The Causal Effect of School Reform: Evidence from California's Quality Education Investment Act*

Paul Burkander†

November 1, 2013

Abstract

Beginning in the 2007-08 school year, California's Quality Education Investment Act required schools selected via lottery to institute reforms including class size reduction, increased average teacher experience, and extra professional training. The act provided additional per-pupil funding for schools to meet these requirements. Conditional on known probabilities of selection, which differed across schools, treatment is uncorrelated with potential outcomes, allowing for non-parametric identification of the causal effect by inverse probability weighting. In the first fully-funded year of the program, math scores in 4^{th} grade increased by 0.32 SD in the population of California school-grade averages, and by the second fully-funded year 5^{th} grade math scores improved by 0.36 SD. By the third fully-funded year of the program, math scores in 2^{nd} grade were 0.28 SD higher in the distribution of California school-grade averages, and 0.27 SD higher in 3^{rd} grade. Selected schools did not increase teacher experience, and had 4 to 4.4 fewer students in the first fully-funded year in 4^{th} and 5^{th} grade. In kindergarten through 3^{rd} grade class sizes were reduced later and less dramatically, by 3 to 4 students by the third fully-funded year, due primarily to unselected schools exiting California's previous class size reduction program. The timing of class size reductions and student achievement gains suggests class size was the driving factor.

JEL Classifications: H75, I21, J45.

Keywords: Class Size, Teacher Labor Markets, Education Production Function.

*I am grateful for guidance provided by Gary Solon, Todd Elder, Michael Conlin, and participants at the University of Michigan Causal Inference in Education Research Seminar. I am also grateful to Marsha Porte and Jim Alford at the California Department of Education for their invaluable assistance. The work reported here was supported in part by a Pre-Doctoral Training Grant from the Institute of Education Sciences, U.S. Department of Education (Award # R305B090011) to Michigan State University. The opinions expressed here are mine and do not represent the views of the Institute or the U.S. Department of Education.

†Department of Economics, Michigan State University, 110 Marshall Adams Hall, East Lansing, MI 48823. burkande@msu.edu

1 Introduction

Current educational policy in the United States is focused on increasing the proportion of students who meet state-determined proficiency levels on standardized tests. There is disagreement about how to achieve this, with some arguing for additional educational inputs, and others for more efficient use of existing inputs. An extensive literature on educational production functions[1] has attempted to resolve this and related questions. However, despite some random experiments and a plethora of natural experiments, no clear consensus has emerged on the question of whether marginal changes in educational resources have any effect on educational outcomes.

California's Quality Education Investment Act (QEIA) offers a unique opportunity to identify the effect of increased inputs on outcomes. In the 2007-2008[2] school year the QEIA went into effect, leading to increased funding and obligatory reforms for about 500 selected schools, which were chosen from 1,260 participating schools. Districts were first required to rank all of their participating schools, and then districts were randomly selected to have their highest ranked schools funded. Once selected, funded schools were required to institute several reforms, e.g., they had to reduce average class size, increase average teacher experience, and provide additional professional training to teachers.

Conditional on districts' rankings, selection of schools was random, though schools differed in their probability of selection. The selection process, and therefore the probabilities of selection, are known, and the average treatment effect of QEIA can therefore be non-parametrically identified using inverse probability weighting (IPW). A drawback of QEIA is that the effects of the individual reforms cannot be separately identified. However, bundled reforms are worth studying in their own right: pressure to improve outcomes often leads to concurrent policy changes so QEIA is reflective of how reforms are actually carried out;

[1] Summaries of the assumptions and methods employed in the education production function literature can be found in Hanushek (1979), Todd and Wolpin (2003), and Rice and Schwartz (2008).

[2] Henceforth, school years are referred to by the year in which the Spring semester occurs. For example, the 2007-2008 school year is referred to as 2008.

it may also be the case that interactive effects cause bundled reforms to be more or less effective than the sum of their constituent parts.

