California Class Size Reduction Reform: New Findings from the NAEP*

Fatih Unlu†
Princeton University
November 2005

Abstract: In 1996, California enacted one of the most extensive and expensive educational reform initiatives ever: the Class Size Reduction (CSR) Program. The CSR program provided substantial extra funds to schools that limited class size to 20 or fewer students in their K-3 classes. A number of studies evaluated the CSR program, but reported mixed results, in part because they lacked pre-program achievement data and used questionable comparison groups. This paper extends the literature by using student-level achievement data from the National Assessment of Educational Progress (NAEP) State samples, which contains comparable test scores prior to the program and afterwards for California and other states. The following empirical strategies are employed: First, I compare test scores of California 4th graders prior to and following the program's implementation. Then, in a difference-in-differences framework, I compare test scores of California 4th graders with test scores of 8th graders, who were not affected by the program using pre and post-program data. Finally, I match California 4th graders with 4th graders from other states by employing propensity score matching to compare their test score changes in a conditional difference-in-differences framework. The results are consistent with the view that the CSR program has had a positive and significant influence on California students' achievement scores. In particular, most specifications suggest that between 1996 and 2000, California 4th graders' NAEP test scores in Mathematics increased by between 0.2 and 0.3 of a standard deviation compared to the increase for closely matched students who were not exposed to the CSR initiative.

* I am indebted to Alan Krueger, for his continuous support and help throughout this project. I am also grateful to Cecilia Rouse and Jesse Rothstein for their helpful suggestions and guidance. I would like to thank participants at the Labor Lunch Seminar at the Industrial Relations Section, in particular to Giovanni Mastrobuoni, Orley Ashenfelter and Jeffrey Kling for their comments and suggestions. I also would like to thank Wayne Dughi from California Department Education, Brian Stecher and Delia Bugliari from RAND for their assistance with data issues. I also thank Leonie Haimson for her editorial comments. The views expressed in this paper are those of the author and any errors are also entirely my own.
† 001 Fisher Hall, Department of Economics, Princeton University, Princeton, NJ 08544. E-mail: funlu@princeton.edu. Revisions are available on www.princeton.edu/~funlu.

1. Introduction

In 1996, California enacted an ambitious and expensive educational policy to improve student achievement, the California Class Size Reduction (CSR) program. The CSR reform promised extra state funds to schools that enrolled 20 or fewer students in their K-3 classes. Although the reform program was voluntary, participation was extensive. Since the program was introduced, almost ten million K-3 students in California have received education in smaller classes.

The effectiveness of class size reduction as a policy instrument is controversial. In his surveys of the research literature, Hanushek (1981, 1986, 1996, 2002 and 2003) has claimed that there is no systematic positive association between smaller classes and better academic performance. Instead of reducing class sizes, he maintains that other reform alternatives are less costly and more effective. Contrary to those arguments, Krueger (2003) has challenged Hanushek's methodology and reanalyzed the same studies that Hanushek used in his meta-level analyses. Krueger found that reductions in class sizes are systematically linked with improved performance.

A number of studies have attempted to determine whether California's CSR program was successful in meeting its goal of better academic achievement. However, apparently inconsistent results reported in these studies have prevented firm conclusions, in part due to the lack of baseline data as there was no state-wide testing program in California prior to the introduction of CSR. One of these studies, Jepsen and Rivkin (2002) highlights this fact: "A number of factors hinder an analysis of the overall effect of CSR. Because California did not administer statewide

examinations until the1997-1998 school year, no baseline measure of achievement prior to CSR is available."

This paper extends the literature by using National Assessment of Educational Progress (NAEP) State Samples to solve the data problems encountered by previous studies. This rich dataset allows one to control for pre-program achievement levels as California students have been assessed by the NAEP since 1990. This paper also improves on the methodologies of previous analyses by employing statistical methods that rely on more plausible assumptions. I employ the following strategies: First, I compare test scores of California 4th graders prior to and following the program's implementation. Then, in a difference-in-differences framework, I compare test scores of California 4th graders to the test scores of California 8th graders who were unaffected by the program, using pre and post-program data. Finally, I match California 4th graders with 4th graders from other states using propensity score matching to compare test scores in a conditional difference-in-differences framework.

The results of these analyses suggest that the CSR program has had a positive and significant influence on student achievement in California. In particular, results show that between 1996 and 2000, California 4th graders' NAEP test scores in Mathematics increased by between 0.2 and 0.3 of a standard deviation. Although the evidence is less clear as to whether there were differential effects by race, ethnicity and free lunch status; black students seem to have benefited from the CSR program more than any other racial or ethnic group. Finally, these results are robust in many of the analyses performed.

The organization of this paper is as follows: Section 2 examines prior class size reduction programs introduced by other states and summarizes their results. It describes California's CSR program in detail and provides a review of previous evaluations. Section 3 introduces and

describes the state NAEP data-set. Section 4 outlines the empirical strategies employed in this paper and reports corresponding estimation results. Section 5 summarizes the findings of this study. Finally, Section 6 presents conclusions and proposes additional topics for further investigation.

2. Background

2.1 Class Size vs. Academic Achievement in the literature

Numerous studies have analyzed the effect of class size on academic achievement. Their findings have generated heated debate in the economics literature. A comprehensive review of the extant literature is beyond the scope of this paper. Hence, I limit my discussion to some of the research related directly to the effects of state enacted class size reduction programs on academic achievement.

2.2 Class Size Reduction Programs in the US

Though when enacted, the California CSR program was one of the most ambitious state CSR programs, California was not the first state to pursue such a policy. Tennessee, Texas, Wisconsin and Nevada are among many states that have enacted CSR policies.

Tennessee's Student/Teacher Achievement Ratio (STAR) study is the most influential, as it was the first and only class size reduction program to employ an experimental design. Tennessee's STAR experiment randomly assigned the students who were entering kindergarten in 79 participating schools to one of the following types of classes: a small class (13-17 students), a regular class (22-25 students) and a regular class with a full-time teacher's aide. Teachers were also randomly assigned to one of the class types. Studies analyzing the effects of the program have found that students who were assigned to smaller classes performed better in

standardized tests during the program (Word et al. (1990), Finn and Achilles (1990)). Krueger (1999) re-analyzed the STAR data, considering whether the experiment deviated from random-assignment and concluded that students of smaller classes enjoyed performance gains, especially in first year they joined the program.

A long term assessment of the students who participated in the STAR experiment was also conducted. Pate-Bain et al. (1997) reported that students who were in smaller classes during the experiment and returned to regular classes in the 4th grade continued to outperform their peers from regular classes through the eighth grade. Krueger and Whitmore (2001) observed the same pattern and they also found that the students who were assigned to smaller classes in the early grades were more likely to take the ACT or SAT exams in the senior year of high school. More recently, Finn et al. (2005) found that students who spent more than three years in smaller class experienced an increase in their high school graduation rates.

Wisconsin's SAGE (Student Achievement Guarantee in Education) initiative was the next noteworthy implementation of K-3 class size reduction. The program began in 1996 and implemented smaller Kindergarten and 1st grade classes (15 at most) in 45 of the low-income schools in the state. Smith et al. (2003) evaluated the program and found that the SAGE students experienced performance gains when compared with students from similar comparison schools.

Nevada enacted a class size reduction program in 1989, which reduced K-3 class sizes of selected schools to 16. Most of the studies evaluating Nevada's program (Snow (1993), Peterson and Rehault (1995), Sturm (1997)) found that the reductions in class sizes had little effect on student achievement.

2.3 California Class Size Reduction Reform

In 1996, then-Governor Pete Wilson proposed channeling the budget surplus into an initiative aimed at reducing class size in the early grades. At the same time, average class size in the state's elementary schools was almost 30, the largest in the nation. In July 1996, California lawmakers, inspired by the Tennessee STAR experiment, authorized the class size reduction reform as a remedy to these problems.

Unlike some other CSR initiatives, the California program is voluntary. The legislation provides additional funding to the districts that participate, with the amount of extra state funding determined by the number of K-3 students placed in classes of 20 or fewer.[3] The additional funding is substantial: In the first year of the program (1996-1997), school districts received $650 for every student in smaller classes. In 2004-2005, the supplementary funding was $928.[4] The California Department of Education reported that the overall cost of the program was $971 million in 1996-1997 and $1.6 billion in 2003-2004.

Although constrained by limited classroom space and increasing enrollment, most California schools were lured by the sizeable opportunities promised by the CSR program. Bohrnstedt and Stecher (2002) reported that some 18,000 new classrooms were created; libraries, computer clusters, labs and auditoriums were also converted into classrooms, especially in the first year. Even though the CSR bill was signed only 6 weeks before the beginning of the 1996-1997 school year, as Table I shows, 88% of California first graders were placed in smaller classes in the first year of the reform. Table I also demonstrates that the implementation of the

[3] The law also specifies an order for participation; schools initially have to reduce first grade class sizes, then second grade class sizes and then they may choose to reduce kindergarten and/or third grade classes.

[4] Per-pupil state expenditure in California before the program was about $6,000. In addition to the per-pupil funds, CSR schools received a one-time facilities grant of $25,000 in 1996-1997 and $40,000 in 1997-1998 and 1998-1999 for each new classroom they created.

CSR reform was almost complete by the 1997-1998 school year for the first and second grades, and by 2000, at least 90 % of all K-3 students were in smaller classes throughout the state. As is also apparent from Table I, CSR participation has recently started to decline somewhat, because for some schools, the cost of keeping class sizes below 20 has exceeded the extra funding provided by the state. Smaller classes, however, continue to be extremely popular among students, parents, and teachers.

2.4 Prior Research on the effects of the CSR Reform

To what extent has the CSR reform fulfilled its goal of increasing academic achievement of K-3 students in California? The California Department of Education assembled a number of major research organizations into a consortium to examine this question.[5] The Consortium conducted an extensive evaluation of the CSR program and published their findings in four reports (Bohrnstedt and Stecher (1999, 2002), Stecher and Bohrnstedt (2000, 2002)).

As there were no state-wide standardized tests given in California to students before the program began, the Consortium used the variation in the CSR participation rates of schools in an attempt to identify the program's effects with just post-implementation data.[6] Bohrnstedt and Stecher (1999) and Stecher and Bohrnstedt (2000), for example, compared test scores of third graders in those schools that implemented smaller classes with the scores of third graders who were in schools that had not implemented the program. In an effort to control for any differences between CSR adopter and non-adopter schools, Stecher and Bohrnstedt (2000) subtracted the difference in fifth graders' test scores between both sets of schools, assuming that fifth-graders

[5] The CSR Consortium was made up by American Institutes for Research (AIR), RAND, WestEd, Policy Analysis for California Education and Edsource. The consortium conducted research about various aspects of the CSR reform from May 1998 until June 2002. They not only investigated how the reform initiative affected students' achievement, but also analyzed the implementation of the program and its effects on the state-wide distribution of qualified teachers, distribution of resources, parental involvement and the way teachers teach. For more information about the CSR consortium and their findings, please visit www.classize.com.

[6] Only after program's initiation, in 1998, was the SAT-9 test first administered in California.

were unaffected by the program. As a result, their analysis documented a positive association between being in smaller classes in the third grade and third grade academic achievement, but the effect was small.[7] In a subsequent study, Bohrnstedt and Stecher (2002) conducted an additional analysis, comparing students who were in smaller classes in the first, second and third grades with those who were in smaller classes only in the second and third grades. The results failed to show a significant effect of spending one more year in smaller classes.

Two widely cited concerns about the CSR program were that schools might not be able to hire qualified teachers for new classes and that the initiative might induce experienced teachers to migrate from schools with large numbers of low-income students to work in wealthier districts. The CSR consortium found that some of the newly appointed teachers were without full credentials and inexperienced, but they did not find evidence of a teacher mobility effect.[8] They also reported that teachers of reduced size classes provided more individual attention to students than did teachers of larger classes. Instructional methods and curriculum, however, did not differ between CSR adopters and non-adopters.

There are two other major studies that investigate the impact of the CSR program. Jepsen and Rivkin (2002) also used the variation in class sizes created by the uneven implementation rate of the program to identify the achievement effects of the reform. They found that students in smaller classes performed better than students in larger classes, with more substantial effects among lower-income and minority students. They also emphasized that qualifications of California elementary school teachers declined as a result of the CSR reform, and argued that

[7] For Reading, Mathematics, Language and Spelling, Stecher and Stecher and Bohrnstedt (2000) report that the "adjusted" effect of 3rd grade CSR participation are: 0.05, 0.1, 0.1 and .04 (in standard deviation units) but they are all statistically significant at the 5% level. They also noted that, in the following school year, fourth graders who had been in smaller classes in the previous year showed better performance than the fourth graders who were in regular sized classes in the third grade.

[8] Bohrnstedt and Stecher (2002) also noted that emergency certified teachers did just as well as the certified ones.

this decline was more alarming in lower income and minority schools, where it partially, and in some cases fully, offset benefits of smaller classes.

Another investigation of the achievement effects of the CSR program was conducted by Sims (2003). Sims argued that the size-20 threshold introduced by the CSR program created an incentive for schools to become eligible to receive funds by assigning students from different grades into combination classes. Sims performed a school level analysis[9] and found that where combination classes were implemented, they significantly decreased achievement levels of second and third graders counteracting the potential benefits of smaller classes.

The analyses undertaken by the CSR Consortium, Jepsen and Rivkin (2002) and Sims (2003) relied on comparing student achievement in schools that implemented CSR to those that did not. This approach, however, may lead to biased results, because these two sets of schools may have unobserved characteristics that affected student achievement in other ways. As Stecher and Bohrnstedt (2000) acknowledged, implementation of the program was slower in inner-city schools, and those with larger numbers of minority and low-income students.

In order to investigate this issue further, I examine the characteristics of CSR participant and non-participant schools whose students were used in these studies. Stecher and Bohrnstedt (2000) used 3rd graders from the 1998-1999 school year in their sample while Jepsen and Rivkin used 3rd graders from the 1997-1998 and 1999-2000 school years and Sims used 2nd graders from 1997 through 2000 and 3rd graders from the 1998-1999 and the 1999-2000 school years. Table II tabulates characteristics of schools of these students using the CSR Consortium's Classification

[9] In this analysis, Sims creates instruments for the variables *average class size* and *percentage of students in combination classes* that he use in his regression models by using the fact that schools are expected to use combination classes more when their enrollments in grades K-3 are far from being a natural multiple of twenty.

of CSR participant and non-participant schools using data collected from the Common Core of Data (CCD) and the Demographics Office of the California Department of Education.[10]

Table II shows that in some years, nearly every school implemented the program. It is apparent from Table II that CSR adopter and non-adopters were systematically different in many respects, including racial composition and location. In some years, teachers of CSR adopter schools are more experienced and more highly educated than are teachers of non-adopters. These figures altogether suggest that there were very few schools that did not implement CSR in some of the samples used by previous studies and it is highly likely that CSR adopter and non-adopter schools have observable and unobservable differences. Hence, comparing the achievement levels of these two sets schools (i.e. using the variation in the CSR participation rates) to evaluate the effects of the CSR program may not be the best strategy.

The CSR Consortium's analysis used 5th graders' test scores in an attempt to account for the differences between the CSR participant and non-participant schools as this approach implicitly assumes that 5th graders were completely unaffected by the CSR program. However, this assumption doesn't hold because the students used by Stecher and Bohrnstedt (2000) were in the 5th grade in the 1998-1999 school year and they were in the third grade in 1996-1997, when CSR was first adopted. Table I indicates that 18% of the third graders were actually in smaller classes in 1996-1997. If the CSR program had positive effects, adjusting the test score

[10] The CSR Consortium kindly shared this data with me (I specifically thank Brian Stecher and Delia Bugliari from RAND for all their help and interest.) Nevertheless, the data that I received includes participation classifications for only 2389 California schools, since the Consortium couldn't determine the participation indicators more than half of California schools. That is one of the reasons why there are only 2350 schools in Table II. CSR participation information was missing for the schools of 1997-1998 third graders, who were used in Bohrnstedt & Stecher (1999) and Jepsen & Rivkin t (2002) so these schools are excluded from Table II.

differences of the third grade CSR participants and non-participants by the test scores of the 5th graders likely attenuates the estimated CSR effect.[11]

By utilizing the NAEP State Samples and more appropriate difference-in-differences strategies, this paper aims to overcome some of the data difficulties and methodological limitations of earlier evaluations of California CSR. The next section describes how the dataset can be employed in this manner.

