveness of incentives as increases to proficiency rates, a sub-sample of data is constructed using only schools that have passed all criteria except proficiency rates; therefore these schools will only fail AYP if they fail their proficiency targets and for no other reason. More specifically, the data looks at schools that have tested 95% of their students, passed the API criteria and passed the graduation requirement (if applicable). This allows the remaining schools to have only one dimension to focus on and the only reason a school would fail NCLB is due to low scores/proficiency rates. Using this kind of approach will provide an upper bound on the incentive effect of NCLB. It is important to note that small schools tend to have the largest variation in scores and score growth. There are also special and complicated rules used to calculate proficiency rates for tiny schools. To account for this, schools with less than 100 students were not included in the sample and growth rates are constrained to be between -100% and 100%. Because

the subgroup is reduced to non-small schools that have passed other criteria, the average proficiency rates tend to be a little higher but the other characteristics are relatively similar. (Table 2)

5 Empirical Strategy

In general, tracking averages or growth rates over time is adequate for getting a broad picture of the current state of student proficiency. The problem with these types of analyses is the lack of a “control” group. One could not identify NCLB as the cause of the changing scores or growth rates since it is impossible to know what the data would have looked like in a world *without* NCLB. Thus, an ideal program evaluation should utilize some experimental setup to identify exactly how the presence of NCLB has changed outcomes. However, all public schools in the United States were subjected to NCLB at the same time and even research strategies like event studies are difficult due to the immense changes in testing rules and data collection over time¹⁰. Instead of identifying the effectiveness of NCLB as increases in proficiency rates across all schools, it might be interesting to look at the effectiveness of NCLB for a subgroup of schools where NCLB has a direct effect. That is, develop an identification strategy such as the regression discontinuity approach, which isolates where NCLB effects are suspected to have the largest impact and testing to see if those impacts are positive.

The regression discontinuity approach is well developed in the program evaluation literature¹¹. The ideal RD design has several components. First, the treatment—defined as

¹⁰In the years before NCLB, California had used different tests and did not keep track of test scores for different racial subgroups. For other states like Texas, English-learners were formerly excluded from school and district level calculations and had to re-classify, re-count, and include them after NCLB.

¹¹See DiNardo and Lee (2004) on unionization, Hahn, Todd and Van der Klaauw (2001) on the theory,

program participation or eligibility—is dependent on an observed and continuous running variable and a target threshold. In a "fuzzy" RD, the probability of treatment conditional on the running variable must have a discrete discontinuity at the target threshold. One possible violation of that condition can be seen as manipulation of the running variable (sorting) near the targets which can be checked by looking at the density function of the running variable¹². Another violation would if treatment was concurrently dependent on other observed characteristics—i.e. racial composition or school size instead of proficiency rates. To check this, one can look for discontinuities along other dimensions and verify that treatment is only dependent on the running variable. Once the first stage is complete, a regression using a polynomial approximation to the underlying conditional expectation function should result in an estimate of the effect of NCLB incentives for failing schools near the target. This is true as long as the polynomial approximation has enough higher order terms to closely approximate the data and one can implement a straightforward test by seeing how the fitted values follow the local averages along the running variable.

Luckily, the NCLB creates its own quasi-experimental setting which allows me to identify incentive effects using an RD approach. This particular setup is appropriate for a regression discontinuity approach and the prediction is that schools just under the cutoff and fail would try harder to avoid punishment and experience higher proficiency rates in the following year. Since NCLB requires schools to meet proficiency rate targets in order to pass, following the progress of failing and passing schools can provide insight into how NCLB has affected student achievement. Without using an RD design, one can simply compare the failing schools against the passing schools; however, that type of analysis has its own

Van der Klaauw (1997) on financial aid, Angrist and Lavy (1999) on class size and test scores, and Lee (forthcoming) on elections.

¹²A more sophisticated test involving local linear density estimators can be utilized as illustrated in McCrary (2005).

drawbacks. On the one hand, schools do not voluntarily choose to fail, so self-selection into participating in Program Improvement is not a concern. Thus, the complications typically involved in program evaluation, like welfare programs or work training programs, is avoided. On the other hand, there are still selection problems that could arise through systematic differences other than proficiency rates that separate the failing and passing schools. For example, if all failing schools were larger and more diverse, then outcomes can potentially be attributed to those differences instead of the effect of being treated or deemed failing by NCLB. Furthermore, schools often alternate between passing and failing from one year to the next, which makes pooling the treated schools over time somewhat uninformative. If one were interested in how NCLB encourages failing schools to do better, one would ideally remove those schools from the fail group that had persistently low scores and low growth—as they have no incentive effect and no chance to make the target. Similarly for the untreated control group—the passing schools—where some will never feel any pressure from NCLB as they are far above the targets and would unlikely fail even with large negative shocks to their scores.

Using the proficiency gains as the indicator of NCLB incentives working, the RD estimate should correspond to where NCLB incentives are most likely to be felt. Data used for this analysis will be the sub-sample where schools have passed all non-test score criteria. Granted, “effectiveness” could have been defined as increases to graduation rates, test participation rates, or growth in API and the same empirical test could be carried out for each indicator. However, graduation rates are limited to high schools, using participation rate targets would result in too few observations and passing the API criteria depends on two indexes (meeting a target or 1 point growth) instead of just one. Taking this particular sub-sample of schools allows for an upper bound estimate on the incentive effect of NCLB

for only schools near the target, as it is not a representative random sample and might not generalize to the entire population.

First, regression discontinuity designs depend on exogenous variation around a threshold to mimic random assignment. Since test results are not perfectly predictable, schools who find themselves near the target are fairly similar and the small shocks to their scores will cause some to pass the target and others to fall short. The typical examples are fire alarms during testing, failing to properly explain the test instructions to the students, or an incidence of the flu excusing several students from taking the test. Otherwise, if teachers and principals could perfectly predict and generate scores, every school would try just hard enough to obtain proficiency rates at the target and no harder. A way to test for this is by looking at the distribution of proficiency rates—where spikes at the targets would indicate that there could be strategic behavior by schools (see Figure 3a, 3b). As the distribution of proficiency rates does not reveal abnormal spikes, this is consistent with non-strategic behavior. Schools can obtain a wide range of test results due to random variation and proficiency rates are calculated as the fraction of students who scored above 350 on the CST out of all the students tested.

As mentioned before, the treatment is going to depend on the distance that schools are from their target (scaled to zero) and this variable is actually based on the worst proficiency rate a school has that year¹³. To verify that treatment is discontinuous in the running variable, Figure 2 shows that it is true that schools over the threshold always pass. Recall that this sample is already controlled for passing all the other criteria. There are some schools below the threshold still passing AYP due to a “safe harbor” rule that is calculated

¹³The "worst proficiency rate" variable will either be for the school or a numerically significant subgroup. Numerical significance is determined by a rule determined by the state department of education which is a minimum of 15% of the school's total valid scores and at least 50 students or a minimum of 100 students.

for schools that appeal and display adequate growth despite being under their targets. Since these “safe harbor” schools represent around 13% of the data, it could potentially put a downward bias on my RD estimate. This situation calls for a "fuzzy" instead of sharp RD design as there exists an "intent to treat" since there is not actual compliance to the treatment. Thus, at the end of the estimation procedure I need to adjust the estimate of the impact of intent to treat for the proportion of schools under the target who do not fail and the proportion of schools that are under the target and fail (receive treatment).

The next step in an RD is to verify that the discontinuity appears only with treatment assignment and not with other variables. To test the validity, variables like percent black, percent Hispanic, percent English Learners, and enrollment are plotted against the cutoff to see if there are additional discontinuities (Figure 4). The regression results for discontinuities by demographic composition also show no discontinuities (Table 3). A good exercise before continuing with the RD is checking the averages near the threshold to see if there is a difference in score growth on either side. Figure 5 shows that there is not a visual discrete jump in a school’s next proficiency rate although an overall positive trend might exist. To actually estimate this effect, one can run the regression¹⁴:

$$\ln(y'_i) = \beta_0 + \beta_1 X_i + \beta_2 D_i + \beta_3 W_i + \beta_4 D_i W_i + e_i \quad (1)$$

where X are school characteristics, D is a binary variable indicating whether a school has passed or failed, and W is the running variable representing the school’s worst proficiency rate that year. The results are reported in Table 4. There does appear to be a large and

¹⁴School characteristics include percent subgroup, enrollment, proficiency levels, and school type (elementary, middle or high). The coefficients for type, enrollment, and some subgroups are very small in magnitude and/or statistically insignificant. Robust standard errors are computed which should also account for clustering at the school level.

statistically significant discontinuity, implying that schools that failed try harder to improve scores the next period. However, when the specification includes higher order polynomials and interactions¹⁵, the effect of failing on growth becomes smaller.

The results from the specification with higher order polynomials and interactions indicate that schools that fail have a slightly higher proficiency rate than schools that just pass. At this stage, it is difficult to interpret the results. A non-discontinuity or a negative discontinuity could indicate a range of underlying causes. One potential explanation is that schools do not understand the system and are not responding to the incentives. Another explanation could be that failed schools are truly increasing their efforts, but they are not seeing the desired results due to the low elasticity of score production to effort. There could also be problems with the assumptions about the exogenous shocks as they could be serially correlated. If schools under the target drew a negative shock in one period and the exogenous shock next period was highly correlated, then any added effort would be mitigated. This would also lead to a small RD estimate and one could not conclude that the NCLB had no incentive effects.

6 Structural Model

At this stage, the small RD estimates could be a result of several possibilities. If teachers do not understand NCLB or do not place a big emphasis on the program, there would be no large incentive effects for schools who fail or barely pass. The punishments may seem too remote or seen as benefits, so schools do not mind failing and would not respond to such

¹⁵

$$\ln(y'_i) = \beta_0 + \beta_1 X_i + \beta_2 D_i + \beta_3 W_i + \beta_4 D_i W_i + \beta_5 W_i^2 + \beta_6 D_i W_i^2 + \beta_5 W_i^3 + \beta_6 D_i W_i^3 + e_i$$

"threats". Another explanation could be that schools are aware of the punishments but due to a low elasticity of the education production function, are unable to achieve results despite putting forth large amounts of efforts to increase proficiency rates.