Moreover, as I find, QEIA caused a reduction in class size of about 4 students per class by the third fully-funded year of the program, but had no discernible effect on the other main policy lever that I observe, teacher experience. Reportedly, the vast majority of elementary schools eligible to participate in QEIA were already required to meet many of its requirements, with the exception of reduced class size, increased teacher experience, and increased professional training. Also, continued participation in QEIA was contingent on schools meeting achievement growth targets; the evidence therefore suggests that the causal effect of QEIA on standardized test scores occurred through some mix of class size reduction, professional training, and increased pressure to raise test scores.

Indeed, QEIA did cause a statistically significant increase in student achievement, as measured by both California's Average Performance Index, and by grade-level results on California's primary standardized test. The API is a weighted school-level average across all tested subjects, grades, and test types. With respect to the population of all elementary schools, the average treatment effect of QEIA on the API by the third fully-funded year of the program was an increase of 0.33 standard deviations, with larger gains for Hispanic and low-SES students. With respect to the population of grade-level averages across all California schools, by the third fully-funded year of the program standardized math scores increased 0.28 standard deviations in 2^{nd} grade, and by 0.44 standard deviations in 5^{th} grade. QEIA caused more modest gains in English language arts, of 0.19 and 0.22 standard deviations in 2^{nd} and 5^{th} grade, respectively.

In what follows, section 2 reviews the relevant literature; section 3 describes QEIA in greater detail; section 4 reviews the data used in this analysis; section 5 outlines the identification strategy; section 6 presents the results. Section 7 concludes.

2 Literature Review

This analysis contributes causal evidence to the aforementioned extensive literature on education production functions, which generally has found mixed results. Meta-analyses that find no clear evidence of an effect of increased school inputs on student outcomes include Hanushek (1986), and Hanushek (1997), though the methods employed in those analyses are criticized by Krueger (2002). In contrast, Greenwald et al. (1996) provide a meta-analysis that finds many school inputs do have positive effects, though their methods are criticized by Hanushek (1996).

Within the education production function literature, this paper contributes most to those strands concerned with the effect of reducing class size, providing professional training to teachers, and increasing accountability. The study of class size effects on student achievement has a rich history, dating back at least a century. As noted by Rockoff (2009), early waves of the literature, which include field experiments as early as the 1920s, tended to find no effect from a reduction of class size.

Recent studies of variation in class size tend to be quasi-experimental, with the notable exception of Tennessee's Project STAR (Student/Teacher Achievement Ratio). Project STAR was a randomized control trial that assigned students to either small classes (13-17 students per class), regular classes (22-26 students per class), or regular classes with a teaching aide. The reduced class size treatment of Project STAR has generally been found to have had positive effects in the short run (Nye et al., 1999, Krueger, 1999), and in longer run outcomes (Krueger and Whitmore, 2001, Chetty et al., 2011a) though non-random attrition, lack of baseline measure of student performance, and little information about teachers and how they were randomized should give us pause in interpreting results (Hanushek, 1999).

Notable natural experiments include Angrist and Lavy (1999), which uses a regression discontinuity design based on a class size rule in Israel. They find significant returns to achievement from class size reduction for math and reading scores for 5th graders. Hoxby (2000) also uses variation in class size generated by class size caps, and exploits exogenous

variation in population, to analyze the effect of class size in Connecticut. She finds no returns to class size reduction, and in fact rules out even modest returns to class size reduction.

The results in Hoxby (2000) are questioned by Jepsen and Rivkin (2009), who note that, in using test scores from the following year, estimates may be attenuated. Jepsen and Rivkin (2009) analyze a previous class size reduction program in California, which was first implemented in 1996. Using a fixed-effects analysis, and with a school-level measure of achievement as the outcome variable, the authors find that a ten-student reduction in class size led to a 0.06 to 0.1 standard deviation improvement in Math, and a 0.04 to 0.6 standard deviation improvement in Reading. An important contribution by Jepsen and Rivkin (2009) is that, unlike previous class size analyses, they explore changes in teacher quality that result from rapidly reducing class size, finding that in the early years of the program class size returns were offset by losses from reduced teacher quality. This issue is explored further by Dieterle (2013), who finds that the reduction in teacher quality was not large enough to account for only modest returns to reduced class size reduction in an anonymous state. Chingos (2012) examines another class size reduction policy implemented in Florida, and finds no effect using a comparative interrupted time series design. Chingos (2012) exploits the fact that many districts already met the class size requirements of Florida's law when it was implemented. Districts that already met the requirement received the same increase in funding, so the counterfactual to increased funding and class size reduction is an unconstrained increase in funding.