3. Data Sources

3.1 State NAEP Assessments

The primary data sources used in this study are the 1996 and 2000 assessments of the State NAEP in Mathematics.[12] I chose the 1996 and 2000 NAEP State Assessments in Mathematics for a couple of reasons. First, almost all previous evaluations utilized students from this period. Second, assessments in other subjects in this period were not as suitable as the math assessment to investigate the effects of the CSR program. In particular, Reading State NAEP was only assessed in 1998 in the first six years of the program and this sample includes very few students exposed to CSR. I did not use the Science NAEP because it was not given to fourth graders in 1996.

California participated in the State NAEP Assessment in Mathematics in the 4th and 8th grades in Spring 1996 and Spring 2000. Hence, test scores of the California 4th graders from 2000 can be used to measure the achievement levels of the students exposed to the CSR

[11] Note that schools have to implement the CSR in the 1st and 2nd grade before implementing it in the 3rd grade and kindergarten classes. Hence, 5th graders from the 3rd grade CSR participating schools in the 1998-1999 school year are more likely to have been exposed to the program in the 1996-1997 school year when they were third graders.
[12] Initiated in 1969, NAEP was designed as an annual national survey to measure and follow academic achievements of American students of ages 9, 13 and 17. State NAEP assessments were introduced to have representative samples of each state and have been carried out on a regular basis since 1996.

program. Fourth graders' test scores from 1996 can be employed to assess the pre-program achievement levels of students, as this cohort was never exposed to the program. In addition, NAEP test scores for the 8th graders in California and 4th graders from other states can be utilized in a Difference-in-Differences (DID) framework to examine the possible effects of the CSR program.

Another advantage of the NAEP is that it is a rich micro-level dataset. A particular State NAEP dataset not only contains information about how each student has performed on the test, but also includes responses to detailed questionnaires given to students, teachers and school administrators.[13] Therefore, NAEP samples facilitate micro-level (student-level) empirical analyses, which have many benefits over school-level analyses. Note that all the previously described evaluations of the CSR reform were conducted at the school-level.[14]

Yet, using State NAEP to evaluate the CSR reform has a few notable drawbacks as well. First, it is not possible to follow individual students' achievement levels over time since the actual students sampled by State NAEP in 1996 and 2000 were different. Nevertheless, it is still possible to compare the results of student groups and schools over time, by selecting those with similar background characteristics. Second, the 4th grade California students who were exposed to the CSR program in earlier grades were no longer in smaller classes when they were given the NAEP in the spring of 2000 in the fourth grade, but had been placed in larger classes for about six months.[15]

[13] See O'Reilly et al (1999) for a complete description of the contents of NAEP student, teacher and school surveys.
[14] This was partly due to the fact that California Department of Education does not give access to the student-level data referring to confidentiality issues. Also, although the CSR Consortium had access to the student-level data, they preferred conducting school level analyses in most cases because the student data cannot be linked over the years.
[15] If the CSR program had positive effects, but those effects declined somewhat over six months, the estimates presented in this paper could be regarded as lower-bound estimates of the true effects of the program.

There are a few other points requiring special attention when working the NAEP data. First of all, the NAEP State sample is a stratified sample and not every student has the same probability of being selected. Therefore, when analyzing the student sample, sampling weights should be used to account for the different probabilities of selection.

Another feature of the NAEP dataset is its use of a "Balanced Incomplete Block Spiral Method," meaning that students are administered hour-long portions of the entire test, in order to increase their motivation to answer the questions asked. One disadvantage of this method is that the content of the test that each particular student is given is limited, and her performance is not sufficient to reveal her true proficiency in the subject. To tackle this, following the methods proposed by Rubin (1987) and Mislevy (1991), the NAEP data set provides a set of estimates for each student's proficiency levels called "plausible values."[16] Plausible values are used as measures of academic achievement in this analysis, as in many others.

3.2 Other Data Sources:

The NAEP data set contains class size information only for the year in which it was administered. For the other years of interest, the California Department of Education provided me with data containing the average class sizes of those schools sampled by the NAEP in 1996 and 2000. I also utilized the Common Core of Data to obtain more information about these schools, for the propensity score matching procedures.[17]

[16] Plausible values are estimated using a student's answers to the test questions and his/her survey responses. A posterior distribution for each student's true achievement is computed and for each student 5 plausible values are drawn from this posterior distribution. 5 set of plausible values are highly correlated with each other.

[17] The Common Core of Data (CCD) is maintained by the National Center for Education Statistics. The information I received from CCD includes pre-treatment characteristics of 2000 NAEP schools such as racial composition, percentage of students who were eligible for reduced price lunch and total enrollment and pupil-to-teacher ratio. For some schools in the NAEP, school data was missing. I used the CCD to determine characteristics of such schools.

Table III reports the descriptive statistics of the NAEP samples of California 4th graders, 4th graders from all other states, and California 8th graders. Table III also tabulates school and teacher characteristics of the corresponding samples. Note that although characteristics of 1996 and 2000 California 4th graders are quite similar, there are major differences between socio-economic attributes of California 4th graders and 4th graders of other states. Similarly, elementary schools in California and other states also differ substantially (see enrollment, locations of schools and teacher-to-pupil ratio). Effects of the CSR Reform on the teacher population can also be observed by the changes in the characteristics of teachers of 4th grade California students. Note the increase in the proportion of 4th grade California teachers who have less than 10 years of teaching experience (from 40.3 % to 56.5 %) and the proportion of the teachers without any certification (from 9.1% to 21.2%) between 1996 and 2000. 8th graders, on the other hand, doesn't seem to be affected by the CSR program.

4. Empirical Models and Estimation Results:

Consider the following framework to estimate the average effect of the program on the 'treated' students: Let Y_{1i} denote the achievement level of a student i if she participated in the CSR program and Y_{0i} be her achievement level without CSR. In addition, let T_i be 1 if student i is a CSR participant and be zero if she is not a program participant. Then the effect of the program on student i (TE_i) and the average effect of treatment on the treated (ATT) is given by:

$$TE_i = Y_{1i} - Y_{0i} \qquad (1)$$

$$ATT = E[TE_i \mid T_i = 1] = E[Y_{1i} \mid T_i = 1] - E[Y_{0i} \mid T_i = 1] \qquad (2)$$

The strategies presented in this paper apply this framework and employ the NAEP Mathematics scores of the 4th graders of California in 2000, who were exposed to the program to estimate the first term of (2). Under several assumptions, test scores of other groups of students

in the NAEP samples can be employed to estimate the second term of (2), the counterfactual scores, that treated students would have obtained had they not been treated.

Figure I displays NAEP math and reading scores for California and all other states. Here, one can observe that California 4th graders were performing far below the national average before CSR, though by 2003, they had narrowed the gap substantially in math, and by 1998, in reading by a lesser degree. The growth in the 4th grade Math NAEP test score observed in California between 1996 and 2003 is the largest seen among all other states in the same period. Another interesting observation is that the gap between California and other states in the 8th grade Math NAEP scores had been growing for 10 years until 2003, when the first cohort of California students who were exposed to the CSR program in earlier years were assessed as 8th graders. Following sections will investigate whether the same patterns can be seen in more sophisticated analyses. Corresponding estimation results will be presented as well.

4.1: Comparing California 4th graders in 1996 and 2000

As using variation in the program participation rates of California schools to identify CSR effects is potentially problematic and invariably provides a small sample, this paper does not employ this strategy. Instead, first, the variation in the average class sizes of schools between 1996 and 2000 is utilized when test scores of California 4th graders in 1996 and 2000 are compared. This strategy assumes the CSR program was the main factor that created the variation in the K-3 class sizes in this period.[18] Consider the following framework:

$$Y_{ijt} = X_{ijt}'\beta + Z_{jt}'\delta + \theta CS_{jt} + \pi D_t + \varepsilon_{ijt} \qquad (3)$$

[18] As stated previously, California 4th graders tested in 2000 were exposed to CSR in the first, second and third grades and 4th graders tested in 1996 were never exposed to the program. Class sizes at the school level are used because it is not possible to determine the size of the classes in which a particular student from NAEP is enrolled.

where t denotes the year in which student i from school j is tested. (t=1996 or 2000). Then, Y_{ijt} denotes her test score.[19] The vector X_{ijt} includes her characteristics and background variables, and the vector Z_{jt} denotes characteristics of school j. D_t is a dummy variable which is 1 if the corresponding observation belongs to the 2000 sample. CS_{jt} is the average of the class size of the first, second and third grades of school j in the years of interest.[20] More specifically, if t is 2000, then CS_{jt} is the average of the first, second and third grade class sizes of school j in the 1996-1997, 1997-1998 and 1998-1999 school years respectively.[21] The coefficient θ measures the effect of a one student change in the average class sizes of the first three grades and it can be used to calculate the effect of the CSR program if the program is the only factor causing the variation in class size in the 1996 and 2000 samples.

This model can be problematic as the class size variable is highly correlated with the year dummy; therefore, including these together in the regression equation poses a "multicollinearity" problem and makes the interpretation of the corresponding coefficient estimates difficult. Considering this, two specifications are estimated: one that includes both the class size variable and the cohort dummy (unrestricted specification) and another which only includes the class size variable (restricted specification). Note that coefficients on average class size variable could also absorb the natural variation in the class sizes due to various factors such as enrollment differences between schools.

[19] In all analyses of this paper, 5 plausible values for each student will be used as test scores. See Appendix I for a detailed presentation of how plausible values are employed to estimate regression coefficients. Appendix I also discusses how standard errors should be calculated when plausible values are used in an analysis.

[20] The reason why school-level class sizes are used is because it is impossible to determine the sizes of classes in which a particular student was enrolled previously. Also note that in this model, school j is the school in which student i was enrolled when she was tested by the NAEP in the fourth grade. To control for the effects for student mobility, for students who indicated that they changed schools in the third and fourth grades (by indicating so in a question in the student questionnaire), the class size variable is assigned to the average observed among the students who indicated that they did not change schools in the third and fourth grades.

[21] Similarly, if t=1996, CS_{jt} is the average of the first, second, third grade class sizes of school j in the 1992-1993, 1993-1994 and 1994-1995 school years. In this model, class size of kindergarten is not considered because California 4th graders tested in 2000 were not exposed to the CSR program in kindergarten.

Weighted Ordinary Least Squares (OLS) estimates of this model are displayed in Table IV. Test scores (plausible values) are normalized, allowing the effect estimates to be interpreted as effect sizes.[22] Columns 1, 3 and 5 use the unrestricted specification while columns 2, 4 and 6 employ the restricted one. All columns include student and school characteristics.[23] Columns 3 and 4 add teacher characteristics (experience, certification status and highest degree gained), and columns 5 and 6 add both teacher characteristics and the 4th grade class size to the independent variables.

Results displayed in Table IV show that the year dummy absorbs all of the differences between the test scores of the two cohorts tested in 1996 and 2000 when controlling for the average class size. The coefficient on the dummy variable is positive and highly significant, showing that in 2000, 4th graders indeed scored higher (at least by almost a half of a standard deviation) than those from 1996. On the other hand, the specifications that exclude the 2000 cohort dummy suggest that the higher scores of students in the 2000 NAEP sample resulted from their assignment to smaller classes in previous years. Table IV, for example, shows that a decrease in the average class size of one student in the first three grades could be linked to a test score increase of about 0.03 of a standard deviation. If we assume the CSR program reduced class sizes of the first three grades by 10 students, the effect of the program is estimated to be 0.3 of a standard deviation by this restricted specification

When interpreting estimates from this 'restricted model', one should consider that it is highly likely that other changes occurred in terms of policies and student background

[22] Plausible values are normalized by the average standard deviation, which is calculated using the variances calculated for each plausible value separately by employing sampling weights. This procedure is also followed in the remainder of the paper.

[23] Student characteristics that are controlled for are sex, race, eligibility for reduced priced lunch, Title 1 funding, limited English proficiency and individualized education plan status, and home environment and disability status. School characteristics controlled for in the analysis include racial composition, total enrollment and region

characteristics that were unrelated to the implementation of CSR between 1996 and 2000. In this model, effects of such developments would be captured by the class size variable and would confound the estimate of θ. Bohrnstedt and Stecher (1999) discuss other changes in education policies that occurred during this period, including the introduction of state-wide assessments, alterations in bilingual education, and teacher certification procedures. These developments may also have had an effect on students' academic performance. The next section will introduce a framework that attempts to isolate the effects of the CSR reform from other policy changes occurring between 1996 and 2000.

4.2 The Difference-in-Differences Estimates:

If state-wide developments in educational policy other than CSR during 1996-2000 confound the estimates presented so far, one way to overcome this problem is to estimate the influence of these policy changes and remove them from the estimates. By examining the NAEP test scores of 8^{th} graders between 1996 and 2000, one can use this group to compare to 4^{th} graders in a Difference-in-Differences (DID) framework. Note that it is plausible to assume that 8^{th} graders were affected by all the educational policy changes made during this period, except for CSR.[24]

The DID framework divides the population into sub-groups: Members affected by the policy intervention form the "treatment group," and those that are not form the "comparison group". The outcome of interest, by which the effect of the intervention is investigated, is evaluated in each group both before and after the intervention. The change of the outcome

[24] In this period, there was no other program than CSR that targeted only K-3 grades. Moreover, to my knowledge, all other initiatives listed by Bohrnstedt and Stecher (1999) were state-wide and thus plausibly affected all students equally.

observed in the treatment group is then adjusted by the change observed in the comparison group. The adjusted difference is considered as a measure of the effect of the intervention.

Here, California 8th graders sampled in 1996 and 2000 can be used as the comparison group and 4th graders from the 1996 and 2000 NAEP samples form the treatment group.[25] An investigation of the descriptive statistics of the 4th and 8th graders in Table III suggest that, overall, these groups are similar, except that 4th graders were more likely to be poor, Hispanic or LEP than 8th graders in both years. In this setting, the difference in the test scores of 8th graders between 1996 and 2000 estimates the effects of all macro-developments but CSR during this period.[26] This difference can then be used to adjust the observed difference between test scores of the 1996 and 2000 4th graders, to isolate the effects of CSR. The identifying assumption of this method is the following: had CSR never been put into effect in California, NAEP test scores in the 4th and 8th grade would have exhibited parallel patterns over the years 1996-2000.

To put these ideas into a more formal framework, consider the following model: The two periods, 1996 and 2000, are represented by t as 0 for the year 1996 and 1 for the year 2000. Let T_i be equal to 1 if student i is a 4th grader, and zero if s/he is an 8th grader. In addition, let M_i be 1 if student i is sampled in 2000, and zero if in 1996.[27] Finally, Y_{it} denotes the test score of student i, who was sampled in period t. The achievement effect of the CSR program (denoted by θ) is then given by:

[25] Note that all 4th graders assessed in 2000 are considered to be exposed to CSR. This is a plausible assumption because as seen in Table I, CSR participation in this cohort was very high and it is very hard to determine which students were never exposed to the program due to student mobility. Nevertheless, when average class sizes of schools are used to determine CSR indicators, only %2.7 of this cohort can be classified as "never been exposed to CSR". The average exposure calculated in the same manner is 2.2 years.

[26] One may think that if the program had any spillover effects (such as changes induced in the teacher population), then this may violate this assumption. Bohrnstedt and Stecher (1999) indicated that the CSR program didn't have any such effects. Descriptive statistics of the 8th graders provided in Table III supports this conclusion as well.