Since interpreting the results from the regression discontinuity can be difficult, it is crucial to think about the theory in a more formal way. The purpose of my structural model is to capture the decision making process of a school's administration regarding how much effort or extra resources to devote to increasing the performance of their students knowing that their effort is costly. In addition, the chosen level of effort in the current period affects whether or not the school gets in trouble in the next period. Once schools choose their optimal response, the results of their decisions should appear in the data. What is particularly interesting about this model, is that for minimal assumptions about discount rates and the distribution of the random error term, there are unique values of the optimal effort or "cost" used by failing and passing schools for given values of the education production elasticity parameter. This model also allows me to decompose the RD estimate into a logical function of parameter values; where the RD estimate is equivalent to the product of the elasticity of "costs" to future proficiency rates and the relative difference in "costs" used by failing and passing schools.

The decision maker in this problem will be the principal of the school. To simplify, the principal receives a payoff of b if employed and s if fired. The principal has to choose a level of effort and resources, costs c , that directly affects test scores. In other words, their total payoff per period is either their employment benefit minus any costs incurred ($b-c$) or s , implying that they are fired. At any point in time, a principal has either passed in the previous period, failed in the previous period, or failed twice in a row and is now fired. Note that a school which fails and then passes does not have a strike against them—

only consecutive strikes causes the principal to get fired. This two-strike model may seem over-simplified, but sanctions under NCLB are actually administered after two consecutive strikes. Therefore, the best opportunity to correct behavior and avoid being punished is after getting the first strike.

If principals are to maximize lifetime payoffs, they are choosing optimal levels of cost such that they take into account the effect that costs have on passing targets in the future. In other words, they solve:

$$\max_c \sum_{t=1}^{\infty} \delta^t E(U_t(b, s, c))$$

Here, δ is a standard discount factor less than 1. For ease of computation, I am normalizing payoffs such that $b=1$ and $s=0$ where the rent from not being fired is normalized to $b-s=1$. Costs \hat{c} will now be defined as a fraction of that rent, $\hat{c} = \frac{c}{b-s}$. Note that this does not change the nature of the maximization problem since maximizing the original problem with respect to cost will be equivalent to solving for the maximum of the normalized problem with respect to \hat{c} (see Appendix A).

The associated value function for this model is:

$$V(S) = \max_{\hat{c}} 1 - \hat{c} + \delta E[V(S')] \tag{2}$$

where 1 is the payoff for not being fired and S is the state variable (number of strikes) taking on the values $\{0,1,2\}$. This simple behavioral model determines a principal's optimal choice of effort given different states and is linear in costs. The associated value function for this

model is:

$$V(0) = 1 - \hat{c}^* + \delta E[V(S') \mid S = 0] \quad (3)$$

$$V(1) = 1 - \hat{c}^{**} + \delta E[V(S') \mid S = 1] \quad (4)$$

$$V(2) = 0 \quad (5)$$

where \hat{c}^* and \hat{c}^{**} are the optimal values of cost for states 0 and 1 respectively. Notice that $E[V(S')]$ has a simple interpretation because if $S=1$, then future states S' can only be 0 or 2. Similarly, if $S=0$, then S' can only be 0 or 1.

For each school, the following year's proficiency rates, y' , are determined by the following production function:

$$y' = a_0 \hat{c}^{a_1} \cdot e^u \quad (6)$$

$$\ln(y') = \ln(a_0) + a_1 \ln(\hat{c}) + u \quad (7)$$

where a_0 is some initial score level (intercept term), y' is continuous and increasing in cost (\hat{c}) and some random noise u (see Appendix B). For easier notation, let:

(a) Probability of passing after having no strikes $P_0 = \Pr[y' > y^* \mid \hat{c}^*]$

(b) Probability of passing after having 1 strike $P_1 = \Pr[y' > y^* \mid \hat{c}^{**}]$

If the threshold for passing is y^* then the probability of passing is simply:

$$\begin{aligned}
P_0 &= \Pr[y' > y^* \mid \widehat{c}^*] \\
&= \Pr[u > \ln(y^*) - \ln(a_0) - a_1 \ln(\widehat{c}^*)] \\
&= \Pr\left[\frac{u - \mu}{\sigma} \leq \frac{1}{\sigma} \ln\left(\frac{a_0}{y^*}\right) + \frac{a_1}{\sigma} \ln(\widehat{c}^*) - \mu\right], \\
&= \Phi\left[\frac{1}{\sigma} \ln\left(\frac{a_0}{y^*}\right) + \frac{a_1}{\sigma} \ln(\widehat{c}^*) - \mu\right] \\
P_1 &= \Phi\left[\frac{1}{\sigma} \ln\left(\frac{a_0}{y^*}\right) + \frac{a_1}{\sigma} \ln(\widehat{c}^{**}) - \mu\right]
\end{aligned}$$

This implies the following:

$$E[V(S') \mid S = 0] = P_0V(0) + (1 - P_0)V(1) \quad (8)$$

$$E[V(S') \mid S = 1] = P_1V(0) + (1 - P_1)V(2) \quad (9)$$

The intuition is that a principal with no strikes would choose c^* effort and the probability of passing the next period given that effort is P_0 . The expectation of the future would be a weighted average of future potential outcomes (a first strike in the next period or no strikes again) where the weights are the probabilities of each scenario occurring. Similar logic holds for the case with one strike. Combining equations (8) and (9) with (3) and (4) yields:

$$V(0) = 1 - \widehat{c}^* + \delta [P_0V(0) + (1 - P_0)V(1)]$$

$$V(1) = 1 - \widehat{c}^{**} + \delta [P_1V(0) + (1 - P_1)V(2)]$$

$$V(2) = 0$$

This system has really has only two equations and two unknowns, so it is possible to solve.

The solutions are functions of \widehat{c}^* , \widehat{c}^{**} , P_1 and P_0 :

$$V(0) = \frac{(1 - \widehat{c}^*)}{1 - P_0\delta - P_1(1 - P_0)\delta^2} + \frac{\delta(1 - P_0)(1 - \widehat{c}^{**})}{1 - P_0\delta - P_1(1 - P_0)\delta^2} \quad (10)$$

$$V(1) = \frac{(1 - \delta P_0)(1 - \widehat{c}^*)}{1 - P_0\delta - P_1(1 - P_0)\delta^2} + \frac{\delta P_1(1 - \widehat{c}^{**})}{1 - P_0\delta - P_1(1 - P_0)\delta^2} \quad (11)$$

The first order conditions imply that for each δ and there is a an optimal pair $\{\widehat{c}^*, \widehat{c}^{**}\}$ that solves the following¹⁶:

$$0 = \frac{\partial V(0)}{\partial \widehat{c}^*} \quad \text{and} \quad 0 = \frac{\partial V(1)}{\partial \widehat{c}^{**}}$$

However, if $u \sim N(\mu, \sigma)$ then

$$\begin{aligned} \frac{\partial P_0}{\partial \widehat{c}^*} &= \frac{a_1}{\sigma} \cdot \frac{1}{\widehat{c}^*} \cdot \phi \left[\frac{1}{\sigma} \ln\left(\frac{a_0}{y^*}\right) + \frac{a_1}{\sigma} \ln(\widehat{c}^*) - \mu \right] \\ \frac{\partial P_1}{\partial \widehat{c}^{**}} &= \frac{a_1}{\sigma} \cdot \frac{1}{\widehat{c}^{**}} \cdot \phi \left[\frac{1}{\sigma} \ln\left(\frac{a_0}{y^*}\right) + \frac{a_1}{\sigma} \ln(\widehat{c}^{**}) - \mu \right] \end{aligned}$$

This simplifies the derivative a little more, and it is apparent that having values for $[a_0, a_1, y^*, \mu, \sigma]$ will allow solving for \widehat{c}^* and \widehat{c}^{**} without a closed form first order condition.

There are a few expected relationships and relative magnitudes of the parameters in this model. First, \widehat{c}^* and \widehat{c}^{**} should both range between 0 and 1. This is because costs should never be greater than the potential rent otherwise there would never be a reason to exert effort at all. The level of cost chosen after receiving a first strike should be lower than the cost chosen after passing, i.e., $\widehat{c}^* < \widehat{c}^{**}$. This implies that $V(0)$ should be greater than $V(1)$ as the expectation of the value function given no strikes is a stream of payoffs where one has

¹⁶Note that $V(2)$ is not part of the optimization problem as the principal is fired and the value function evaluated at two strikes is not a function of cost.

lower costs. Due to the complexity of the formulas, it is not obvious that these relationships are true. Plugging in some sample parameters can easily illustrate how the model works.

This numerical example uses a discount factor, $\delta = .9$, and assumes that the random component of scores is distributed $u \sim N(0, 1)$. In this case, I will set the schools exactly at the cutoff $\ln(y^*) = \ln(a_0)$ where the probability of passing reduces to $\Pr[u > a_1 \ln(c)]$. This particular functional form for the production function has a constant elasticity and exhibits properties of diminishing returns as long as $0 < a_1 < 1$ (Figure 6). As equation (10) is the value function for schools with no strikes, the derivative is taken with respect to \hat{c}^* (optimizing level of cost given no strikes) at the different levels of \hat{c}^{**} and it is shown graphically in Figure 7a. The function is well behaved and crosses the marginal cost at 1. [Similarly for maximizing equation (11) in Figure 7b] Since \hat{c}^* and \hat{c}^{**} range from 0 to 1, one can map out the solutions to (10) and (11) and the intersection of the two will solve the entire model (Figure 8). In this case where $a_1 = .5$ and $a_0 = y^*$, the solutions are $\hat{c}^* = .16$ and $\hat{c}^{**} = .31$ implying that schools with one strike will use twice as much effort or equivalently, use up 15% more rent than similar schools with no strikes near the target threshold.

As the model is set-up, decreasing δ will make both \hat{c}^* and \hat{c}^{**} smaller as future payoffs become less relevant (see Table 5). The gap between \hat{c}^* and \hat{c}^{**} will decrease as δ increases since principals who are forward looking realize that a bad draw could potentially cause them to fail and invest more as a response. If principals are myopic (low discount rate) then they will accept a lucky year and reduce efforts, disregarding its effect on future outcomes. Decreasing a_1 has the same effect, but for a different reason. The parameter a_1 captures the ability of schools to influence scores. When the elasticity is high, an increase in cost has a larger effect on scores, even at high levels of cost. When the elasticity is lower, the function

has more curvature and the effectiveness of cost on scores is greatly reduced. For certain values of a_1 , the maximization problem produces corner solutions of $\hat{c}^*=0.01$ and $\hat{c}^{**}=0.01$. These will arise when the elasticity of cost to proficiency gains is too low and even full effort would not help a school get over the threshold so no school has any incentive to try¹⁷. The other extreme is that very high elasticities allow schools to secure a future pass with minimal effort leading to cost choices of the lowest value.

To summarize, this model captures the decision making process of school administrators under a two-strike system where optimal costs (\hat{c}^* and \hat{c}^{**}) for failing and passing schools can be calculated for given values of elasticity (a_1) of the education production function and how close schools are to the cutoffs (a_0) where the discount rate (δ) and distribution of the random shocks (u) are assumed.

7 Structural Model and Empirical Estimates

Now that it has been verified that solutions to this model can exist, we can calibrate it with the RD framework. For two similar schools near the threshold, their expected proficiency rates the next period are approximated by:

$$E[\ln(y'_{pass})] = E[\ln(a_0) + a_1 \ln(\hat{c}^*)] \tag{12}$$

$$E[\ln(y'_{fail})] = E[\ln(a_0) + a_1 \ln(\hat{c}^{**})] \tag{13}$$

¹⁷The actual lowest amount of cost would be zero, however the logarithmic functional form will not allow that choice thus any optimal solution of $c=0.01$ would only be a local solution.

and since these schools are assumed to be similarly endowed, the $\ln(a_0)$ terms are equivalent in expectation. The expected difference between proficiency levels the next period is just:

$$\begin{aligned} E[\ln(y'_{fail})] - E[\ln(y'_{pass})] &= [\ln(a_0) + a_1 \ln(\widehat{c}^{**}) - \ln(a_0) - a_1 \ln(\widehat{c}^*)] \\ \widehat{\beta}_2 &= a_1 [\ln(\widehat{c}^{**}) - \ln(\widehat{c}^*)] \end{aligned} \quad (14)$$

(where $\widehat{\beta}_2$ is the regression discontinuity estimate from equation (1))

This implies that a passing and failing school near the threshold (and near each other) will have different growth rates and it will be completely attributed to the predicted difference in effort/cost and the level of elasticity. Currently, the left hand sides of the equations (12), (13) and (14) can be estimated from the data.

Recall that the structural model takes values of a_0 , a_1 to compute P_0 and P_1 and solves for optimal costs for failing and passing schools (\widehat{c}^* and \widehat{c}^{**}). Earlier, the parameters P_0 and P_1 were defined as:

$$\begin{aligned} P_0 &= \Phi \left[\frac{1}{\sigma} \ln\left(\frac{a_0}{y^*}\right) + \frac{a_1}{\sigma} \ln(\widehat{c}^*) - \mu \right] \\ P_1 &= \Phi \left[\frac{1}{\sigma} \ln\left(\frac{a_0}{y^*}\right) + \frac{a_1}{\sigma} \ln(\widehat{c}^{**}) - \mu \right] \end{aligned}$$

Notice that if all schools choose the same level of cost, then the probabilities of passing are equal. The distribution of the random shock u can be approximated by the data using the mean μ and standard deviation σ of the residuals from the regression:

$$\ln(y') = \lambda_0 + \lambda_1 y + e$$

The next step in the calibration takes the values of a_0 and a_1 and the corresponding optimal

pair \hat{c}^* and \hat{c}^{**} to solve equations (12) (13) and (14). The model is now completely identified and it is possible to estimate it using a generalized method of moments framework. This type of estimation actually resembles a grid search over potential values of (a_0, a_1) where each (a_0, a_1) produces a unique pair of optimal \hat{c}^{**} and \hat{c}^* . With candidate values of $(a_0, a_1, \hat{c}^*, \hat{c}^{**})$, the optimal set of parameters should make the left hand sides of the equations (12),(13) and (14) equal to the right hand sides of those same equations. The only remaining assumptions are choosing a discount factor δ and letting u be distributed $N(\mu, \sigma)$.

To give the NCLB the benefit of the doubt, the highest expected percentage gains in proficiency rates are around 4.8% (which is from the 95% confidence interval in the fifth order specification). After adjusting for the safe-harbor schools, the estimate is 5.6%¹⁸ As a point of comparison, proficiency targets are raised between 40-50% from 2006 to 2007. This implies that at best, the incentive effect covers around 10% of the gains needed just to keep up with the targets. To calibrate the effect of NCLB to the structural model, I also assume $\delta = .9$, and use the estimated values of $\mu = 0$ and $\sigma = .2339$. Since the grid search over the parameters involved numbers in the hundredths, there is not enough precision to isolate unique solutions. For example, it is difficult to find sets of $(a_0, a_1, \hat{c}^*, \hat{c}^{**})$ that allowed a_0 from the equations (12) and (13) to be exactly identical. Thus, I made the constraint from equation (10) be binding and then minimized the difference in the a_0 estimates (see "Difference" column in Table 6a). I have chosen the solution where $a_1 = 0.04$, $\hat{c}^* = .07$, and $\hat{c}^{**} = .28$ with $a_0 \approx 0.3$ (see Table 6a). This solution implies an RD estimate of 5.55% where the a_0 values from equations (12) and (13) are relatively close. The solution implies a

¹⁸It has been shown that the local average treatment effect can be computed by dividing the intent to treat estimate by the fraction of the treatment group that actually receive treatment. In this study, 13% of schools received safe harbor—thereby circumventing the treatment (being deemed failing) while they were technically in the treatment group (having proficiency rates lower than target). Thus the RD estimate can be "inflated" by dividing by 87% to take into account the non-compliance.

low elasticity of cost to proficiency rates, where schools near their target and fail will spend around 20% more rent than similar schools that pass. One could argue for other candidate solutions, but it is clear that elasticities over 0.1 would be ruled out as the predicted RD values increase beyond the 5.6% estimate. For the future, this grid search method can be expanded and refined to allow more precision with more decimal places especially since larger values can be ruled out—which would reduce computation time.

In reality, the fifth order specification passed the F -test for joint significance in the polynomial and interaction terms and yielded a small estimate of 1% that was statistically insignificant. I can not rule out the possibility that the true incentive effect is actually zero and there is either no difference in costs or near-zero elasticities. One adjustment to the model could potentially allow for negative RD estimates and it would involve changing the payoffs such that the payoff after one strike is lowered and the payoffs get progressively higher for consecutive passes. In this setup, principals who get their first strike essentially stop trying since incurring costs only decrease payoffs further and principals who pass are encouraged to keep their proficiency rates up. While this idea does not mirror the NCLB structure, perhaps there are non-stipulated benefits to passing (better reputation, pride, or promotional opportunities) and unmeasured costs to failure (negative stigma or low morale) that are not in the model.

It is also important to keep in mind that the use of effort or cost has not been literal. At this moment, there is no data on actual expenditures used by schools each year for improving scores and the \hat{c}^* and \hat{c}^{**} values are a relative fraction of some unobservable rent. With this simple model, it is impossible to say what would happen if the payoff b is doubled or if the magnitude of the rent is large or small in dollar terms.

8 Structural Model Extensions

Using all the possible criteria, schools that fail in one year pass the following year only 28% of the time and schools that pass in one year pass again 85% of the time. This could imply that there could be something happening to the schools that initially failed, forcing them to fail more often. One potential explanation is a signaling or “creaming” effect where after a school fails the first time, parents and their children are reluctant to return and some of the better students with more resources may leave. It may even be true that it discourages high performing new students to enroll the next year. From a longer-term perspective, the punishments from Program Improvement status seem disruptive and could potentially harm student performance. More research is needed to determine if the NCLB has accidentally set up a failure trap which prevents failing schools from being able to raise achievement.

Other extensions to the model can include more levels of strikes, more dimensions (i.e. graduation rates, test participation levels) to better approximate NCLB and all of its incentives. Further analysis could be carried out to investigate if the random variation in year to year scores is different for the pass and fail groups. That is, if there are properties about the error term that violate the assumption that schools near the threshold share a similar distribution of random shocks. In future extensions, more robustness checks can be done to verify and refine the RD estimate as well as constructing additional RD estimates using the other criteria.

Lastly, the pairing of the RD design and the structural model provides a framework for thinking about policy implications. Given that there is most likely a low elasticity of production, policy makers can change the timing of sanctions, the size of the sanctions, the targets, the rules for determining pass/fail, and other policies to try and maximize incentive effects. For example, instead of proficiency-level targets, the policy could be changed to

percentage growth targets. A future exercise can be carried out where hypothetical growth targets can be used in the model given the estimated elasticity and see how it would affect the optimal cost choices. There will presumably be more observations of schools close to the cutoffs as fewer schools would find themselves with extremely large or negative growth rates, in contrast to the current analysis, where fewer schools are right near the cutoffs since the data is in levels.

9 Conclusion

Without the structural model, a basic RD framework would not have addressed the elasticity of the education production function or the underlying cost/effort differences that would have been driving the results. The RD estimates agree with others in the literature that finds little or no effect, which does not necessarily indicate that NCLB has no incentive effects at all. The theoretical model in this research has shown that extra effort by failing schools should be expected and the largest incentives are felt only when the elasticity of cost to outcome is high. Thus, putting the results from the data and the model together, the overall conclusion is that—under the most favorable conditions—schools in the NCLB system with one strike against them will try harder and use about 2% more of the rent they receive, although there is most likely a low elasticity (less than 0.01) of increasing proficiency rates with respect to cost.