[27] By defining M_i and T_i in this manner, I implicitly assume that a 4th year grader was not sampled again as an 8th grader. Given that there are over 1 million students in California in a specific grade, this is a plausible assumption.

$$\theta = \{E[Y_i \mid T_i = 1, M_i = 1]\text{-}E[Y_i \mid T_i = 1, M_i = 0]\} - \{E[Y_i \mid T_i = 0, M_i = 1]\text{-}E[Y_i \mid T_i = 0, M_i = 0]\} \qquad (4)$$

Equation (4) is a modified version of the equation (2), which estimates the CSR effect in the most general sense. In (4), the change in the test scores of the 8th graders between 1996 and 2000 (E [Y_i | T_i=0, M_i=1] – E [Y_i | T_i=0, M_i=0]) is the "counterfactual change" that would have been observed in 4th grade test scores in the same period if the CSR program had not been introduced. [4] can also be represented by the following regression equation:

$$Y_{ijt} = X_{ijt}'\beta + Z_{jt}'\delta + \varphi M_i + \alpha T_i + \theta M_i T_i + \varepsilon_{ijt} \qquad (5)$$

The coefficient on the interaction of M_i and T_i gives the DID estimate of the program effects, θ. Table V presents Weighted OLS estimates of equation (5). Here, the dummy variable M_i is referred to as "after" and the label "treated" corresponds to the dummy T_i. As before, three specifications are estimated. The first includes only student and school characteristics and is displayed in column 1. The specification shown in column 2 adds teacher characteristics, and the specification in column 3 includes variables reflecting teacher characteristics and class size in 4th grade at the time of the exam. Estimates of the CSR program are displayed in the 3rd row.

The results suggest that the CSR program indeed had a positive and statistically significant influence on student test scores. For instance, column 1 shows that the CSR program led to nearly a 0.25 of a standard deviation increase in the test scores of 4th graders. Columns 2 and 3 display almost exactly the same pattern, with highly significant effect estimates of .243 and .249 respectively. As before, teacher characteristics and class sizes in 4th grade have little explanatory effect.

4.3 Models Using Propensity Score Matching in the Difference-in-Differences framework:

4.3.1 Matching and Conditional Difference-in-Differences Methods

To determine what the effects of a program might be, a valid comparison group is required, which is made up of individuals who were unaffected by the program and who otherwise exhibit similar characteristics to those exposed to the program. For this purpose, many studies (such as Dehejia and Wahba (1999, 2002), Agodini and Dynarski (2004), Blundell et. al (2003), Heckman et. al (1998), etc.) have employed matching procedures. For each treated individual, matching identifies a number of individuals with similar pre-treatment characteristics.[28]

When the average effect of a program on the treated (*ATT*) is of interest, the validity of the matching procedure relies on the following assumption: The outcome that would prevail in the absence of treatment would be the same in both treated and matched comparison populations, once all relevant observable characteristics of both groups are controlled for in the matching process (Abadie and Imbens (2005)). This is the "conditional independence assumption" (CIA), which can be hard to satisfy, especially when there are unobservable characteristics that may affect individuals' treatment status. If these unobservable characteristics do not change over time, matching in the DID framework may offer a solution to this problem.

Since the DID framework does not use actual outcomes, CIA can be relaxed when matching is used within the DID framework. The "relaxed" CIA states that *the change* in the outcome that would prevail in the absence of treatment would be the same in both treated and

[28] Matching is generally performed along pre-treatment characteristics in order to isolate the matching process from effects of the treatment.

matched comparison groups, once the relevant observable characteristics of both groups are controlled for in the matching process. This method of using matching procedures in the DID framework is often referred to as the Conditional Difference-in-Difference method (CDID) (Blundell and Costa Dias (2000), Heckman et. al (1997)).

To address these issues, let us consider the application of CDID to a panel data set to estimate the effects of a general reform program. There are two periods and each period is represented by t=0 and t=1. Next, let Y_{1it} be the outcome of student i if she participated in the program at time t. Similarly let Y_{0it} be the outcome she would experience at time t in the absence of the program. In addition, T_{it} represents whether she participated in the program at t. We can drop the time index on T_{it} if the program is introduced after period 0. In this case, T_i equals one if i is treated, and zero if i is untreated in period 1. Finally let X_i denote all pre-treatment characteristics of individual i that will be used in the matching. In this framework, *ATT* can be calculated by the following equation:

$$ATT = E\left\{E[Y_{11} - Y_{00} \mid X = x, T = 1] - E[Y_{01} - Y_{00} \mid X = x, T = 0] \mid T = 1\right\} \quad (6)$$

See Appendix II for the derivation of this equation. In (6), 'i' is dropped for simplicity. Since the NAEP samples are repeated cross sections, it is not possible to observe the same student both before and after being treated. Hence, (6) should be slightly modified to be used to estimate CSR effects in the CDID framework as:

$$E\left\{E[Y_{11} \mid X, T = 1] - E[Y_{00} \mid X, T = 1] + E[Y_{01} \mid X, T = 0] - E[Y_{00} \mid X, T = 0] \mid T = 1\right\} \quad (6')$$

The first term of (6'), $E\,[Y_{11} \mid X,\ T=1]$, can be estimated by using the 2000 NAEP scores of the California 4th graders. Estimating $E\,[Y_{00} \mid X,\ T=1]$ directly is impossible because the

treated students of 2000 were not tested before the program. This problem can be solved by matching these students with California 4th graders tested in 1996 and using 1996 4th grade test scores to estimate this term. Matching is preferred since 1996 and 2000 samples may have different characteristics and this procedure will be referred to as the "first matching procedure" (or the first matching step). For the estimation of the third term, $E\,[Y_{01} \mid X,\ T=0]$, a subset of the 4th graders from other states' 2000 NAEP samples can be used. This subset consists of students who are matched with the 2000 4th graders from California (the second matching procedure). Finally, the last term in (6') can be calculated by using test scores of the other states' 1996 NAEP 4th graders, who have similar characteristics with the California 4th graders from 2000 (the third matching procedure). Using three matching procedures for implementing CDID with repeated cross section data is suggested by Blundell and Costa Dias (2001).

4.3.2 Matching Models

In this paper, several matching models are employed. Each model is characterized by three attributes: the level at which matching is performed, whether matching is carried out with or without replacement and how many non-treated units are matched with each treated unit. First, let us consider the levels of matching. One option is performing matching at the school level. Test scores of the students from the matched schools can then be used to estimate the effects of the CSR program. Alternatively, matching can be performed at the student level. Although matching at the school level may be more suitable as schools decide whether or not to adopt the program, in this paper, matching methods performed at both levels are presented.

Note that as the number of pretreatment variables used in the matching increases, matching becomes more difficult as more untreated observations are needed to be exact matches for the treated units. This "curse of dimensionality" problem is solved by performing the

matching on a function of the pretreatment variables instead of targeting an exact match on the covariates. Rosenbaum and Rubin (1983) showed that if CIA holds for a vector of pretreatment characteristics, X, it also holds for a specific function of X, p(X). They specify p(X) to be the probability of being assigned to treatment as the *propensity score*.[29]

The second attribute that can be used to differentiate matching methods is whether matching is performed with or without replacement. Matching with replacement may use an untreated unit repeatedly whereas matching without replacement may use an untreated unit only once. Matching methods also differ by how many untreated units are used and how they are chosen in the matching process. In this paper, three different matching techniques are used: The first is the "*one to one"* matching technique, which separately sorts treated and untreated units with respect to their propensity scores to create an index in both groups, and then matches each treated unit with an untreated unit which has the same index as the treated unit.[30] The second matching technique matches each treated unit with the four most similar untreated units and untreated units can be used more than once. This method is referred as the "*nearest 4"* matching method. Finally, the third technique uses most of the untreated observations by specially weighting them, so that overall, the treated population and the weighted untreated population look alike in terms of matching characteristics.[31] This method is called "*kernel matching"* as a kernel function is used to calculate the special weights.

4.3.3 How to perform the matching and how to check its quality?

[29] All matching methods presented in this paper perform the matching process by using propensity scores.
[30] In this paper, I sorted the observations in descending order.
[31] In this weighting scheme, untreated units that have more similar characteristics (i.e. closer propensity scores) with the treated units are weighted more. A common practice of this method is discarding the untreated units that do not have a propensity score that falls into the score space spanned by propensity scores of the treated units, which is also carried out in this study.

When performing matching, first propensity scores are estimated by a logit model, which utilizes pre-treatment characteristics as the independent variables. A dummy variable which is set to one if the corresponding unit has received treatment is the dependent variable.[32] The model is then estimated and by using the estimated coefficients, the conditional probability of receiving treatment is estimated for each observation, which is the propensity score.

In the logit model, school-level matching procedures match on the following attributes: racial composition, pupil-to-teacher ratio, total enrollment, location of the school (central city, urban city, rural area) and percentage of students eligible for reduced price lunch.[33] For each school, the 1996 values of these variables are used, since these are from the pre-treatment period. For student-level matching, the following student characteristics are utilized in addition to the previously mentioned pre-treatment characteristics of their schools: sex, race, eligibility for reduced priced lunch, eligibility for Title1 funding, limited English proficiency status, individualized education plan status and home environment. [34]

Next, the three matching procedures (or matching steps) are separately carried out. When matching is used with a non-random sample, it is a common practice to assign the sampling weights of treated individuals to their matched untreated pairs (Bryson et al. 2002). In this paper, the same procedure is followed. When the *'nearest 4'* matching method at the school level is

[32] When performing matching at the school level, for example, the logit model is defined on the combined sample of 1996 and 2000 NAEP 4th grade schools of California and other states. Note that only California schools from year 2000 are treated so the dependent variable is set to one for these schools. For the other schools, it is set to zero.
[33] As the CSR program mainly affected class sizes, it's important to make sure that the treated group and matched untreated group had similar class sizes prior to the program. Although there are many studies stating that class size and pupil-teacher-ratio are different concepts, the best feasible proxy to use for class size in this project is pupil-to-teacher ratio since it's included in the Common Core Data
[34] Note that it is not possible to utilize the values of the relevant student characteristics from the pre-treatment period for those who were sampled in 2000. The variable home environment is created by students' response to the following questions: "Does you family get a newspaper regularly?", "Is there an encyclopedia in your home?", "Are there more than 25 books in your home?" and "Does your family get any magazines regularly?"

utilized, for an untreated unit denoted by index i and matched with m_i treated units ($j=1,2,\ldots,m_i$), the adjusted weight (w_i) is calculated by:

$$w_i = \sum_{j=1}^{m_i} \frac{1}{4} w_j \qquad (7)$$

where w_j is the weight of the j^{th} treated unit.[35] For each one of the three matching procedures, this process is separately carried out. Finally an assessment of match quality must be performed. Note that a high quality match would produce statistically similar treated and matched untreated groups.

In the literature, it is common to evaluate the quality of a matching process by first sorting treated and the matched untreated units by their propensity scores, then dividing observations into strata of equal score range and finally performing a t-test to see whether the average propensity scores of the treated and untreated units in each stratum are similar and a number of t-tests to check whether the treated and untreated units from each stratum are balanced along each characteristic that is used in the matching(Dehejia and Wahba (1999, 2002), Agodini and Dynarski (2004).)[36]

If all tests show that there are no significant differences between the treated units and matched untreated units, the matching procedure is finished. Otherwise, this algorithm is repeated with more subgroups. If there are still differences when tested within even finer strata, the logit model is re-specified by adding interactions and higher-order terms of the matching variables and propensity scores are re-estimated. This process, which I will refer to as the 'divide

[35] Student weights are also adjusted after corresponding school weights are re-calculated. In student level matching, the same weight adjusting scheme is used.

[36] Agodini and Dynarski (2004) perform an F-test of the similarity of the collection of the pretreatment attributes of the treatment and comparison units. In this paper, by performing separate t-tests, I aim to exhibit for which characteristics matching works and for which it doesn't.

and modify' algorithm from now on, is carried out a number of times until treatment and matched comparison groups look statistically similar.

4.3.4 School Level Matching Procedures and Corresponding CDID Estimates

In this paper, two sets of untreated schools are used. To minimize the effects of macro policy and demographic developments during 1996-2000 that cannot be controlled for in the analysis, first, schools from states close to California (Oregon, Nevada, Arizona, New Mexico, Texas and Utah[37] ("nearby states")) are used. The second set consists of schools from all states (except California) assessed by the NAEP in 1996 and 2000. This second approach attempts to increase the variety of the characteristics used in the matching procedures so that the matching quality increases.

In Figures II and III, histograms of the estimated propensity scores of the treated and untreated schools for the two sets are presented. Both figures show that although many untreated schools have smaller propensity scores, there are enough untreated schools to be used as matches for the treated schools. Leaving the detailed discussion of how well each matching method works for Appendix III, here I present a summary. According to my statistical tests, one-to-one matching is the least satisfactory in terms of balancing the pretreatment characteristics. Moreover, when matching the 2000 California schools with those in 1996 (the first matching procedure), total enrollment cannot be balanced. Similarly, California schools and schools of other states differ substantially in terms of their percentages of Hispanic students and those eligible for reduced priced lunch, these dissimilarities cannot be eliminated by the second and third matching procedures. Lastly, "the divide and modify'' algorithm produces no better results.

[37] Although Idaho and Washington are also close to California, I couldn't use them in my analysis because these states were only sampled once in 1996 and 2000.

The fact that there are still a few attributes that cannot be balanced between treated and matched untreated schools should be taken into account when estimating the effects of the CSR program; otherwise the estimates will be biased. Hence, instead of estimating $E\,[Y_{11} \mid X,\ T=1]$, $E\,[Y_{00} \mid X,\ T=1]$, and $E\,[Y_{00} \mid X,\ T=0]$ in (6') separately by using the treated units and three groups of matched untreated units, I use the following regression framework, which utilizes students from treated and matched untreated schools:

$$Y_{ijt} = X_{ijt}'\beta + Z_{jt}'\delta + \varphi M_i + \alpha T_i + \theta M_i T_i + \varepsilon_{ijt} \qquad (8)$$

In (8), M_i denotes a dummy set to 1 if student i is from the 2000 California NAEP sample or from an untreated school sampled in the same year in a nearby or other state's NAEP sample. Similarly, T_i is a dummy variable that has value 1 if student i is from a California NAEP school in 2000 or a matched 1996 California school. Then θ is the CDID estimate of the CSR effect. Note that in this framework school characteristics are also controlled for.[38]

Table VI presents Weighted OLS estimates of this model.[39] First, observe that all of the estimates of the program effects are positive and statistically significant. The smallest of the estimates is from column 4 (.193 with t-value 2.32). This estimate can be interpreted as follows: The change in the test scores of the California 4th graders is estimated to be .193 of a standard

[38] Note that values of the characteristics that are used in the matching are from the pretreatment period provided by the CCD. In the regressions I use the *current* values of the school characteristics, which are provided by the NAEP.

[39] As before, sampling weights and the jack-knife method are used in the estimation. A recent paper by Abadie and Imbens (2005) suggests that the bootstrapping method may lead to biased standard error estimates when used in the matching framework. The Jack-knife procedure, which is theoretically similar to the bootstrapping methods. Therefore, there is a chance that the standard errors calculated for the matching models may be biased as well. Instead of bootstrapping, Abadie and Imbens (2005) suggested using a closed form estimate for the standard errors of matching estimators. This suggestion could not directly be applied to this question since in this exercise three matching steps are performed and the dataset used includes sampling and replicate weights. Therefore, I utilized their suggestion for an estimate that is found by only using the first matching procedure and I observed that results from the jack-knife method and the ones from the closed form are not much different. I leave further investigation of this matter for future work.

deviation greater than the change in the test scores of the 4th graders, who have similar characteristics, from nearby states in the period 1996-2000. This change can be attributed to the CSR program if test scores of California students and students from nearby states would have followed parallel patterns in the absence of the program. By selecting similar treated and untreated schools, matching procedures increase the plausibility of this assumption. The last three columns of Table VI indicate that using schools from all states in the matching process does not change the results, except in the nearest 4 matching method.

4.3.5 Student-Level Matching Procedures and Corresponding CDID Estimates

Student-level matching procedures also used the two groups of untreated 4th graders: students from nearby states and students from all states (except California). Both student and school characteristics are utilized in the matching procedures. Although a detailed discussion of the quality of the matches is in Appendix IV, a few points are worth mentioning. School attributes are often matched better than student attributes, but when viewed as a whole, there is sufficient evidence to conclude that adequate matching was not achieved. Adding students from all other states and the 'divide and modify' procedure do not improve the results.