This paper is unable to address what would happen if benefits to passing were raised or punishments for failure were increased. The only recommendation is that finding ways to increase the effectiveness of cost on scores would lead to higher levels of effort. If cost and proficiency rate relationships are weak, there is no reason for schools to respond despite

harsh consequences or great rewards. However, tying performance to growth rates instead of levels will encourage more schools to put forth more effort.

While it is true that schools that just fail often try harder and attain higher growth rates, there seems to be a general pattern that schools who fail often fail again. Something else could be happening to those schools that fail—perhaps higher turnover of teachers and students—that is causing future failure. Utilizing more of the data made available from the California Department of Education (for example, staff demographics) may help answer that question.

The complexity of NCLB makes evaluation and assessment extremely difficult. High turnover in schools make year to year level comparisons inadequate measures of real change in achievement. At the current rate, NCLB goals are not attainable by the year 2014. Analyzing incentives will be critical for reform and the design of future accountability programs and understanding the underlying education production function might enable federal and state education agencies to get the results desired.

9.1 Appendix

Appendix A. The principals of a school want to maximize lifetime utility:

$$\max_c \sum_{t=1}^{\infty} \delta^t E(U_t(b, s, c))$$

Without loss of generality, I can add and subtract the term $\sum \delta^t s$ to the expectation:

$$\begin{aligned} \sum_{t=1}^{\infty} \delta^t E(U_t) &= \sum_{t=1}^{\infty} \delta^t s - \sum_{t=1}^{\infty} \delta^t s + \sum_{t=1}^{\infty} \delta^t E(U_t) \\ &= \sum_{t=1}^{\infty} \delta^t s + \sum_{t=1}^{\infty} \delta^t E(U_t - s) \end{aligned}$$

Furthermore, dividing by $b-s$ will not change the equation:

$$\frac{1}{b-s} \sum_{t=1}^{\infty} \delta^t E(U_t) = \sum_{t=1}^{\infty} \delta^t \frac{s}{b-s} + \sum_{t=1}^{\infty} \delta^t E\left(\frac{U_t - s}{b-s}\right)$$

This particular form of the original maximization problem reveals that choosing optimal values of cost will only effect the second term on the right. That is, maximizing $\sum_{t=1}^{\infty} \delta^t E(U_t)$ is equivalent to maximizing $\sum_{t=1}^{\infty} \delta^t E\left(\frac{U_t - s}{b-s}\right)$. Note that, $\frac{U_t - s}{b-s}$ will only take on the values $1 - \frac{c^*}{b-s}$, $1 - \frac{c^{**}}{b-s}$, or 0. This is because the utility in a given time period randomly takes on one of the following three forms depending on the number of strikes, and c^* and c^{**} are the

optimized values of the utility given no strikes or one strike respectively:

$$U_t^0 = b - c^*$$

$$U_t^1 = b - c^{**}$$

$$U_t^2 = s$$

The optimal values c^* and c^{**} that solve the equation $\sum_{t=1}^{\infty} \delta^t E(U_t)$ will be the same as optimizing $\sum_{t=1}^{\infty} \delta^t E(\frac{U_t - s}{b - s})$ with respect to $\hat{c}^* = \frac{c^*}{b - s}$ and $\hat{c}^{**} = \frac{c^{**}}{b - s}$. The main difference between the original maximization problem and the normalized maximization problem is that the interpretation of the payoffs for working (b) and being fired (s) is now going to be a rent, where 1 is the maximum value and costs are now a fraction of that rent.

Appendix B. Recall that costs were normalized to represent the fraction of the rent received by a principal from not being fired. In reality

$$y' = \tilde{a}_0 c^{a_1} e^u$$

$$\ln(y') = \ln(\tilde{a}_0) + a_1 \ln(c) + u.$$

However, I can add and subtract the term $[a_1 \ln(b - s)]$ to the right hand side without changing the equation:

$$\ln(y') = \ln(\tilde{a}_0) + a_1 \ln(b - s) + a_1 \ln(c) - a_1 \ln(b - s) + u$$

which can be re-written as

$$\begin{aligned}\ln(y') &= \ln(a_0) + a_1 \ln\left(\frac{c}{b-s}\right) + u \\ &= \ln(a_0) + a_1 \ln(\hat{c}) + u\end{aligned}$$

where $\ln(a_0) = \ln(\tilde{a}_0) + a_1 \ln(b-s)$. The interpretation of a_0 is the “intercept” where if costs were at the maximum level $\hat{c} = 1$, then the expected score for the next year would be exactly a_0 .

References

- Angrist, J and V. Lavy (1999) "Using Maimonides' Rule to Estimate the Effect of Class Size on Student Achievement." Quarterly Journal of Economics. May: 533-575.
- California Department of Education, *Adequate Yearly Progress Report Information Guide*. 2002, 2003, 2004, 2005, 2006. Sacramento, CA.
- Chay, K., P. McEwan & M. Urquiola. (2005) "The Central Role of Noise in Evaluating Interventions That Use Test Scores to Rank Schools." American Economic Review. September, 1237-1258.
- DiNardo, J. and D. S. Lee. (2004) "Economic Impacts of New Unionization on Private Sector Employers: 1984-2001". Quarterly Journal of Economics. 119: 1383-1442.
- Forbes, T. R. Gordon, A. Rosean, & N. Schwartz. (2006) "An Evaluation of the No Child Left Behind Act: Measuring the Effects of First-Year School Failure in California." Unpublished. April.
- Friedman, J. & S. Kelman. (2006) "Effort as Investment: Analyzing Incentives in the Public Sector". As presented in the Public Finance Seminar at UC-Berkeley, October 9.
- Hahn, J., W. van der Klaauw, and P. Todd. (2001) "Identification and Estimation of Treatment Effects with a Regression Discontinuity Design". Econometrica. 69 (1): 201-209.
- Hanushek, E. (1986) "The Economics of Schooling: Production and Efficiency in Public Schools". Journal of Economic Literature. 24: 1141-1177.
- Hanushek, E. (2003) "The Failure of Input-Based Schooling Policies". Economic Journal. Royal Economic Society, 113(485), F64-F98.
- Hanushek, E. & M. Raymond. (2004) "Does School Accountability Lead to Improved Student Performance?" NBER Working Paper 10591, National Bureau of Economic Research.
- Hanushek, E. & S. Rivkin. (1996) "Understanding the 20th Century Growth in U.S. School Spending". NBER Working Paper 5547, National Bureau of Economic Research.
- Kane, T.J., & D. Staiger. (2002) "Racial Subgroup Rules in School Accountability Systems" Paper presented at Taking Account of Accountability Conference held at Kennedy School of Government, Harvard University, June 9-10, 2002.

- Kane, T.J., & D. Staiger. (2002) "The Promise and the Pitfalls of Using Imprecise School Accountability Measures". Journal of Economic Perspectives. 16(4).
- Lee, D. S. (2006) "Randomized Experiments from Non-random Selection in U. S. House Elections". Forthcoming Journal of Econometrics.
- McCrary, J. (2005) "Manipulation of the Running Variable in the Regression Discontinuity Design" Forthcoming Journal of Econometrics.
- National Center for Education Statistics. <http://nces.ed.gov/surveys/pisa/>
- OECD's Programme for International Students Assessment. <http://www.pisa.oecd.org>
- Rivkin S., E. Hanushek & John F. Kain. (2005) "Teachers, Schools, and Academic Achievement". Econometrica. 73(2): 417-458.
- Rouse, C. & D. Figlio. (2005) "Do Accountability and Voucher Threats Improve Low Performing Schools?" NBER Working Paper 11597, National Bureau of Economic Research.
- Van der Klaauw, W. (1996) "A Regression-Discontinuity Evaluation of the Effect of Financial Aid Offers on Enrollment." Unpublished manuscript. New York University.
- U.S. House Committee on Education & the Workforce. *No Child Left Behind Funding: Pumping Gas into a Flooded Engine?* Available from: <http://edworkforce.house.gov/issues/108th/education/nclb/nclbfundingreport.pdf>, 2004.

Figure 1: Schedule of Proficiency Targets

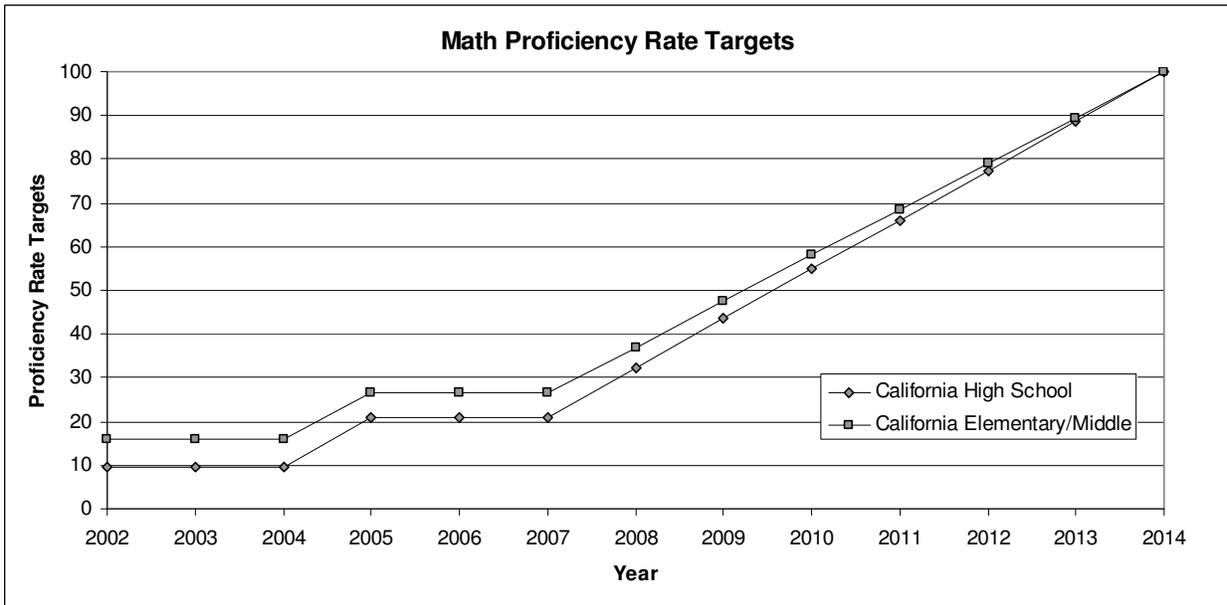
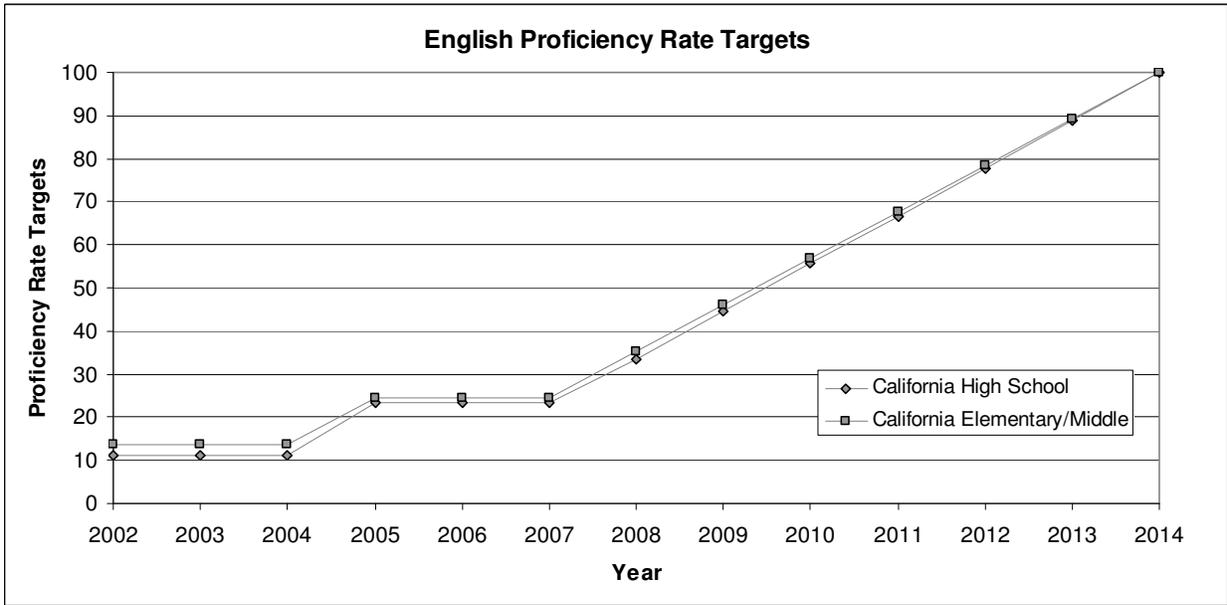


Figure 2: AYP Passing Rates and Distance to Target

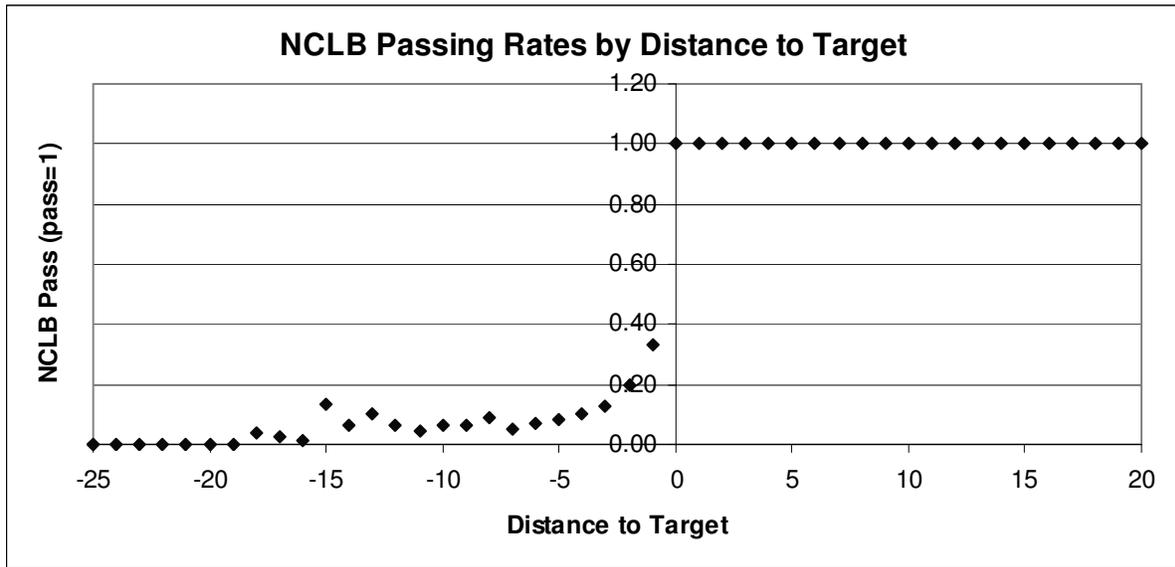
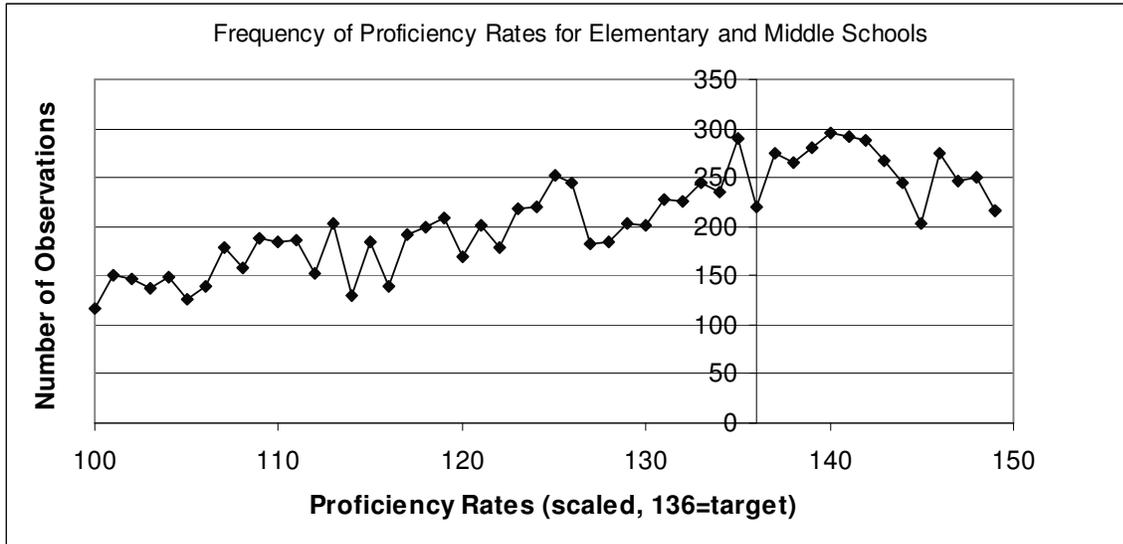


Figure 3a: Proficiency Rates for Elementary and Middle Schools (2002-2004)



Note: These are frequencies of observations of English proficiency rates by distance to target. The target is 13.6% for elementary and middle Schools before 2005 where there was an increase in the target.

Figure 3b: Histogram of English proficiency (<2005) for Elementary and Middle Schools

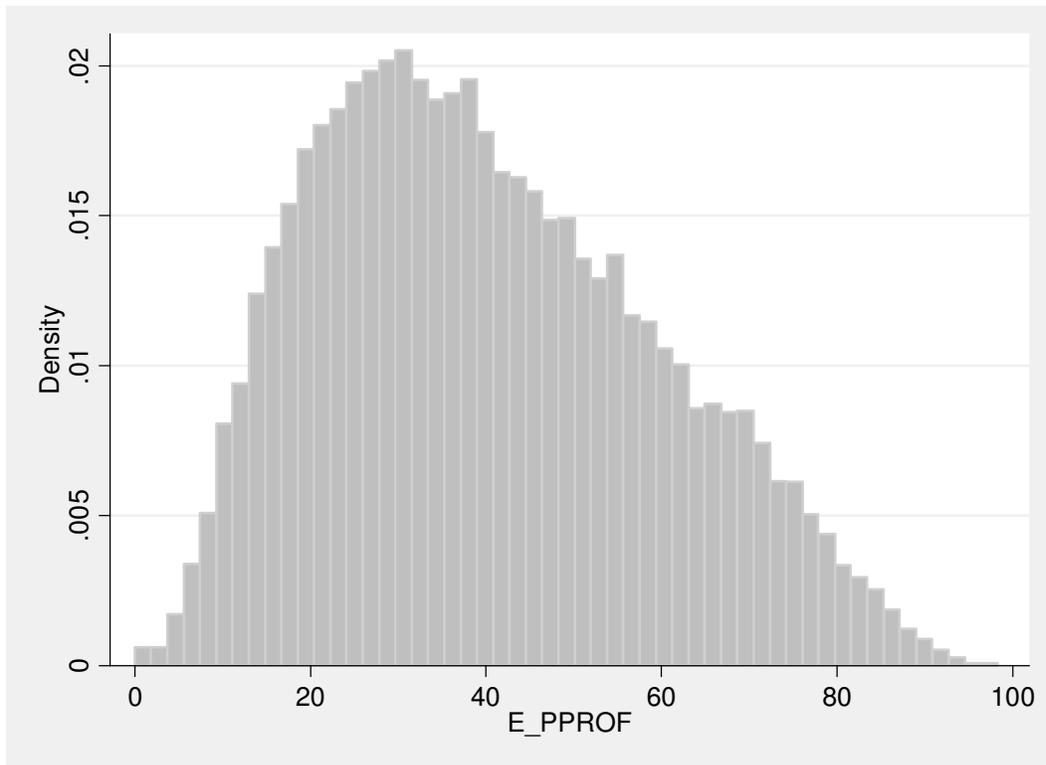


Figure 4: Demographic Composition and Distance to Target (scaled)