To tackle this problem, a new strategy is followed in which students are matched using only student attributes. As one would expect, this strategy increases the effectiveness of all matching procedures as measured by the similarity of the student characteristics in the matched groups. Moreover, when students from nearby states are used, this strategy yields matched groups that are strikingly balanced along school attributes as well. The pattern, however, disappears when students from all states are employed, leaving only student attributes balanced between treated and matched comparison groups.

In order to estimate the effects of the CSR program by using the treated and matched untreated students, the same framework is employed as in (8). By using the regression analysis, the effects of the CSR program are isolated from other factors that could not be balanced between treated and matched untreated students. Corresponding estimation results are presented in Tables VII and VIII.

Table VII indicates that when both student and school characteristics are used in the matching, estimates of program effects are more robust and closer to the estimates of previous methods than the estimates that employ school-level matching procedures. For instance, column 3 of Table VII suggests that when compared with their peers from nearby states, California students enjoyed almost a 0.25 of a standard deviation more test score growth in math between 1996 and 2000. Note that very similar figures were suggested by the DID approach, in which the comparison group was California's 8th graders.

Table VIII displays estimates from the matching procedures that only use student attributes in the matching. Here, the estimates are smaller but still positive and four of them are statistically significant. When students from all untreated states are used, estimates are lower still. One reason behind this pattern could be that school attributes are no longer matched between the treated and comparison groups, especially, pupil-to-teacher ratios differ considerably.[40]

[40] These differences suggest that although student characteristics are matched, the students come from different schools and this could reduce (and bias) the estimates down. Note that in column 2, where the matching method produced balanced treatment and comparison groups even when the school attributes are considered, the estimate is 0.193, still comparable to the previous estimates.

4.4 Heterogeneity in the effects of the CSR Program:

Often policy interventions like class size reduction affect subpopulations in varying amounts. Indeed, the Tennessee STAR study and the previous evaluations of the CSR program showed that students who were poor and/or black saw the greatest gains from smaller classes. To investigate whether this was the case here as well, I explored how various sub-groups of California students were affected by the CSR program by performing analyses on samples of each of these subgroups. Three models previously described were employed to answer this question: the DID model, Nearest 4 CDID model and Kernel CDID model. Both CDID models performed matching at the student-level, using students from nearby states and controlling for both student and school characteristics. The sub-groups used in the analysis were the following: male students, female students, student eligible for free lunch, students ineligible for free lunch, black students, Hispanic students, white students and students from urban areas.

Table IX displays coefficient estimates indicating how much each group was affected by the program. These estimates suggest that girls may have been more positively affected by CSR than boys and those in urban areas experienced greater benefits than their peers from large cities. Although the evidence is less clear as to whether there were differential effects by race, ethnicity and free lunch status, black students seem to have benefited from the CSR program more than any other racial or ethnic group. When interpreting these estimates, it is important to keep in mind that implementation of the CSR program was slower in schools that have a large population of minority students and/or eligible for reduced price lunch students.

5. Summary of Results and Discussion:

In this paper, various empirical strategies are used to estimate the achievement effects of the CSR program using data from the NAEP State Assessments. First, math scores of California 4^{th} graders in 1996 and 2000 (before and after the CSR program is introduced) were compared in a model that utilizes the average of the first, second and third grade class sizes. In this model, a dummy variable for the 2000 sample was also used to capture the uncontrolled (time-invariant) differences between the 1996 and 2000 4^{th} graders. This variable, however, was highly correlated with the class size variable. When used, the dummy variable absorbed all of the increase in the test scores between 1996 and 2000. When the dummy variable was left out, estimated class size effects were sizeable and also statistically significant, suggesting that a decrease in the average class size of one student in the first three grades could be linked to a test score increase of at least .03 of a standard deviation, which corresponds to a 0.3 of a standard deviation increase in the scores of 2000 4^{th} graders assuming the CSR program led to an approximately ten student decrease in the K-3 class sizes.

The models employing 1996 and 2000 NAEP samples of the California 4^{th} graders, particularly those that exclude the 2000 dummy, assume that there were no unobservable (or uncontrolled for) macro changes occurred between 1996 and 2000 affecting California students. This is a strict assumption that can be easily violated and hence, a new approach depending on a more plausible assumption was introduced. The new approach utilized 8^{th} graders from California sampled by the NAEP in 1996 and 2000 as a comparison group in a DID framework, where 4^{th} graders of California from 1996 and 2000 NAEP samples made up the treatment group. DID estimates suggested that CSR program significantly and positively affected

California students and the size of the effects are almost a quarter of a standard deviation in the test scores of 4th graders.

It may be argued that comparing the changes in the test scores of 4th and 8th graders may not be appropriate, although descriptive statistics provide some supporting evidence. To address this potential problem, propensity score matching models were used to find untreated students in other states who shared similar pre-treatment characteristics with the California students. Potential untreated matches were chosen either from nearby states or from all other states, before and after CSR and several matching procedures were carried out.

Quality tests of matching procedures showed that some attributes of California schools and students (such as pupil-to-teacher ratio and ethnicity) differed from potential untreated schools and students so greatly that in some cases perfect matches could not be achieved. To account for these differences, treated and matched untreated students were compared, in a Conditional DID regression framework.

Corresponding estimates suggested that the CSR program positively affected California 4th graders when compared with 4th graders from both nearby and all other states. Moreover, the estimates from the models that match students at the school level and those that match students at the student level using both school and student attributes are similar to the previous estimates: between 0.2 and 0.3 of a standard deviation increase in test scores. Estimates from student-level matching methods that only control for student attributes were smaller but still significant and positive.

Finally, whether the CSR program has had heterogeneous effects is investigated. Although the evidence is less clear as to whether there were differential effects by race, ethnicity

and free lunch status; black students seem to have benefited from the CSR program more than any other racial or ethnic group.

6. Conclusion:

The California Class Size Reduction program was one of the most ambitious educational reforms ever enacted in the United States. Since 1996, the program has affected more than 1 million students each year, with annual operating costs of more than $1 billion. A number of earlier studies have evaluated the program with mixed results. A compelling assessment of the achievement effects of the program has been hindered by the absence of any statewide exams given before the program began. As a result, previous analyses had to compare achievement levels of schools that implemented class size reduction to those that did not, which is problematic as there is evidence to suggest that there were systematic differences between the participating schools and the small number of non-participating schools. Stecher and Bohrnstedt (2000) attempted to deal with this problem by subtracting the differential between fifth graders at participating and non-participating schools from the differential between third graders, in order to eliminate all the other possibly confounding differences between these two sets of schools. Yet by doing so, they could not fully solve the problem, since 18% of fifth graders in California had also been in smaller classes in earlier years.

This study addresses the problem of finding adequate baseline test data by employing State NAEP scores in Mathematics of California students before and after CSR, and uses student-level achievement in the analysis. With several techniques, I isolate the effects of smaller classes from other educational changes that were occurring in California concurrently.

I find that the achievement effects of the program were statistically significant and positive. For example, when California 8th graders are used as a comparison group, estimates from the DID framework reveal that test scores of California 4th graders, who were affected by CSR grew by at least 0.25 of a standard deviation between 1996 and 2000. Although the evidence is less clear as to whether there were differential effects by race, ethnicity and free lunch status; black students seem to have benefited from the CSR program more than any other racial or ethnic group. Finally, the estimated effects are fairly robust in all the various specifications and approaches.

Although this study provides an extensive analysis of the achievement effects of the California CSR program, it does not answer the question of whether this program has been cost-effective. Therefore, a cost-benefit analysis will be required to draw a more complete picture to policymakers who are considering whether to implement class size reduction programs. Other future topics ripe for investigation include analyzing whether the California CSR program has had long-lasting effects beyond third grade, and whether the introduction of smaller classes in K-3 has influenced other measures of school climate or educational outcomes, such as school crime, disciplinary problems, grade retention, attendance and/or dropout rates. An extension of these results using test scores from recent years (after 2000) and other test subjects, such as reading would also be useful.

REFERENCES

Abedie, A. (2005). Semiparametric difference-in-differences estimators. *Review of Economic Studies 72(1)*, 1-19 (2005).

Abedie, A. and G. Imbens (2005). Large sample properties of matching estimators for average treatment effects. *Mimeo*

Abedie, A. and G. Imbens (2004). On the failure of the bootstrap for matching estimators. *Mimeo*.

Agodini, R. and M. Dynarski (2004). Are experiments the only option? A look at dropout prevention programs. *The Review of Economics and Statistics 86(1),* 180-194.

Blundell, R., M. C. Dias, C. Meghir and J. V. Reenen (2003). Evaluating the employment impact of a mandatory job search assistance program. *Mimeo*

Blundell, R. and M. C. Dias (2000). Evaluation methods for experimental data. *Fiscal Studies, 21(4),* 427–468.

Bohrnstedt, G. W. and B. M. Stecher (Eds.) (2002). What we have learned about class size reduction. Sacramento, CA: California Department of Education.

Bohrnstedt, G. W. and B. M. Stecher (Eds.) (1999). Class size reduction in California: Early evaluation findings, 1996–1998. Palo Alto, CA: American Institutes for Research.

Bryson, A., R. Dorsett, and S. Purdon (2002). The use of propensity score matching in the evaluation of active labour market policies. *Department for Work and Pensions, Working Paper No.4.*

Dehejia, R. H. and S. Wahba (2002). Propensity score matching methods for non-experimental causal studies. *Review of Economics and Statistic*s 8*4*, 151-161.

Dehejia, R. H. and S. Wahba (1999). Casual effects in non-experimental studies: reevaluating the evaluation of training Programs. *Journal of the American Statistical Association 94(448)*, 1053-1062.

Finn, J. D. and C. M. Achilles (1990). Answers and questions about class size: A statewide experiment. *American Educational Research Journal 28*, 557-577.

Hanushek, E. A. (2003). The failure of input-based schooling policies. *Economic Journal 113*, 64-98.

Hanushek, E. A. (2002). Evidence, Politics, and the Class Size Debate. In L. Mishel and R. Rothstein (Eds.), *The Class Size Debate*, 37-65. Washington, DC: Economic Policy Institute.

Hanushek, E. A. (1996). School resources and student performance. In G. Burtless (Eds.) *Does Money Matter? The Effect of School Resources on Student Achievement and Adult Success*, Washington, D.C.: Brookings Institution.

Hanushek, E. A. (1986). The economics of schooling: Production and efficiency in public schools. *Journal of Economic Literature 24*, 1141-1177.

Hanushek, E. A. (1981). Throwing money at schools. *Journal of Policy Analysis and Management 1(1),* 19-41.

Heckman, J. J., H. Ichimura, J. Smith, and P. E. Todd (1998). Characterizing selection bias using experimental data. *Econometrica, 66(5)*.

Heckman, J. J., and H. Ichimura and P. E. Todd (1997). Matching as an econometric evaluation estimator: evidence from evaluating a job training program. *Review of Economic Studies 64(4)*, 605-54.

Krueger, A. B. (2003). Economic considerations and class size. *Economic Journal 113*, 34-63.

Krueger, A. B. (1999). Experimental estimates of education production functions. *Quarterly Journal of Economics 1144*, 497-532.

Krueger, A. B. and D. M. Whitmore (2001). The effect of attending a small class in the early grades on college-test taking and middle school test results: Evidence from Project STAR. *Economic Journal 111*.

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika 56*, 177-196.

Pate-Bain, H., J. Boyd-Zaharias, V. A. Cain, E. Word and E. Binkley (1997). STAR follow-up studies, 1996-1997: The Student/Teacher Achievement Ratio (STAR) Project. Heros, Inc. Lebanon Tennessee.

Peterson, M. L. and K. Rheault (1995). The Nevada Class Size Reduction Evaluation Study. *Mimeo*

Rivkin S. G. and C. Jepsen (2002). What is the tradeoff between smaller classes and teacher quality. *NBER Working Paper # 9205*

Rosenbaum, P. R. and D. B. Rubin (1983). The central role of the propensity score in observational studies for causal effects. *Biometrica 70*, 41-55

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: John Wiley and Sons.

Sims, D. P. (2003). How flexible is educational production? Combination classes versus class size. *Mimeo*

Smith, P., A. Molnar and J. Zahorik (2003) Class size reduction in Wisconsin: A fresh look at the data. Education Policy Research Unit, Arizona State University.

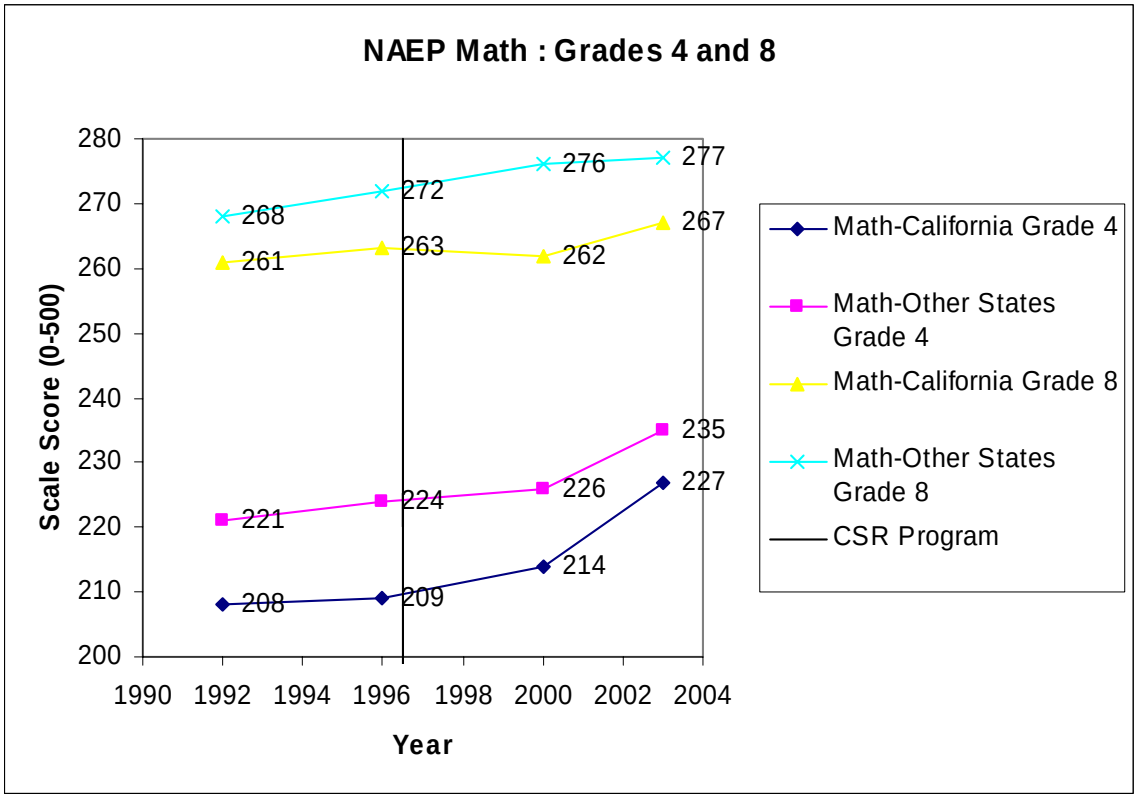
Snow, M. (1993). The 1993 class size reduction evaluation study. Nevada Department of Education.

Stecher, B. M. and G. .W. Bohrnstedt (Eds.) (2002). Class size reduction in California: Findings from 1999–00 and 2000–01. Sacramento, CA: California Department of Education. (2002).

Stecher, B. M. and G. W. Bohrnstedt (Eds.) (2000). Class size reduction in California: The 1998–99 evaluation findings. Sacramento, CA: California Department of Education.

Sturm, H. P. (1997). Nevada's class size reduction program. Carson City, NV: Nevada Legislative Counsel Bureau.

Word, E. et al. (1990). Student/Teacher Achievement Ratio (STAR) Tennessee's K–3 Class Size Study. Final Summary Report.



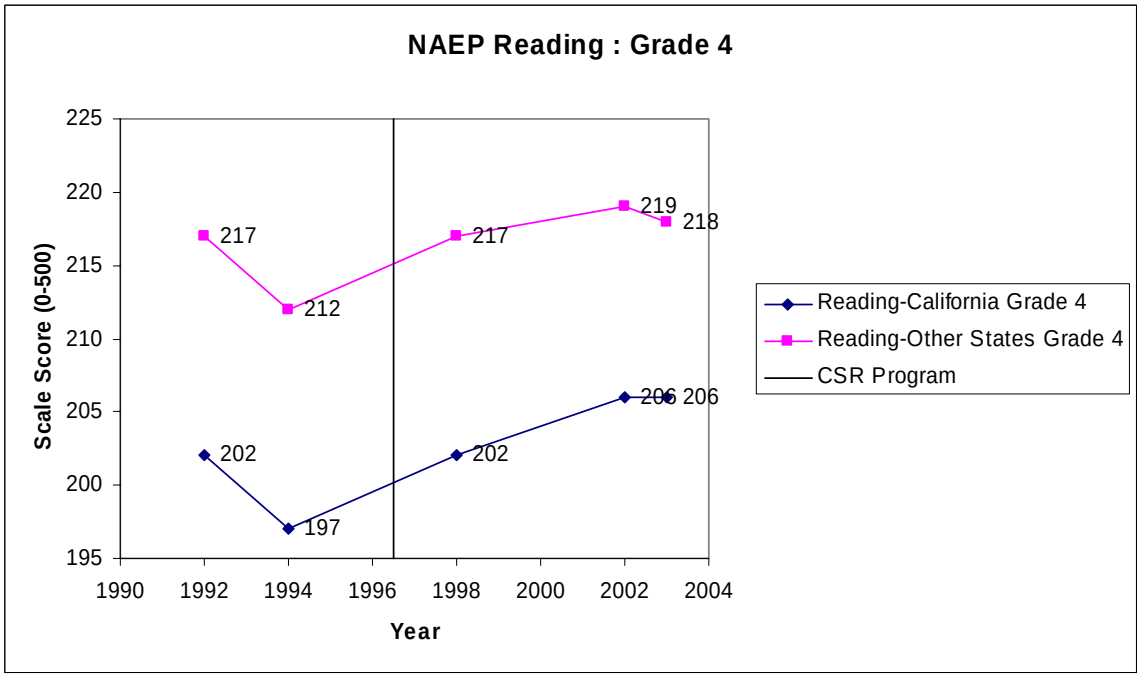


Figure I: Trends in NAEP Mathematics and Reading Tests: California and Other States are compared. Source: http://nces.ed.gov/nationsreportcard/

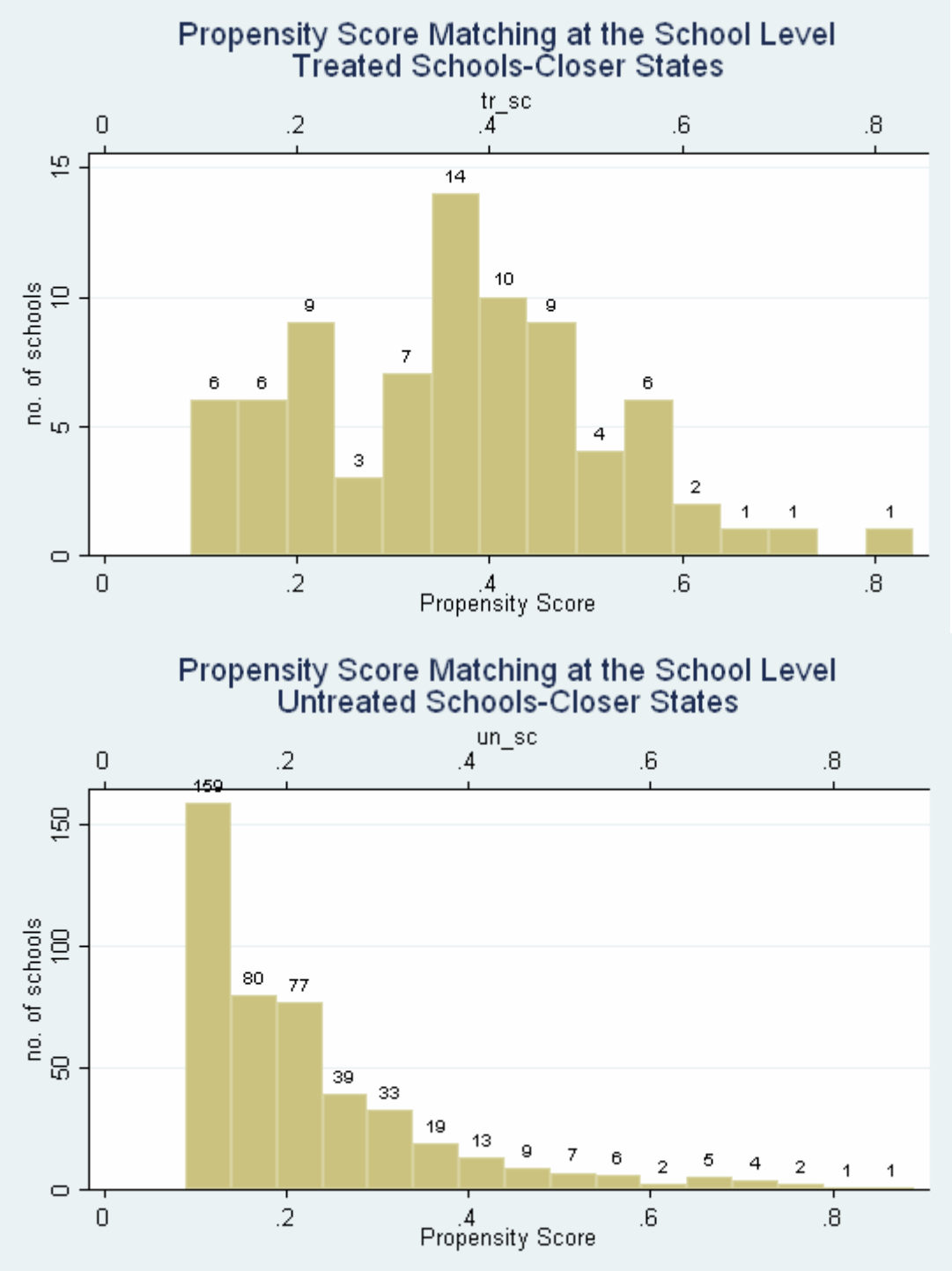


Figure II: Histogram of the propensity scores of the treated and untreated schools. Only nearby states to California are used and matching is performed at the school level.

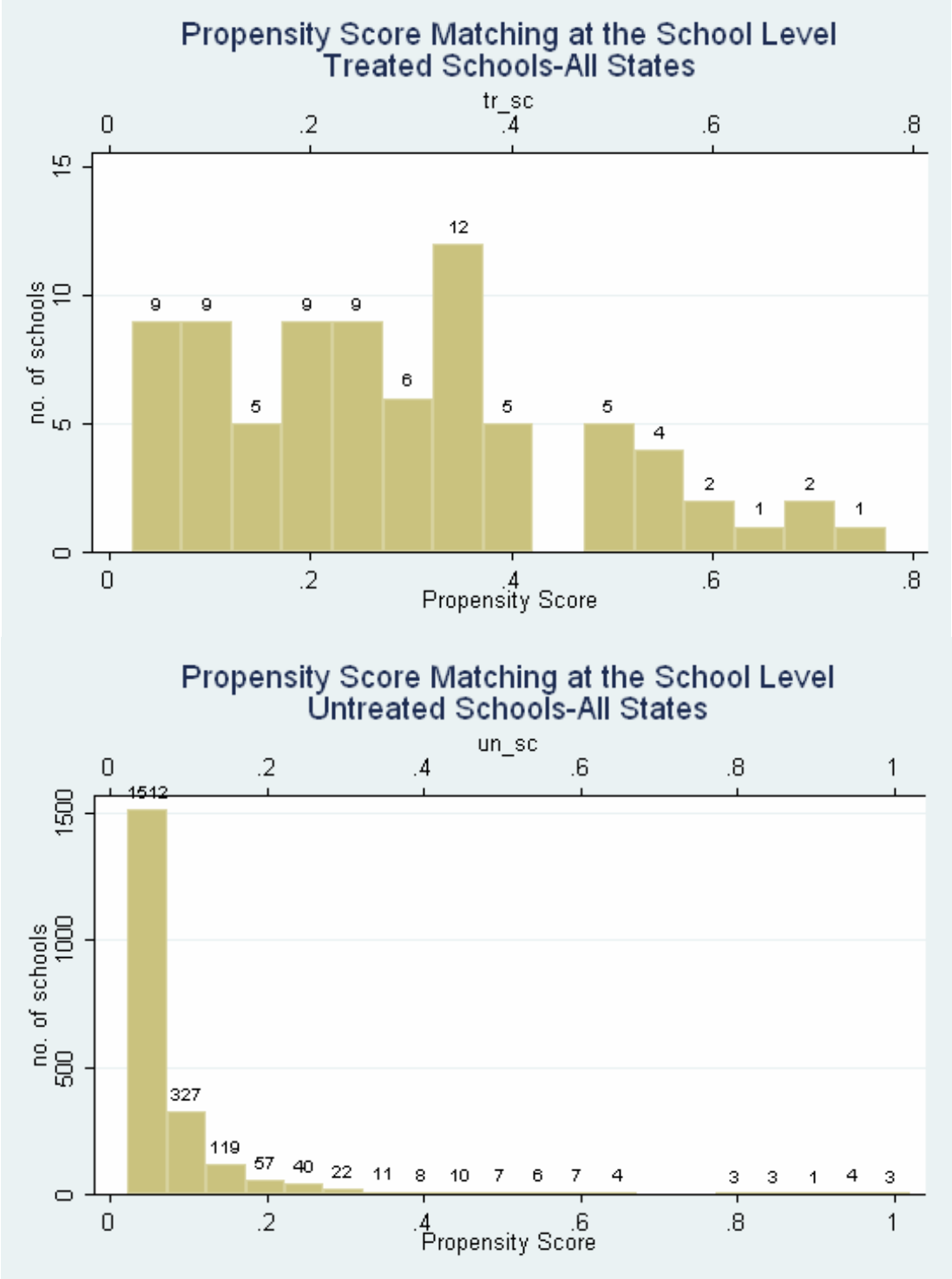


Figure III: Histogram of the propensity scores of the treated and untreated schools. All states are used and matching is performed at the school level.

Table I : Percentage of California Students in Reduced Classes by Year and Grade					
School year	Grade Levels				Districts not participated
	Kindergarten	Grade 1	Grade 2	Grade 3	
1996-1997	14	88	57	18	56
1997-1998	69	99	96	67	20
1998-1999	86	99	98	84	13
1999-2000	96	99	97	91	9
2000-2001	96	99	97	91	9
2001-2002	96	98	97	92	6
2002-2003	94	97	98	93	7
2003-2004	94	97	96	86	13
2004-2005	94	97	95	87	10

Notes: There are about 900 school districts in California.
Source: Fingertip Facts, California Department of Education
http://www.cde.ca.gov/ls/cs/k3/facts.asp

Table II: California Schools-CSR Adopters and Non-Adopters
2nd grade CSR Participation in 1997-1998, 1998-1999 and 1999-2000 School Years

	1997-1998		1998-1999		1999-2000[1]	
	CSR (N=2298)	non-CSR (N=48)	CSR (N=2342)	non-CSR (N=7)	CSR (N=2349)	non-CSR (N=0)
School Characteristics:						
Total enrollment	664.74 (263.66)	740.70 (203.46)	664.74 (263.66)	882.29 (227.46)	668.05 (278.75)	-
Percentage Black	8.34 (11.80)	4.22 (6.72)	8.34 (11.80)	9.07 (8.20)	8.05 (11.62)	-
Percentage Hispanic	41.51 (28.50)	72.93 (20.85)	41.51 (28.50)	73.08 (7.69)	44.23 (29.09)	-
Percentage eligible free lunch	54.73 (29.63)	73.23 (20.31)	54.73 (29.63)	66.77 (22.9)	46.59 (28.57)	-
Pupil-teacher ratio	20.96 (2.09)	24.73 (2.73)	20.96 (2.09)	23.8 (2.68)	20.04 (2.77)	-
Percentage located in large city	44.56 (49.71)	25.0 (43.80)	44.56 (49.71)	0.0	39.16 (48.82)	-
Percentage located in urban city	51.49 (49.98)	72.72 (45.05)	51.49 (49.98)	85.72 (37.79)	52.16 (49.96)	-
Percentage located in rural area	3.94 (19.46)	2.27 (15.07)	3.94 (19.46)	14.28 (37.79)	8.68 (28.17)	-
Teacher Characteristics:						
Percentage highest bachelor degree	73.58 (14.66)	70.55 (16.93)	74.12 (14.46)	77.62 (7.84)	74.00 (14.25)	-
Percentage tenured	65.57 (14.80)	66.85 (15.11)	66.55 (17.01)	61.15 (14.21)	68.73 (16.86)	-
Experience	12.70 (3.26)	12.74 (3.21)	12.44 (3.22)	12.06 (2.19)	12.48 (3.22)	-
Percentage full credentials	87.24 (12.82)	81.30 (20.53)	87.22 (13.35)	79.73 (14.71)	86.94 (13.45)	-
1996 NAEP Math score (4th grade)	205.02 (31.10) NS=957	200.14 (31.20) NS=45	204.75 (31.25) NS=1002	-	204.75 (31.25) NS=1002	-

Notes: Standard deviations are displayed in parenthesis. School Characteristics are obtained from the common core data and teacher characteristics are gathered from California Department of Education. CSR Consortium's classification of schools with respect to CSR

[1] There are no schools that did not adopt the CSR program in the 1999-2000 school year at the second grade. These cohorts of students are only used by Sims (2003).

Table II-cont'd: California Schools-CSR Adopters and Non-Adopters
3rd grade CSR Participation in the 1998-1999 and 1999-2000 School Years

	1998-1999		1999-2000	
Used By	Sims (2003), Stecher & Bohrnstedt (2000)		Sims(2003), Jepsen & Rivkin (2002)	
	CSR (N=2142)	non-CSR (N=206)	CSR (N=2294)	non-CSR (N=55)
School Characteristics:				
Total enrollment	661.00 (273.24)	734.26 (200.65)	667.12 (272.06)	703.6 (212.23)
Percentage Black	8.43 (12.02)	5.34 (5.92)	8.16 (11.74)	3.57 (3.71)
Percentage Hispanic	42.21 (28.94)	52.72 (25.81)	43.78 (29.04)	61.09 (25.59)
Percentage eligible free lunch	46.23 (28.36)	48.28 (25.05)	46.47 (28.67)	51.00 (24.16)
Pupil-to-teacher ratio	19.9 (1.88)	22.31 (2.27)	19.98 (2.78)	22.12 (1.28)
Percentage located in large city	39.66 (48.93)	34.71 (43.80)	39.41 (48.88)	29.09 (45.84)
Percentage located in urban city	51.55 (49.98)	59.06 (49.29)	51.85 (49.98)	63.67 (48.55)
Percentage located in rural area	8.77 (28.30)	6.21 (4.42)	8.72 (28.22)	7.27 (26.20)
Teacher Characteristics:				
Percentage highest bachelor degree	74.10 (14.62)	74.42 (12.71)	73.97 (14.27)	76.86 (12.23)
Percentage tenured	66.26 (17.23)	69.62 (13.89)	68.75 (16.83)	67.18 (17.74)
Experience	12.41 (3.27)	12.80 (2.94)	12.48 (3.22)	12.05 (3.09)
Percentage full credentials	87.29 (13.53)	87.61 (11.20)	86.94 (13.45)	86.29 (13.04)
1996 NAEP Math score (4th grade)	204.71 (31.22) NS=957	205.40 (31.65) NS=69	204.75 (31.25) NS=1002	-

Notes: Standard deviations are displayed in parenthesis. School Characteristics are obtained from the common core data and teacher characteristics are gathered from California Department of Education. CSR Consortium's classification of schools with respect to CSR participation in the third grade in a given year is used to group the schools. In the last row, N represents total number of students who took the NAEP 1996 Math exam in the corresponding schools.