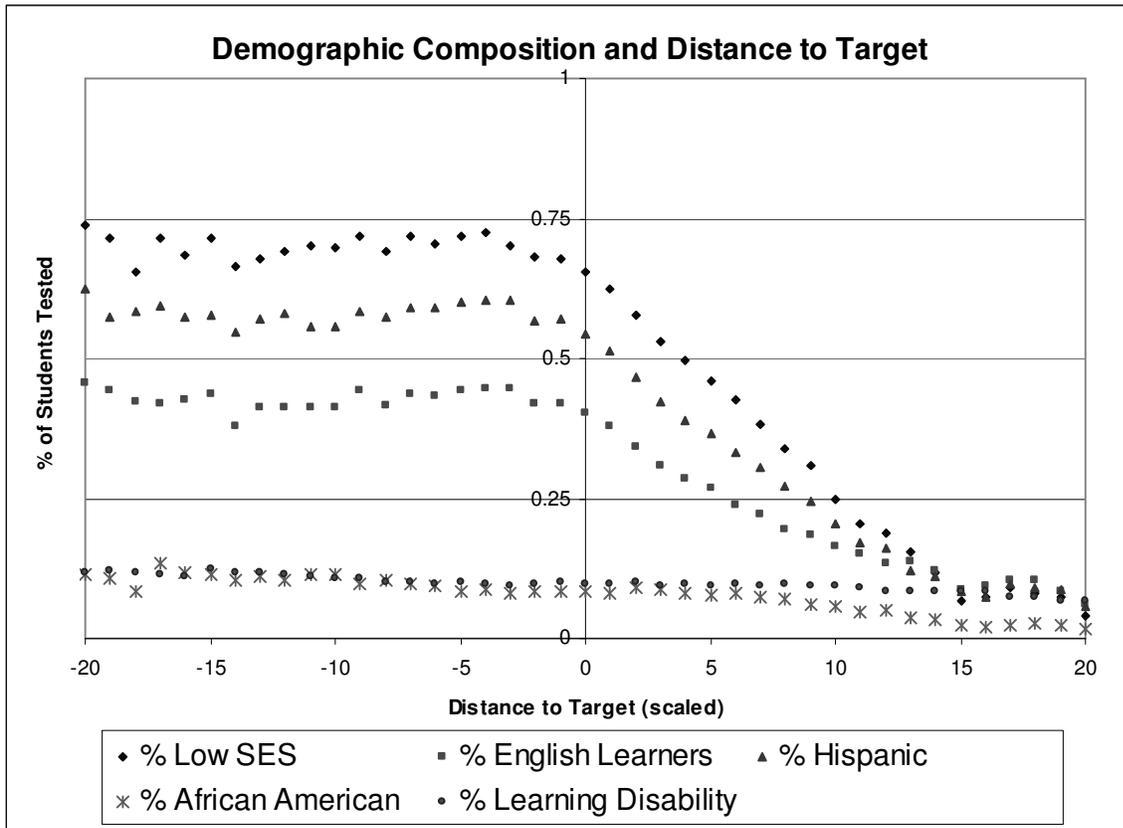
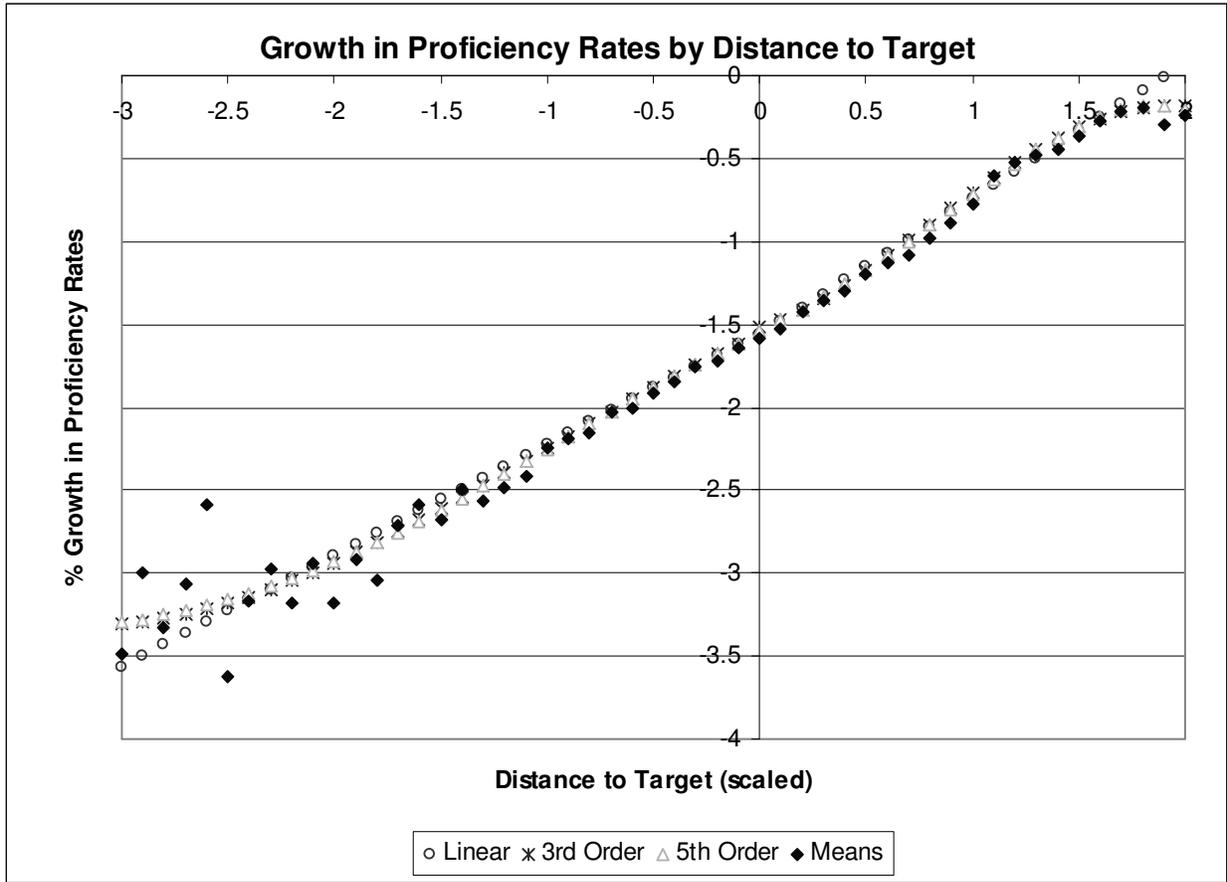
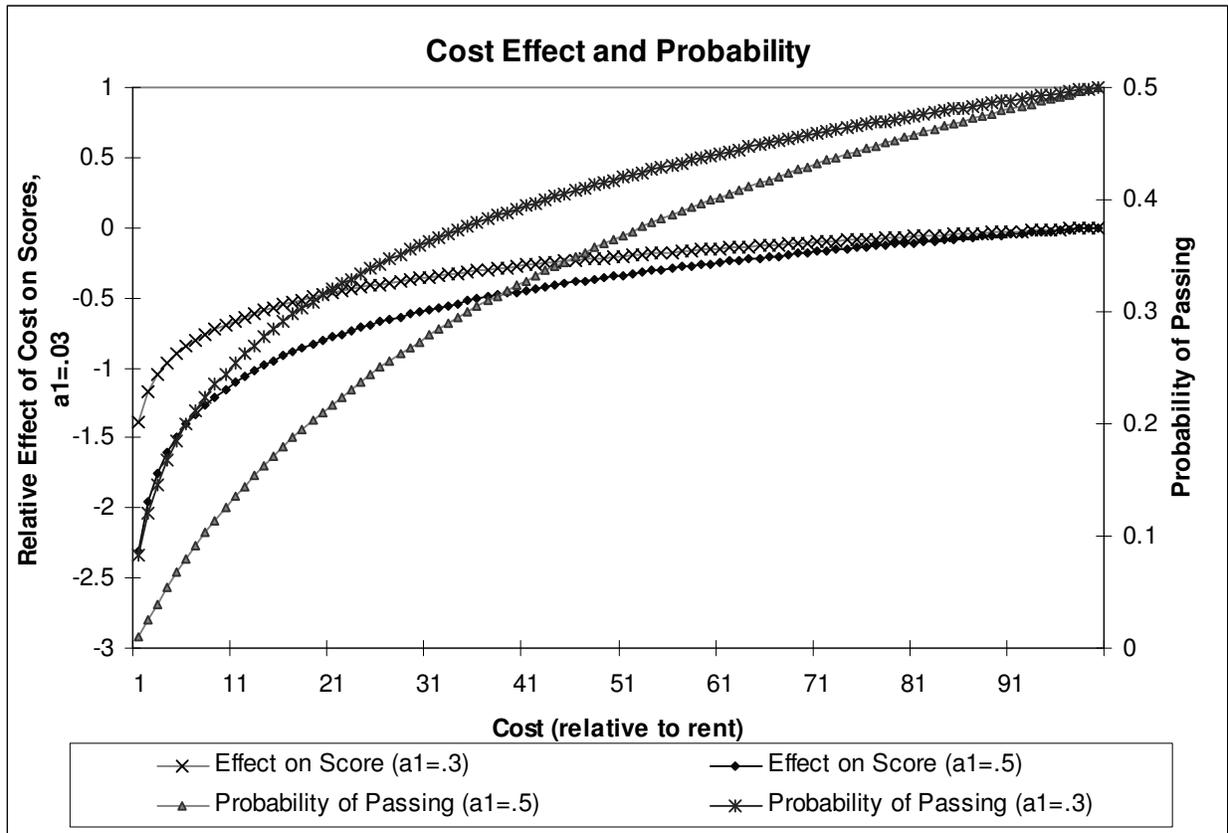


Figure 5: Average Log Proficiency Rates (next year) and Predicted Proficiency Rates



Note: This is for the sub-sample constrained to schools with more than 100 enrolled students, where percentage growth is limited to -100% and +100%.