Table III: Descriptive Statistics

	4th Grade California		4th Grade All Other States		8th Grade California	
	1996 (N=2319)	2000 (N=1656)	1996 (N=116636)	2000 (N=93327)	1996 (N=2522)	2000 (N=1628)
Student Characteristics						
Male	.513 (.50)	.504 (.50)	.507 (.50)	0.494 (.50)	.485 (.05)	.507 (.05)
White	.421 (.49)	.356 (.478)	.65 (.474)	.615 (.486)	.411 (.492)	.342 (.474)
Black	.075 (.263)	.086 (.282)	.157 (.364)	.178 (.382)	.078 (0.269)	.072 (.259)
Hispanic	.373 (.483)	.41 (.491)	.133 (.339)	.15 (.356)	.364 (.481)	.435 (.495)
Asian	.107 (.309)	.114 (.317)	.02 (.160)	.0291 (.168)	.128 (.333)	.139 (.347)
Limited English proficient	.146 (.353)	.217 (.412)	.025 (.157)	.025 (.157)	.066 (.247)	.16 (.366)
Individualized education plan	.035 (.18)	.054 (.226)	.057 (0.231)	.048 (.214)	.039 (.193)	.049 (.217)
Reduced priced lunch eligible	.391 (.489)	.487 (.499)	.362 (.480)	.385 (.487)	.325 (.469)	.352 (.481)
Home environment						
0-2 types	.375 (.484)	.469 (.499)	.292 (.454)	.379 (.485)	.278 (.449)	.293 (.455)
3 types	.304 (.458)	.277 (.477)	.312 (.463)	.320 (.466)	.306 (0.46)	.302 (.459)

Notes: All statistics are calculated using sampling weights. Standard deviations are displayed in parenthesis. Home environment tabulates how many times the student answers the following questions as "Yes": "Does you family get a newspaper regularly?", "Is there an encyclopedia in your home?", "More than 25 books?" and "Get any magazines regularly?" Category 4 is left out for this variable. If the label denoting the characteristic is a dummy, then corresponding values in the cells are proportions.

Table III: Descriptive Statistics-cont'd

	4th Grade California		4th Grade All Other States		8th Grade California	
	1996 (N=2319)	2000 (N=1656)	1996 (N=116636)	2000 (N=93327)	1996 (N=2522)	2000 (N=1628)
School Characteristics						
Enrollment	679.77 (298.8)	688.04 (350.79)	656.64 (716.5)	572.30 (411)	1163.48 (1072.4)	1130.92 (610.69)
Percentage Black	7.7 (10.6)	9.85 (14.67)	16.58 (24.08)	19.03 (27.72)	7.46 (9.05)	8.13 (11.16)
Percentage Hispanic	33.58 (25.49)	37.42 (28.09)	8.64 (18.21)	10.03 (20.10)	30.62 (22.91)	37.59 (25.86)
Pupil-to-teacher ratio	25.70 (2.69)	--	18.05 (3.10)	--	--	--
Large city	.415 (.493)	.388 (.487)	.325 (.468)	.327 (.469)	.422 (.494)	.389 (.487)
Urban city	.519 (.499)	.578 (.493)	.399 (.490)	.369 (.482)	.533 (.498)	.578 (.494)
Rural region	.065 (.247)	.034 (.180)	.272 (0.445)	.301 (.459)	.045 (.206)	.033 (.179)
teacher characteristics						
Female	.754 (.43)	.676 (.475)	.82 (0.384)	.817 (0.387)	.477 (.499)	.498 (.500)
Experience: 0-9 years	.403 (.49)	.565 (.495)	.357 (0.479)	.384 (0.486)	.501 (.500)	.534 (.499)
Experience: 10-25 years	.334 (.471)	.183 (.386)	.397 (0.489)	.334 (0.472)	.293 (.455)	.262 (.440)
Highest degree: bachelor	.645 (.479)	.640 (.48)	.511 (0.499)	.499 (0.50)	.645 (.476)	.70 (.454)
Certified	.91 (.826)	.788 (1.13)	.942 (1.14)	.927 (1.16)	.92 (.767)	.87 (.699)

Notes: All statistics are calculated using sampling weights. Standard deviations are displayed in parenthesis. If the label denoting the characteristic is a dummy, then corresponding values in the cells are proportions.

Table IV: Effects of the CSR Reform-1966 and 2000 California 4th grade NAEP samples
Average Class Size is used

	I	II	III	IV	V	VI
Explanatory Variables						
Grades 1-3 average class size	.018 (.017)	-.033*** (.006)	.015 (.015)	-.034*** (.006)	.017 (.015)	-.032*** (.006)
Year 2000 dummy	.499*** (.137)	-	.479*** (.137)	-	.493*** (.139)	-
Student & school characteristics	Yes	Yes	Yes	Yes	Yes	Yes
Teacher characteristics	No	No	Yes	Yes	Yes	Yes
4th grade class size	No	No	No	No	Yes	Yes
Sample size	2995	2995	2995	2995	2995	2995
R^2	.38	.38	.39	.39	.39	.38

Notes: *** denote statistical significance at 1% level. All regressions are weighted by sampling weights. Standard errors are calculated by the jack-knife method and displayed in parentheses. Each column corresponds to a separate regression. Student Characteristics include sex, race, and eligibility for reduced priced lunch and Title1 funding, limited English proficiency status, disability and individualized education plan status. School characteristics are racial composition, total enrollment and region. Teacher characteristics include certification status, experience and degree. The year 2000 dummy is set to one if the observation is from 2000 NAEP sample and zero otherwise.

Table V: Difference in Difference Estimates of the Effects of the CSR Reform
1996 & 2000 NAEP California 8th Graders Used as the Comparison Group

	I	II	III
Explanatory Variables:			
After	.08 (.05)	.086* (.049)	.086* (.048)
Treated	-.405*** (.056)	-.410*** (.056)	-.419*** (.057)
After*Treated	.230*** (.074)	.243*** (.075)	.249*** (.076)
Student and school characteristics	Yes	Yes	Yes
Teacher characteristics	No	Yes	Yes
Class size at the time of the test	No	No	Yes
Sample size	7729	7729	7729
R^2	.41	.41	.42

Notes: *, and *** denote statistical significance at the 10% and 1% levels respectively. All regressions are weighted by sampling weights. Standard errors are calculated by the jack-knife method and displayed in parentheses. Each column corresponds to a separate regression. Student Characteristics include sex, race, and eligibility for reduced priced lunch and Title1 funding, limited English proficiency status, disability and individualized education plan status. School characteristics are racial composition, total enrollment and region. Included teacher characteristics control for teachers' experience, certification and degree. Dummy variable after is set to 1 if the corresponding observation is from year 2000. Dummy variable Treated is set to 1 if the student is from the 4th graders NAEP sample.

Table VI: School Level Matching-1996 and 2000 NAEP Samples
Conditional Difference in Difference Estimation

	Untreated schools from nearby states			Untreated schools from all states		
Matching method	One-to-One	Nearest 4	Kernel	One-to-One	Nearest 4	Kernel
Explanatory Variables:						
After	.101 (.075)	.105 (.068)	.052 (.086)	.163*** (.061)	-.04 (.089)	-.004 (.068)
Treated	-.149** (.067)	-.085 (.105)	-.178 (.120)	-.236*** (.061)	-.312*** (.092)	-.315*** (.077)
After*Treated	.235*** (.086)	.211** (.097)	.336*** (.125)	.193** (.083)	.359*** (115)	.333*** (.090)
Student and school characteristics	Yes	Yes	Yes	Yes	Yes	Yes
Teacher characteristics	No	No	No	No	No	No
Sample size	7017	7669	11948	6843	9775	51230
R^2	.36	.38	.39	.41	.38	.40

Notes: **, *** denote statistical significance at the 5% and 1% levels respectively. All regressions are weighted by sampling weights. Standard errors are calculated by the jack-knife method and displayed in parentheses. Each column corresponds to a separate regression. Student Characteristics include sex, race, and eligibility for reduced priced lunch and Title1 funding, limited English proficiency status, disability and individualized education plan status. School attributes are racial composition, total enrollment and region. Matching procedures are performed at the school level, regression are estimated at the student level.

Table VII : Student Level Matching – Student and School Attributes Used in the Matching- 1996 and 2000 NAEP Samples
Conditional Difference in Difference Estimation

	Untreated students from nearby states			Untreated students from all states		
Matching Method	One-to-One	Nearest 4	Kernel	One-to-One	Nearest 4	Kernel
Matching Variables:						
Student characteristics	Yes	Yes	Yes	Yes	Yes	Yes
School characteristics	Yes	Yes	Yes	Yes	Yes	Yes
Explanatory Variables:						
After	.141***	.135	.110	.103**	.099	.097*
	(.051)	(.093)	(.079)	(.046)	(.063)	(.056)
Treated	-.188***	-.221***	-.181**	-.178***	-.189**	-.192***
	(.053)	(.080)	(.073)	(.056)	(.072)	(.060)
After*Treated	.191**	.244**	.234**	.250***	.245**	.255***
	(.075)	(110)	(.096)	(.074)	(.095)	(.087)
Student and school characteristics	Yes	Yes	Yes	Yes	Yes	Yes
Teacher characteristics	No	No	No	No	No	No
Sample size	6164	5821	14463	6164	6894	73430
R^2	.35	.35	.36	.39	.37	.40

Notes: *, **, *** denote statistical significance at the 10%, 5% and 1% levels respectively. All regressions are weighted by sampling weights. Standard Errors are calculated by the jack-knife method and displayed in parentheses. Each column corresponds to a separate regression. Matching variables are the attributes used in the matching procedures and explanatory variables are the ones used in the regression estimation. Student Characteristics include sex, race, and eligibility for reduced priced lunch and Title1 funding, limited English proficiency status, disability and individualized education plan status. School characteristics are racial composition, total enrollment and region. For the matching procedures, values of school characteristics from the pre-treatment period are used. For regression estimations, original values for the corresponding variables are employed.

Table VIII : Student Level Matching – Only Student Attributes Used in the Matching- 1996 and 2000 NAEP Samples
Conditional Difference in Difference Estimation

	Untreated students from nearby states			Untreated students from all states		
Matching method	One-to-One	Nearest 4	Kernel	One-to-One	Nearest 4	Kernel
Matching Variables:						
Student Characteristics	Yes	Yes	Yes	Yes	Yes	Yes
School Characteristics	No	No	No	No	No	No
Explanatory Variables:						
After	.146***	.153*	.195***	.204***	.230***	.174***
	(.04)	(.083)	(.028)	(.041)	(.085)	(.020)
Treated	-.272***	-.264***	-.255***	-.247***	-.232**	-.225***
	(.052)	(.095)	(.050)	(.056)	(.095)	(.051)
After*Treated	.180***	.193*	.144**	.130*	.102	.143**
	(.067)	(.115)	(.062)	(.072)	(.107)	(.061)
Student and School Characteristics	Yes	Yes	Yes	Yes	Yes	Yes
Teacher Characteristics	No	No	No	No	No	No
Sample Size	6276	3968	29664	6276	4193	194262
R^2	.34	.33	.34	.35	.35	.35

Notes: *, **, *** denote statistical significance at the 10%, 5% and 1% levels respectively. All regressions are weighted by sampling weights. Standard Errors are calculated by the jack-knife method and displayed in parentheses. Each column corresponds to a separate regression. Matching variables are the attributes used in the matching procedures and explanatory variables are the ones used in the regression estimation. Student Characteristics include sex, race, and eligibility for reduced priced lunch and Title1 funding, limited English proficiency status, disability and individualized education plan status. School characteristics are racial composition, total enrollment and region.

Table IX: Effects of the CSR Program on Various Sub-Populations-1996 and 2000 NAEP Samples
DID and CDID Estimates

Estimation Method	(I) DID	(II) CDID-Nearest 4	(III) CDID-Kernel
Subpopulation:			
Male	.198**	.208	.304**
	(.092)	(.153)	(.133)
Sample Size	3812	2916	7310
Female	.267***	.334***	.332***
	(.08)	(.127)	(.100)
Sample Size	3917	3017	6855
Free Lunch Eligible	.249***	.247*	.257*
	(.095)	(.156)	(.147)
Sample Size	2852	2549	6702
Free Lunch Non-Eligible	.182*	.302***	.236***
	(.096)	(105)	(.087)
Sample Size	3614	2754	8707
White	.236***	.314***	.293***
	(.084)	(.114)	(.095)
Sample Size	3177	2549	7942
Black	.290*	.401	.325
	(.157)	(.359)	(.350)
Sample Size	650	523	983
Hispanic	.192**	.127	.192
	(.097)	(.151)	(.153)
Sample Size	2813	1983	4408
Urban City	.265***	.334**	.363***
	(.097)	(.140)	(.138)
Sample Size	4127	3150	7865
Large City	.120	.064	.038
	(.128)	(.170)	(.158)
Sample Size	3188	2169	4348

Notes: *, **, *** denote statistical significance at the %10, %5 and %1 levels respectively. All regressions are weighted by sampling weights. Standard Errors are calculated by the jack-knife method and displayed in parenthesis. Each cell corresponds to a separate regression. In the last two columns, matching is performed at the student level using students from nearby states.

APPENDIX I: Jack-Knife Variance Estimation Procedure

Assume there are M sets of plausible values and N reweighed sub-samples and assume that we are estimating a statistic t (can be the mean of a variable, a regression coefficient, etc). Then t_m denotes the corresponding statistic when the m^{th} plausible value (m=1,2,…,M) is used.

The final estimate of t will be:

$$t^* = \frac{1}{M}\sum_{m=1}^{M} t_m$$

To calculate the sample variance for the m^{th} estimate of t, first calculate N sets of estimates using the N sub-samples and denote them by t^n_m. Then, variance of t_m (denoted by U_m) is given by:

$$U_m = \sum_{n=1}^{N} (t_m^n - t_m)^2$$

Then, compute the average sampling variance over the M sets of plausible values:

$$U^* = \frac{1}{M}\sum_{m=1}^{M} U_m$$

There is also variability caused by the plausible values, which is denoted by B:

$$B^* = \frac{1}{M-1}\sum_{m=1}^{M} (t_m - t^*)^2$$

The overall variance of the estimate t* is then given by:

$V=U^*+(1+M^{-1})B^*$

APPENDIX II: CDID Estimator

Consider the following model to calculate the average effect of treatment on the treated individuals in the conditional difference-in-differences framework. There are two periods and each period is represented by t, having values 0 and 1. Next, let Y_{1it} be the outcome of student i if she participated in the program of interest at time t. Similarly let Y_{0it} be the outcome she would experience at time t in the absence of the program. In addition, T_{it} represents whether i participated in the program at time t. That is, T_{it} is equal to 1 if i is affected by the program at time t. The time index on T_{it} can be dropped if the program is introduced after period 0. In this case, T_i equals one if i is treated, and zero if i is untreated in period 1. Finally let X_i denote all pre-treatment characteristics of individual i that will be used in the matching. For individual i, the treatment effect can then be represented as:

$$TE_i = Y_{1i1} - Y_{0i1} \tag{AII-I}$$

Then, the average effect of the CSR on the treated population will be:

$$ATT = E[TE_i \mid T_i = 1] = E[Y_{1i1} \mid T_i = 1] - E[Y_{0i1} \mid T_i = 1] \tag{AII-II}$$

$$ATT = E[Y_{1i1} - Y_{0i0} \mid T_i = 1] - E[Y_{0i1} - Y_{0i0} \mid T_i = 1] \tag{AII-II'}$$

(AII-II') adds and subtracts the term $E[Y_{0i0} \mid T_i = 1]$ to (AII-II). In this setting, the relaxed conditional independence assumption (rCIA) can be represented by the following statement:

$$E[Y_{0i1} - Y_{0i0} \mid X_i, T_i = 1] = E[Y_{0i1} - Y_{0i0} \mid X_i, T_i = 0] \tag{rCIA}$$

Next, let's calculate the ATT for the treated subpopulation represented by X=x:

$$ATT(x) = E[Y_{1i1} - Y_{0i0} \mid X = x, T_i = 1] - E[Y_{0i1} - Y_{0i0} \mid X = x, T_i = 0] \quad \text{(AII-III)}$$

For this group, matching finds a subpopulation among the untreated that has similar pretreatment characteristics (X=x). If rCIA is satisfied for these two subgroups, (AII-III) can be rewritten as:

$$ATT(x) = E[Y_{11} - Y_{00} \mid X = x, T = 1] - E[Y_{01} - Y_{00} \mid X = x, T = 0] \quad \text{(AII-IV)}$$

In (AII-IV), the individual index 'i' is dropped for simplicity. Finally ATT the whole treated population can be found by evaluating $ATT(x)$ on the distribution of X conditional on T:

$$ATT = E\left\{E[Y_{11} - Y_{00} \mid X = x, T = 1] - E[Y_{01} - Y_{00} \mid X = x, T = 0] \mid T = 1\right\} \quad \text{(AII-V)}$$

APPENDIX III: Testing the Quality of the School Level Matching Methods

In this section, I investigate how well the treated schools of California are matched with the untreated schools that are used in the CDID procedures. Table AIII-I, I present results from the t-tests that check the quality of the matches. The First panel of Table AIII-I shows that California Schools from 2000 and 1996 differ a lot in terms of total enrollment, and that, matching couldn't overcome this problem. Similarly, most school characteristics (such as percentage of Hispanic and Black students, pupil-teacher ratio and percentage of students eligible for reduced price lunch) tend to differ between treated schools and matched untreated schools in the second and third matching procedures when one-to-one matching is used. These differences persist even when the logit specification is modified and even when the t-test are carried out among smaller treated and comparison groups.

Second and third panels of Table AIII-I, on the other hand, indicate that nearest 4 and kernel matching methods yield more balanced treated and untreated groups. In particular, in almost every nearest 4 and kernel matching procedure, the propensity score as well as the region dummy, percentage of black students, total enrollment, percentage of students eligible for reduced priced lunch and pupil-to-teacher ratio are statistically the same between treated and matched untreated groups. Nevertheless, treated and matched untreated units still display dissimilarities in the 'problematic' characters (total enrollment in the first matching procedure, percentage Hispanic and percentage eligible for reduced price lunch in the second and third matching procedure).

Similar patterns to the ones described above are observed when the unrestricted (i.e. schools of all states) sample of untreated schools are used in the matching procedures. More specifically, one-to-one matching is the worst performing procedure in this case as well. In addition, the nearest 4 and kernel matching procedures still offer better matched treatment and comparison groups. Moreover, the third matching procedure that uses the nearest 4 matching method yields a perfectly balanced match (Table AIII-V) and when the second matching procedure employing the same method, treated and untreated schools are almost perfectly matched (at the 4% significance level). Finally, it is useful to note that the 'divide and modify' algorithm doesn't make the differences between matched groups disappear either.

Table AIII- I: P-values from the t-tests Between Treated and Matched Untreated Groups
One-to-one, Nearest 4 and Kernel Matching Methods Performed with Schools from Nearby States to California

Matching Method	One-to-One			Nearest 4			Kernel		
Matching Step	Step I	Step II	Step III	Step I	Step II	Step III	Step I	Step II	Step III
Matching Characteristics									
Central city dummy	0.714	0.811	0.193	0.474	0.528	0.678	0.877	0.920	0.951
Urban city dummy	0.459	0.298	0.511	0.626	0.768	0.589	0.848	0.708	0.930
Rural city dummy	0.345	0.046	0.291	0.556	0.385	0.770	0.315	0.445	0.910
Percentage Black	0.458	0.124	0.004	0.857	0.636	0.701	0.693	0.760	0.782
Percentage Hispanic	0.324	0.000	0.000	0.626	0.001	0.011	0.447	0.066	0.000
Total enrollment	0.01	0.377	0.986	0.006	0.429	0.412	0.021	0.445	0.822
Pupil-teacher ratio	0.499	0.012	0.042	0.212	0.672	0.305	0.489	0.686	0.914
Percentage free lunch	0.419	0.000	0.029	0.405	0.000	0.094	0.437	0.007	0.537
Propensity score	0.332	0.000	0.000	0.991	0.744	0.974	0.988	0.761	0.907
Sample size	158	158	158	168	160	240	171	261	258

Notes: Each cell is a p-value from a t-test which investigates whether the corresponding variable is balanced between the treated and matched untreated groups for the corresponding matching procedure. Sampling weights and jack-knife standard error estimation method are used. Step I corresponds to matching 2000 California Schools (treated schools) with 1996 California Schools (the first matching procedure). Step II refers to matching of 2000 California Schools with untreated schools from the 2000 NAEP 4th grade samples of nearby states (the second matching procedure). Finally, in Step III, treated schools are matched with 1996 untreated schools from nearby states (third matching procedure).

Table AIII- II: P-values from the t-tests Between Treated and Matched Untreated Groups
One-to-one, Nearest 4 and Kernel Matching Methods Performed with Schools from all States

Matching Method	One-to-One			Nearest 4			Kernel		
Matching Step	Step I	Step II	Step III	Step I	Step II	Step III	Step I	Step II	Step III
Matching Characteristics									
Central city dummy	0.911	0.068	0.761	0.249	0.282	0.275	0.722	0.368	0.148
Urban city dummy	0.911	0.074	0.612	0.317	0.378	0.420	0.838	0.431	0.165
Rural city dummy	0.997	0.899	0.707	0.714	0.609	0.535	0.746	0.807	0.656
Percentage Black	0.553	0.043	0.854	0.784	0.093	0.426	0.561	0.127	0.080
Percentage Hispanic	0.662	0.042	0.001	0.729	0.038	0.075	0.304	0.022	0.000
Total enrollment	0.000	0.056	0.202	0.006	0.122	0.962	0.000	0.067	0.152
Pupil-teacher ratio	0.642	0.389	0.315	0.530	0.458	0.455	0.373	0.784	0.411
Percentage free lunch	0.473	0.133	0.075	0.494	0.205	0.187	0.298	0.265	0.590
Propensity score	0.499	0.043	0.001	0.947	0.964	0.965	0.974	0.831	0.829
Sample size	158	158	158	164	218	274	177	1036	1165

Notes: Each cell is a p-value from a t-test which investigates whether the corresponding variable is balanced between the treated and matched untreated groups for the corresponding matching procedure. Sampling weights and jack-knife standard error estimation method are used. Step I corresponds to matching 2000 California Schools (treated schools) with 1996 California Schools (the first matching procedure). Step II refers to matching of 2000 California Schools with untreated schools from the 2000 NAEP 4^{th} grade samples of all states (the second matching procedure). Finally, in Step III, treated schools are matched with 1996 untreated schools from nearby states (third matching procedure).

APPENDIX IV: Testing the Quality of the Student Level Matching Methods

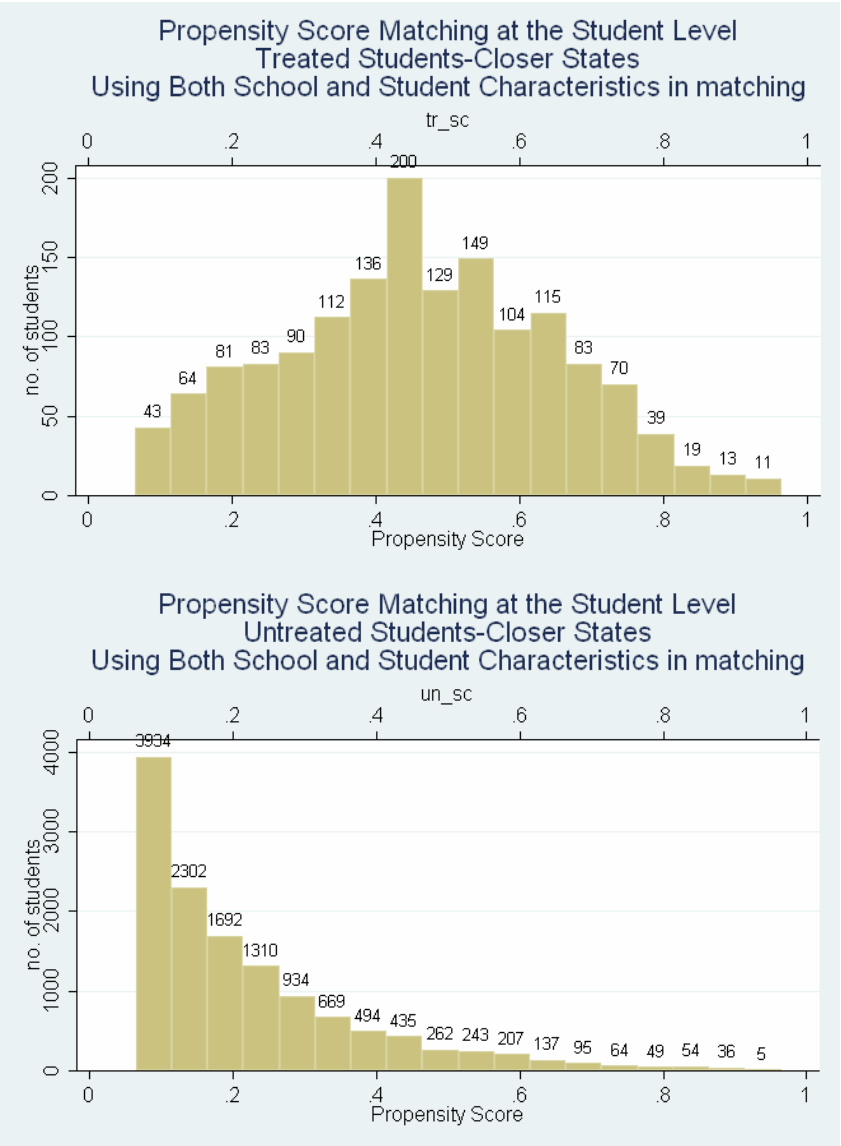
In this section, I investigate how well matching procedures at the student level perform. As mentioned before, initially I utilize both student and school characteristics when conducting the matching procedures. In Figure AIV-I, I display histograms of the propensity scores of the treated students of California and untreated students from nearby(closer)/all other states. I provide results (p-values) of the t-tests performed between the treated and matched untreated groups formed at each step of the procedures in tables Table AIV-I and AIV-II. Results indicate that when students from nearby states are used, one-to-one matching method is again the worst performing method. In steps II and III (second and third matching procedure), even the propensity scores aren't balanced between the treated and matched untreated units. Some of the student characteristics (such as race, home environment, Title1 funding and reduced price lunch eligibility) show evidence of dissimilarity. A few school characteristics differ between these groups as well. When the matching performance of the other two methods are considered, it is easy to observe that perfect matching couldn't be achieved along the same problematic characteristics. Neither adding students for all other states to the pooled untreated population nor the 'divide and modify algorithm' seemed effective in eliminating these differences.

When the new strategy of matching students by using only student characteristics is carried out, it is easy to see that this strategy increases performance of all matching procedures as measured by the similarity of the student characteristics of the treatment and comparison groups (Tables AIV-III and AIV-IV). Table AIV-III indicates that even the one-to-one matching technique produces perfectly balanced matched populations in the second and third matching procedures in terms of student characteristics. Moreover, in almost all of the matching procedures which use students from only nearby states to California, this strategy produces treatment and comparison groups that are strikingly balanced along school characteristics as well as along student attributes.

At first glance, this may seem to be an unexpected outcome since school attributes aren't taken into account during matching. On the other hand, note that only students from nearby (closer) states to California are used in this part of the analysis as the untreated group. Therefore one may think that similarities between California and closer states' schools may be partially responsible for this outcome. Actually, a closer inspection of Tables AIV-III reveals that when only student attributes are controlled for in the matching, the variable *student-to-teacher ratio* ends up being significantly different between

treatment and matched untreated groups in the last two matching procedures. Therefore, when controlled for in the matching process, this variable could be the reason why other characteristics turned out to be not balanced between treatment and comparison groups.

When students from all states are included as untreated students in the matching procedures and when only student attributes are controlled for in the matching, matched student pairs are more similar in terms of student attributes (Table AIV-IV). Features of the matched pairs' schools, however, are no longer well matched. This is not an unexpected result because now California students are matched with students from all other states. The untreated students therefore come from a larger pool of school that may differ substantially from California. Therefore when school characteristics are not controlled for in the matching, the matched treatment and comparison groups have fewer tendencies to display similarities along school attributes.



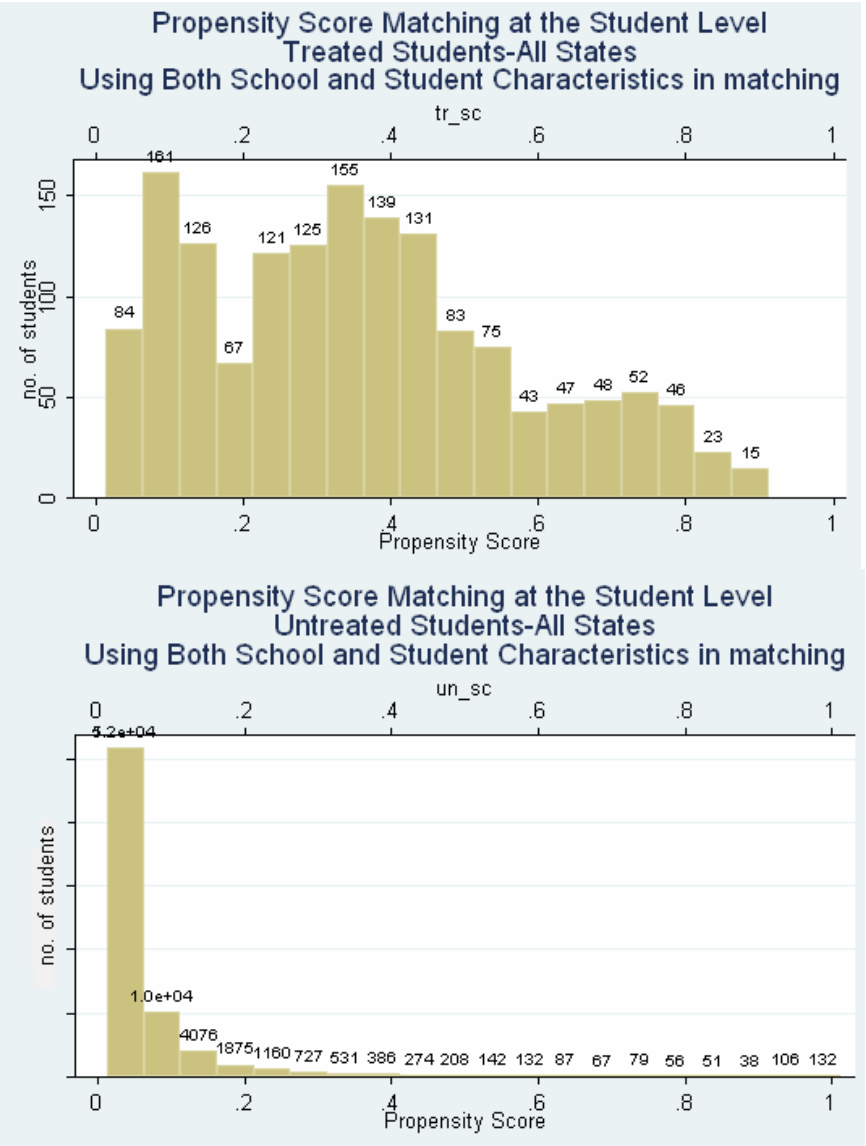


Figure AIV-I: Histogram of the propensity scores of the treated and untreated students. Nearby (closer) states to California and all other states are used and matching is performed at the student level using both student and school characteristics.