Figure 6: Effect of Cost on Proficiency Rates and Effect of Cost on Probability of Passing



(at different values of a1)

Figure 7a: Marginal Effects of Cost on the Value Function for No Strike Schools

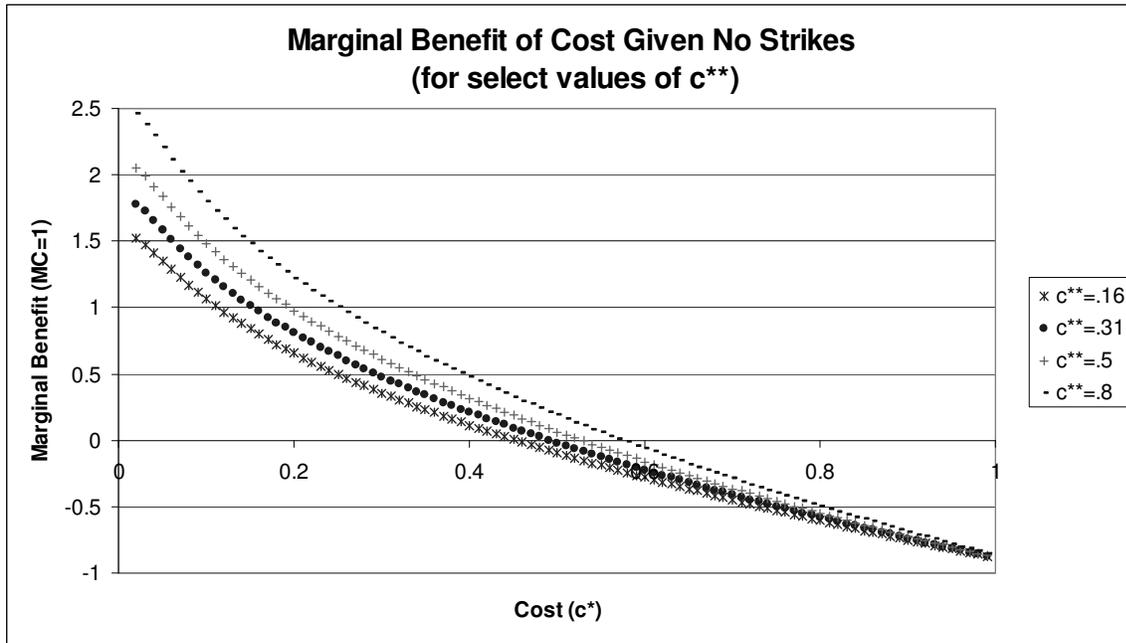


Figure 7b: Marginal Effects of Cost on the Value Function for 1 Strike Schools

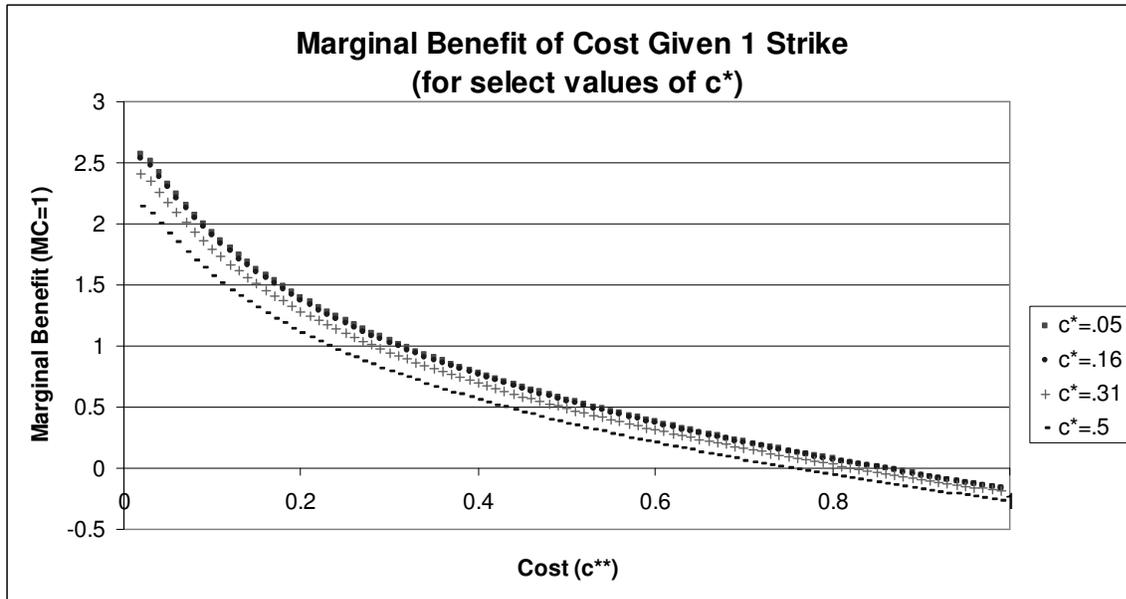


Figure 8: Joint solution to No Strike and 1 Strike Optimization Problem

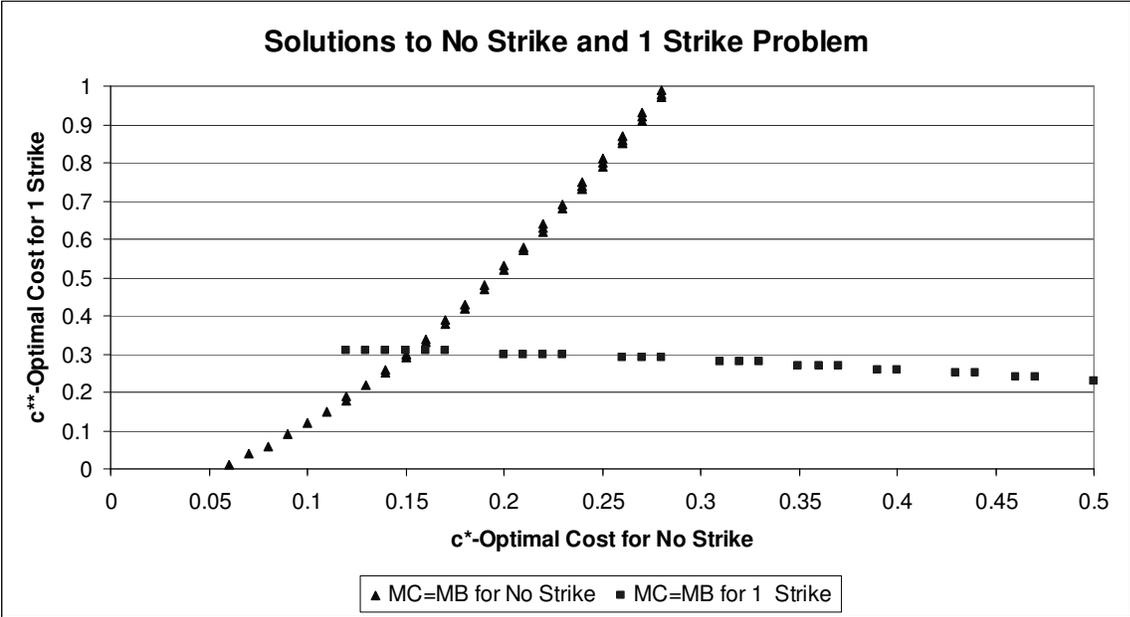


Table 1: NCLB Program Improvement Requirements Chart (for schools)

Number of Years School Does Not Make Adequate Yearly Progress (AYP)						
One	Two	Three	Four	Five	Six	Seven
		School Improvement		Corrective Action	Restructuring	
		<i>PI Year 1</i>	<i>PI Year 2</i>	<i>PI Year 3</i>	<i>PI Year 4</i>	<i>PI Year 5</i>
Did not make AYP	Did not make AYP	<p><u>Local Educational Agency (LEA):</u></p> <ul style="list-style-type: none"> • Provide technical assistance to PI school • Notify parents of PI status of school and school choice • Set aside minimum 5% for professional development to meet highly qualified staff requirements • Provide choice to attend another public school in the LEA that is not PI (LEA is responsible for transportation costs) • Establish peer review process to review revised school plan <p><u>School:</u></p> <ul style="list-style-type: none"> • Revise school plan within 3 months to cover 2-year period • Use 10% of Title I school funds for staff professional development • Implement plan promptly 	<p><u>LEA Continue:</u></p> <ul style="list-style-type: none"> • Technical assistance • Parent notification of PI status of school, school choice, supplemental services • Professional development • School choice <p><u>LEA Add:</u></p> <ul style="list-style-type: none"> • Provide supplemental educational services to all eligible students <p><u>School Continue:</u></p> <ul style="list-style-type: none"> • Plan implementation • Professional development 	<p><u>LEA Continue:</u></p> <ul style="list-style-type: none"> • Technical assistance • Parent notification of PI status of school, school choice, supplemental services • Professional development • School choice • Supplemental services <p><u>LEA Add:</u></p> <p>LEA identifies school for corrective action and does at least one of the following:</p> <ul style="list-style-type: none"> • Replaces school staff • Implements new curriculum • Decreases management authority at school level • Appoints outside expert • Extends school year or day • Restructures internal organizational structure of school <p>LEAs may give direct technical assistance to school site councils in developing school plans. LEA informs parents and public of corrective action and allows comment.</p> <p><u>School Continue:</u></p> <ul style="list-style-type: none"> • Professional development • Collaboration with district to improve student achievement 	<p><u>LEA Continue:</u></p> <ul style="list-style-type: none"> • Technical assistance • Parent notification of PI status of school, school choice, supplemental services • Professional development • School choice • Supplemental services <p><u>LEA and School Add:</u></p> <p>During Year 4, prepare plan for alternative governance of school. Select one of the following:</p> <ul style="list-style-type: none"> • Reopen school as a charter • Replace all or most staff including principal • Contract with outside entity to manage school • State takeover • Any other major restructuring <p>LEA provides notice to parents and teachers and allows comment.</p> <p><u>School Continue:</u></p> <ul style="list-style-type: none"> • Professional development • Collaboration with district to improve student achievement 	<p><u>LEA Continue:</u></p> <ul style="list-style-type: none"> • Technical assistance • Parent notification of PI status of school, school choice, supplemental services • Professional development • School choice • Supplemental services <p><u>LEA and School Add:</u></p> <ul style="list-style-type: none"> • Implement alternative governance plan developed in Year 4 <p>School continues in PI, and LEA offers choice and supplemental services until school makes AYP for two consecutive years. School exits PI after two consecutive years</p>

Table 2: Summary of Data

Summary Statistics Variable	Full Sample		Subsample*	
	Observations	Mean	Observations	Mean
Enrollment	236841	533.51	22004	511.22
		347.12		295.96
Participation Rate	236841	98.20	22012	99.16
		4.28		0.97
English Proficiency Rates	236814	40.90	22012	41.45
		19.57		19.95
English % Proficient (subgroup)	236347	34.04	21563	31.05
		21.77		22.29
Math % Proficient	236777	44.41	21563	46.13
		19.71		19.11
Math % Proficient (subgroup)	236250	38.59	21563	37.40
		22.63		21.38
% Learning Disabled	236836	0.10	20530	0.09
		0.09		0.04
% English Learner	236841	0.31	20531	0.30
		0.23		0.25
% Low SES	236841	0.52	20531	0.51
		0.30		0.32
% African American	236841	0.08	20531	0.07
		0.11		0.11
% Hispanic	236841	0.42	20531	0.42
		0.28		0.29
% White	236841	0.35	20531	0.37
		0.27		0.28
API Score	232639	724.42	21432	736.51
		101.06		94.34
Graduation Rate	21897	88.98	1190	94.86
		14.92		5.77

* Only includes schools that pass graduation, API, and participation rates but not proficiency rates.

Small schools with enrollment less than 100 students are removed.

Table 3: Estimated Discontinuities for School Characteristics

Estimated Discontinuities (3rd order polynomial)	
Enrollment	31.33
	6.16
% Hispanic	0.0012
	0.0045
% Low SES	-0.0057
	0.0045
% African American	0.0011
	0.0021
% White	0.0083
	-0.0083
% Learning Disability	0.0016
	0.0010
% English Learners	-0.0010
	0.0040

Table 4: Regression Discontinuity Results

Effect of Failing on Next-Year Proficiency Rates

(current year)	Next-Year (Log) Proficiency Rates						
	Linear	2nd Order	3rd Order	4th Order	5th Order	5th Order	
Failing	0.0802 ** (.008)	0.0740 ** (.011)	-0.0386 ** (.013)	-0.0356 ** (.015)	0.0118 (.017)	0.0142 (.017)	
Distance to Target	0.5277 ** (.006)	0.5057 ** (.017)	0.0352 (.035)	0.1966 ** (.067)	0.8445 ** (.118)	0.8718 ** (.118)	
Distance^2		0.0126 (.008)	0.6738 ** (.040)	0.2854 ** (.134)	-2.0923 ** (.357)	-2.1669 ** (.358)	
Distance^3			-0.2419 ** (.013)	0.0751 (.098)	3.3559 ** (.443)	3.4576 ** (.444)	
Distance^4				-0.0821 ** (.024)	-1.9686 ** (.237)	-2.0244 ** (.238)	
Distance^5					0.3815 ** (.046)	0.3921 ** (.046)	
%White	-0.0951 (.091)	-0.0933 (.091)	-0.0871 (.092)	-0.0890 (.092)	-0.0861 (.093)		
%Black	0.0021 (.094)	0.0038 (.094)	0.0246 (.025)	0.0144 (.095)	0.0166 (.096)		
%Hispanic	-0.1464 (.091)	-0.1452 (.091)	-0.1312 (.092)	-0.1320 (.092)	-0.1293 (.092)		
%English Learner	0.0256 (.023)	0.0250 (.023)	0.0081 (.023)	0.0066 (.023)	0.0065 (.023)		
%Disability	-0.3587 ** (.104)	-0.3575 ** (.104)	-0.3567 ** (.104)	-0.3568 ** (.104)	-0.3535 ** (.103)	-0.3479 ** (.102)	
% Low SES	-0.0370 (.020)	-0.0364 ** (.020)	-0.0109 (.020)	-0.0101 (.020)	-0.0085 (.020)		
Constant	-0.7168 ** (.090)	-0.7114 ** (.090)	-0.6573 ** (.092)	-0.6875 ** (.092)	-0.7149 ** (.092)	-0.7973 ** (.022)	
Subgroup Controls	yes	yes	yes	yes	yes	no	
Interaction Terms	yes	yes	yes	yes	yes	yes	
R2	0.7858	0.7858	0.7882	0.7883	0.7887	0.7878	
Obs	22634	22634	22634	22634	22634	22634	

*All coefficients are statistically significant at the 5% level are marked **, at the 10% level marked *.
 School type, proficiency levels, and enrollment are also controlled, results not reported.
 Other subgroup coefficients are available, only a few a reported here for illustration.
 Only %disability was included for last specification. Reported standard errors are robust standard errors.

Table 5: Changing Parameters and Optimized c^* , c^{**}

Changing Elasticity $d=.9$			Changing Discount Rates $a_1=.5$		
a_1	c^*	c^{**}	Discount	c^*	c^{**}
0	0.01	0.01	$d=.95$	0.22	0.36
0.1	0.06	0.11	$d=.9$	0.16	0.31
0.2	0.11	0.19	$d=.85$	0.11	0.27
0.3	0.14	0.24	$d=.8$	0.07	0.23
0.4	0.16	0.28	$d=.75$	0.03	0.19
0.5	0.16	0.31	$d=.7$	0.01	0.16
0.6	0.14	0.34			

Lag y =proficiency rate, y^* =target, cost= c
 $y=a_0+c^{a_1}$, $\ln(a_0)=\ln(y^*)$

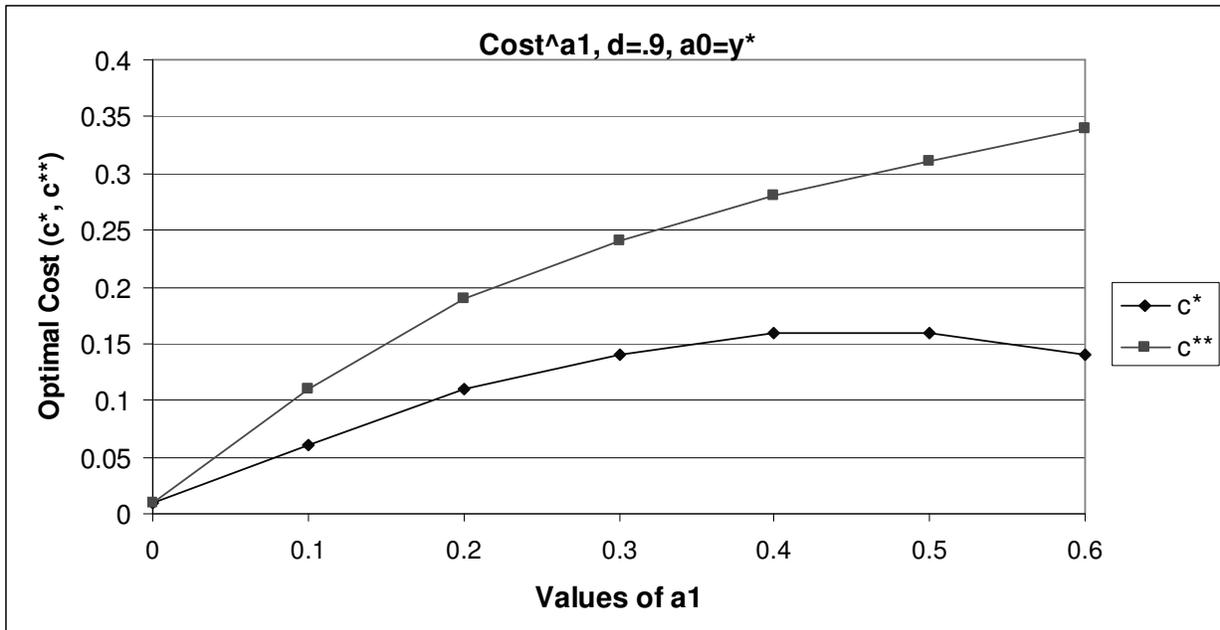


Table 6a: Grid Search Results for the Structural/Empirical Model Solution

Candidate values of (a_1 , a_0 , c^* , c^{}) and the corresponding predicted RD estimate.**

a_1	(Eq 10)	(Eq 11)	Difference	c^*	c^{**}	Predicted Values from Structural Model Discount Rate=.9		
	a_0	a_0				P0	P1	Imputed RD
0.06	0.314	0.300	-0.013	0.16	0.37	0.62	0.70	0.0503
0.03	0.312	0.298	-0.014	0.03	0.18	0.84	0.89	0.0538
0.04	0.312	0.298	-0.015	0.07	0.28	0.75	0.82	0.0555
0.06	0.347	0.332	-0.015	0.03	0.07	0.06	0.08	0.0508
0.10	0.323	0.308	-0.015	0.25	0.43	0.43	0.52	0.0542
0.06	0.336	0.321	-0.015	0.05	0.12	0.11	0.16	0.0525
0.03	0.316	0.300	-0.016	0.02	0.14	0.89	0.93	0.0584
0.04	0.314	0.299	-0.016	0.06	0.26	0.79	0.85	0.0587
0.09	0.333	0.317	-0.016	0.15	0.28	0.23	0.31	0.0562
0.07	0.317	0.301	-0.016	0.18	0.42	0.60	0.70	0.0593
0.05	0.315	0.299	-0.016	0.1	0.33	0.74	0.81	0.0597
0.10	0.337	0.320	-0.017	0.16	0.29	0.22	0.30	0.0595
0.08	0.341	0.323	-0.017	0.09	0.19	0.15	0.22	0.0598

Table 6b

Changes to solution by varying discount rate.

a_1	a_0	c^*	c^{**}	P0	P1	RD	Discount Rate
0.04	0.31	0.11	0.43	0.77	0.84	0.0545	0.95
0.04	0.31	0.07	0.28	0.75	0.82	0.0555	0.9
0.04	0.31	0.05	0.21	0.73	0.80	0.0574	0.85