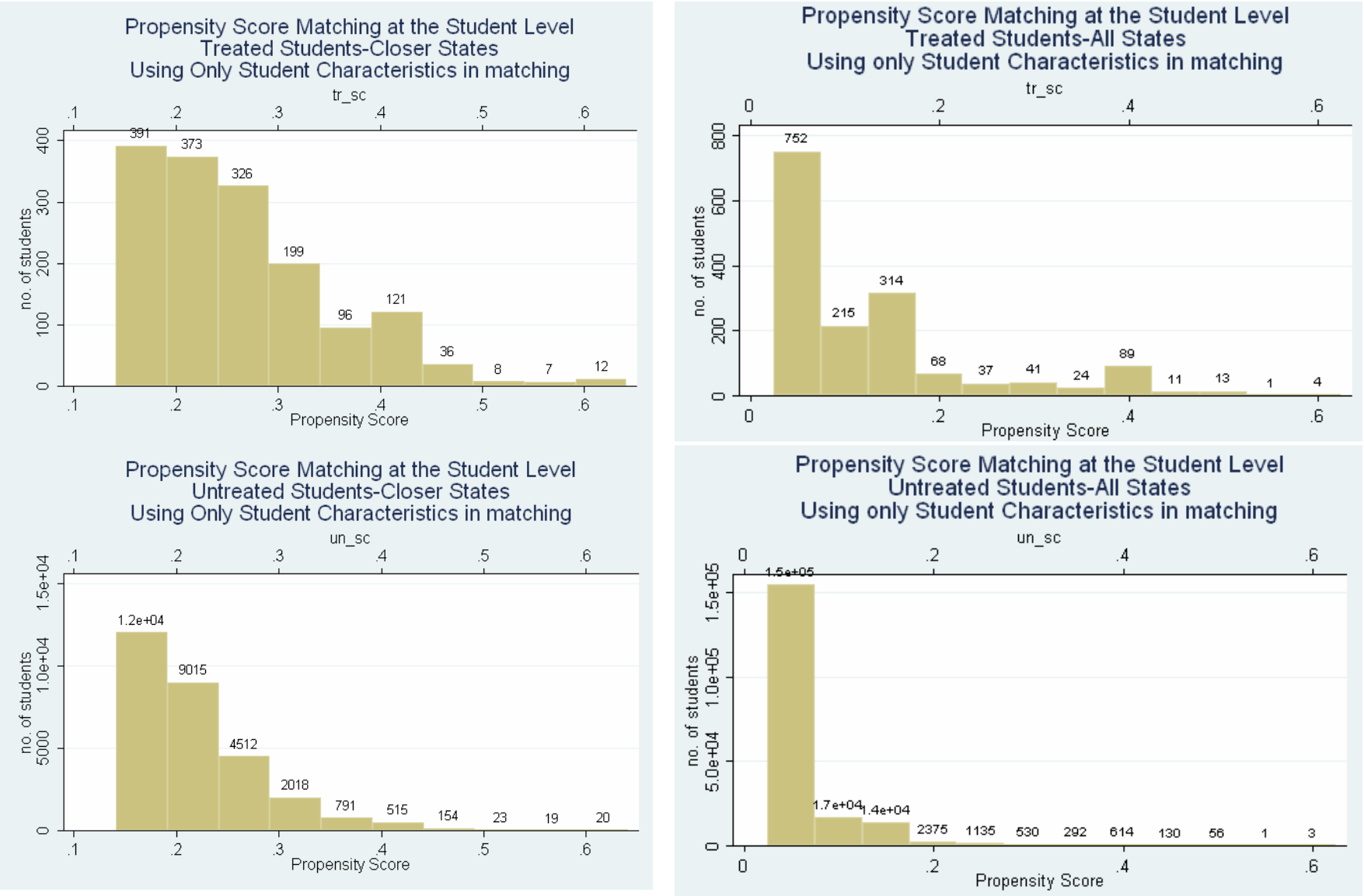


Figure AIII-II: Histogram of the Propensity scores of the treated and untreated students. Nearby (closer) states to California and all other states are used and matching is performed at the student level using only student characteristics.

Table AIV- I: P-values from the t-tests Between Treated and Matched Untreated Groups-Nearby States
Matching Methods Performed at the Student Level Using Both Student and School Characteristics

Matching Method	One-to-One			Nearest 4			Kernel		
Matching Step	Step I	Step II	Step III	Step I	Step II	Step III	Step I	Step II	Step III
Matching Characteristics									
Student:									
Female dummy	0.840	1.000	0.011	0.895	0.041	0.034	0.696	0.000	0.001
Black dummy	0.608	0.176	0.045	0.374	0.495	0.851	0.741	0.400	0.520
Hispanic dummy	0.051	0.000	0.000	0.105	0.000	0.000	0.059	0.000	0.000
Asian dummy	0.520	0.000	0.001	0.355	0.355	0.011	0.508	0.002	0.100
Amer. Indian dummy	0.525	0.248	0.029	0.339	0.088	0.139	0.341	0.196	0.395
Ind. educ. plan dummy	0.010	0.107	0.459	0.032	0.328	0.097	0.008	0.044	0.335
Lim. English dummy	0.001	0.000	0.000	0.001	0.000	0.003	0.001	0.003	0.000
Title 1 funds eligib. d.	0.008	0.218	0.000	0.032	0.000	0.753	0.010	0.985	0.000
Red. price lunch d.	0.001	0.000	0.000	0.001	0.000	0.003	0.001	0.003	0.000
Home env. cat 2 d.	0.550	0.000	0.000	0.543	0.227	0.251	0.377	0.093	0.082
Home env. cat 3 d.	0.007	0.923	0.004	0.097	0.842	0.065	0.015	0.102	0.707
School:									
Large city dummy	0.274	0.200	0.003	0.384	0.784	0.894	0.498	0.848	0.209
Urban city dummy	0.016	0.339	0.168	0.151	0.678	0.020	0.012	0.062	0.793
Rural area dummy	0.340	0.167	0.001	0.408	0.902	0.917	0.432	0.622	0.830
Pupil-teacher ratio	0.533	0.755	0.002	0.607	0.946	0.659	0.566	0.969	0.868
Percentage Black	0.708	0.081	0.007	0.814	0.459	0.954	0.650	0.728	0.481
Percentage Hispanic	0.995	0.000	0.000	0.661	0.000	0.000	0.975	0.000	0.000
Total enrollment	0.440	0.045	0.182	0.307	0.134	0.004	0.499	0.020	0.120
Propensity score	0.592	0.000	0.000	1.000	0.969	0.936	0.976	0.859	0.917
Sample size	3082	3082	3082	3077	2917	2909	3523	7048	6974

Notes: Each cell is the p-value from a t-test which investigates whether the corresponding variable is balanced between the treated groups (California 2000 students) with the matched untreated group for the corresponding matching procedure. Sampling weights and jack-knife method are used. Home environment tabulates how many times the student answers the following questions "Yes": "Does you family get a newspaper regularly?", "Is there an encyclopedia in your home?", "More than 25 books?" and "Get any magazines regularly?" Reduced price lunch d. label refers to the variable that denotes whether the student is eligible for reduced price lunch.

Table AIV- II: P-values from the t-tests Between Treated and Matched Untreated Groups-Nearby States
Matching Methods Performed at the Student Level Using only Student Characteristics

Matching Method	One-to-One			Nearest 4			Kernel		
Matching Step	Step I	Step II	Step III	Step I	Step II	Step III	Step I	Step II	Step III
Matching Characteristics									
Student:									
Female dummy	0.979	0.997	0.812	0.957	0.951	0.992	0.645	0.965	0.482
Black dummy	0.651	0.926	0.890	0.941	0.882	0.955	0.750	0.120	0.031
Hispanic dummy	0.268	0.772	0.771	0.919	0.908	0.954	0.770	0.070	0.491
Asian dummy	0.016	0.732	0.733	0.407	0.800	0.684	0.017	0.623	0.387
Amer. Indian dummy	0.147	0.978	0.758	0.143	0.987	0.836	0.052	0.020	0.031
Ind. educ. plan dummy	0.013	0.705	0.728	0.002	0.679	0.477	0.008	0.047	0.074
Lim. English dummy	0.179	0.729	0.999	0.766	0.773	0.834	0.301	0.158	0.471
Title 1 funds eligib. d.	0.090	0.947	0.771	0.557	0.928	0.813	0.107	0.997	0.171
Red. price lunch d.	0.179	0.729	0.999	0.766	0.773	0.834	0.301	0.158	0.471
Home env. cat 2 d.	0.486	0.783	0.896	0.922	0.851	0.831	0.715	0.294	0.161
Home env. cat 3 d.	0.667	0.634	0.833	0.538	0.694	0.815	0.423	0.012	0.520
School:									
Large city dummy	*0.366*	*0.967*	*0.876*	*0.803*	*0.979*	*0.827*	*0.830*	*0.232*	*0.882*
Urban city dummy	*0.796*	*0.518*	*0.887*	*0.578*	*0.570*	*0.935*	*0.214*	*0.009*	*0.415*
Rural area dummy	*0.180*	*0.000*	*0.084*	*0.393*	*0.037*	*0.105*	*0.174*	*0.009*	*0.089*
Pupil-teacher ratio	*0.117*	*0.000*	*0.000*	*0.402*	*0.000*	*0.000*	*0.138*	*0.000*	*0.000*
Percentage Black	*0.768*	*0.224*	*0.346*	*0.973*	*0.132*	*0.374*	*0.759*	*0.102*	*0.091*
Percentage Hispanic	*0.297*	*0.003*	*0.213*	*0.059*	*0.128*	*0.629*	*0.690*	*0.001*	*0.002*
Total enrollment	*0.655*	*0.804*	*0.303*	*0.520*	*0.426*	*0.108*	*0.423*	*0.245*	*0.631*
Propensity score	0.698	0.887	0.834	0.997	0.997	0.992	0.837	0.753	0.636
Sample size	3138	3138	3138	2286	2418	2409	3836	14043	15936

Notes: Each cell is the p-value from a t-test which investigates whether the corresponding variable is balanced between the treated groups (California 2000 students) with the matched untreated group for the corresponding matching procedure. Sampling weights and jack-knife method are used. Home environment tabulates how many times the student answers the following questions "Yes": "Does you family get a newspaper regularly?", "Is there an encyclopedia in your home?", "More than 25 books?" and "Get any magazines regularly?" Reduced price lunch d. label refers to the variable that denotes whether the student is eligible for reduced price lunch.

Table AIV- III: P-values from the t-tests Between Treated and Matched Untreated Groups-All States
Matching Methods Performed at the Student Level Using both Student and School Characteristics

Matching Method	One-to-One			Nearest 4			Kernel		
Matching Step	Step I	Step II	Step III	Step I	Step II	Step III	Step I	Step II	Step III
Matching Characteristics									
Student:									
Female dummy	0.593	0.310	0.252	0.819	0.540	0.938	0.709	0.230	0.839
Black dummy	0.851	0.063	0.095	0.836	0.117	0.121	0.689	0.192	0.082
Hispanic dummy	0.245	0.001	0.000	0.415	0.000	0.025	0.165	0.000	0.001
Asian dummy	0.885	0.465	0.657	0.554	0.603	0.872	0.853	0.917	0.625
Amer. Indian dummy	0.682	0.014	0.009	0.985	0.001	0.171	0.454	0.005	0.250
Ind. educ. plan dummy	0.008	0.074	0.500	0.044	0.008	0.710	0.005	0.012	0.961
Lim. English dummy	0.001	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000
Title 1 funds eligib. d.	0.010	0.170	0.002	0.006	0.087	0.105	0.015	0.065	0.076
Red. price lunch d.	0.001	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000
Home env. cat 2 d.	0.820	0.310	0.190	0.726	0.084	0.664	0.713	0.262	0.971
Home env. cat 3 d.	0.022	0.043	0.212	0.030	0.104	0.387	0.010	0.025	0.902
School:									
Large city dummy	0.319	0.594	0.081	0.362	0.635	0.628	0.446	0.253	0.819
Urban city dummy	0.038	0.020	0.913	0.051	0.078	0.423	0.010	0.047	0.978
Rural area dummy	0.148	0.235	0.124	0.244	0.081	0.161	0.264	0.135	0.161
Pupil-teacher ratio	0.249	0.010	0.251	0.400	0.022	0.254	0.339	0.008	0.257
Percentage Black	0.555	0.021	0.112	0.598	0.052	0.140	0.525	0.088	0.089
Percentage Hispanic	0.556	0.000	0.000	0.392	0.000	0.000	0.500	0.000	0.000
Total enrollment	0.343	0.063	0.005	0.294	0.004	0.054	0.305	0.031	0.062
Propensity score	0.867	0.046	0.001	0.999	0.996	1.000	0.982	0.928	0.929
Sample size	3082	3082	3082	3129	3519	3328	3532	34168	38812

Notes: Each cell is the p-value from a t-test which investigates whether the corresponding variable is balanced between the treated groups (California 2000 students) with the matched untreated group for the corresponding matching procedure. Sampling weights and jack-knife method are used. Home environment tabulates how many times the student answers the following questions "Yes": "Does you family get a newspaper regularly?", "Is there an encyclopedia in your home?", "More than 25 books?" and "Get any magazines regularly?" Reduced price lunch d. label refers to the variable that denotes whether the student is eligible for reduced price lunch.

Table AIV- IV: P-values from the t-tests Between Treated and Matched Untreated Groups-All States
Matching Methods Performed at the Student Level Using Only Student Characteristics

Matching Method	One-to-One			Nearest 4			Kernel		
Matching Step	Step I	Step II	Step III	Step I	Step II	Step III	Step I	Step II	Step III
Matching Characteristics									
Student:									
Female dummy	0.376	0.935	0.838	0.976	0.938	0.902	0.494	0.844	0.703
Black dummy	0.752	1.000	0.933	0.947	0.963	0.936	0.423	0.403	0.703
Hispanic dummy	0.751	0.925	0.971	0.963	0.961	0.988	0.621	0.005	0.007
Asian dummy	0.167	0.757	0.782	0.452	0.784	0.853	0.146	0.001	0.000
Amer. Indian dummy	0.249	0.874	0.915	0.067	0.919	0.875	0.129	0.874	0.882
Ind. educ. plan dummy	0.002	0.926	0.918	0.002	0.958	0.779	0.005	0.048	0.038
Lim. English dummy	0.348	1.000	0.978	0.801	0.945	0.955	0.829	0.263	0.512
Title 1 funds eligib. d.	0.159	0.867	0.879	0.993	0.872	0.887	0.042	0.926	0.003
Red. price lunch d.	0.348	1.000	0.978	0.801	0.945	0.955	0.829	0.263	0.512
Home env. cat 2 d.	0.891	0.817	0.985	0.949	0.894	0.976	0.622	0.675	0.495
Home env. cat 3 d.	0.806	0.912	0.999	0.777	0.859	0.929	0.108	0.283	0.162
School:									
Large city dummy	*0.825*	*0.975*	*0.957*	*0.856*	*0.950*	*0.938*	*0.839*	*0.594*	*0.847*
Urban city dummy	*0.900*	*0.864*	*0.956*	*0.549*	*0.864*	*0.979*	*0.027*	*0.298*	*0.066*
Rural area dummy	*0.155*	*0.494*	*0.186*	*0.173*	*0.143*	*0.271*	*0.172*	*0.117*	*0.477*
Pupil-teacher ratio	*0.087*	*0.000*	*0.000*	*0.195*	*0.000*	*0.000*	*0.129*	*0.000*	*0.000*
Percentage Black	*0.772*	*0.001*	*0.000*	*0.726*	*0.022*	*0.000*	*0.597*	*0.000*	*0.001*
Percentage Hispanic	*0.195*	*0.000*	*0.000*	*0.079*	*0.000*	*0.000*	*0.606*	*0.000*	*0.000*
Total enrollment	*0.560*	*0.003*	*0.509*	*0.325*	*0.000*	*0.201*	*0.450*	*0.009*	*0.488*
Propensity score	0.594	0.976	0.977	0.964	0.989	0.992	*0.831*	*0.726*	*0.605*
Sample size	3138	3138	3138	2223	2505	2498	0.494	0.844	0.703

Notes: Each cell is the p-value from a t-test which investigates whether the corresponding variable is balanced between the treated groups (California 2000 students) with the matched untreated group for the corresponding matching procedure. Sampling weights and jack-knife method are used. Home environment tabulates how many times the student answers the following questions "Yes": "Does you family get a newspaper regularly?", "Is there an encyclopedia in your home?", "More than 25 books?" and "Get any magazines regularly?" Reduced price lunch d. label refers to the variable that denotes whether the student is eligible for reduced price lunch.