The literature on effects of teacher professional development is less-well developed. Yoon et al. (2007) reviews over 1,300 studies conducted between 1986 and 2006, and found only nine to rigorously examine the effect of teacher professional development on student achievement. Among these, the range of effects was quite large, from -0.53 to 2.39 standard deviations,[3] with the smallest and largest effect coming from the same study. The student assessment tools were generally closely aligned with teacher training, and within study there was wide

[3]Presumably the standard deviations are with respect to the population of students studied, though neither Yoon et al. (2007) nor the source material clarifies the point.

variation of effects. In a more recent study, Barrett et al. (2012) uses a propensity score model to test whether less effective teachers are more likely to select into professional development, and whether accounting for this selection affects estimates of effectiveness of such programs. They find that pre-treatment value added scores are an important predictor of participation, and controlling for selection professional development increases student test scores by 0.08 standard deviations in elementary school.

Linking incentives to test scores has been shown to improve medium term math outcomes (Chiang, 2009), and long-term outcomes of low-performing students (Cohodes et al., 2013). A growing body of literature considers potential erosion of signal quality of student assessments when those assessments are linked to incentives. There is evidence that schools manipulate the population of test takers (Figlio and Getzler, 2006, Jacob, 2005), shift resources towards marginal students (Neal and Schanzenbach, 2010), and that teachers manipulate test results (Jacob and Levitt, 2003). The QEIA link between incentives and test scores differs from those studied above in at least two ways: using the API, an average of scores across all students, instead of percent proficient removes the incentive to teach to the marginal student, and scores on tests for cognitively impaired students count toward a school's API. I nonetheless test below whether the population of test takers changes.

With conflicting results in meta-analyses and natural experiments, and few randomized control trials, it seems clear that after a century of research into education production functions, more research is needed. States such as California in 1996, Florida in 2003, and Ohio in 2009 have passed class size reduction laws, devoting resources toward increasing inputs that may or may not improve outcomes. If increased inputs can lead to improved output, it must be determined how to move closer to an optimal mix of inputs. To address these questions, more natural experiments with credible exogenous variation are needed. QEIA provides such credible evidence.

3 Policy Description

The QEIA was preceded in California by a larger and more ambitious class size reduction policy. That policy was enacted in 1996, and continues nominally to this day. Participation was voluntary, but incentivized: districts received $650 in the first year[4] per student in a K-3 class of 20 or fewer students. However, this incentive has diminished twice over time. In 2004 the maximum qualifying average class size was increased to just under 22, and as of February 2009 classes of 25 or more students are still eligible for 70% of the per-pupil funds, though funding is given for no more than 20 students.[5]

The QEIA came about as the consequence of litigation against then California Governor Arnold Schwarzenegger. The plaintiffs in the case argued successfully that the state paid less than the legislated minimum amount to kindergarten through 12th grade public schools in the 2005[6] and 2006 school years. As a result, the state was required to pay back approximately $2.7 billion to K-12 schools.

Rather than distribute the money equally across all schools, legislators decided to focus on a subset of low-performing schools. The subset was chosen on a semi-random basis, and the number of schools was chosen such that per-student funding would increase by $500 in grades K-3,[7] $900 in grades 4-8, and $1,000 in high school from 2009-2014, and by half as much in 2008.[8]

Schools were deemed eligible to participate in QEIA if they were in the bottom quintile of the state's 2005 academic performance distribution, as determined by the API.[9] Eligible schools had to commit to meeting the requirements of QEIA before they could participate

[4]This number was adjusted for inflation in subsequent years.

[5]For instance, a class of 25 or more students would receive $0.70 \times 20 \times$ Full Per-Pupil Amount

[6]Governor Schwarzenegger had reached an agreement with a coalition including the California Teachers Association to underfund education by $2 billion below the amount guaranteed by Proposition 98, which pegs education funding to growth in general funds. However, state revenue exceeded expectations, and education funding was not updated to reflect this. For more information, see Bluth (2005).

[7]For comparison, in the first full year of QEIA funding the per-pupil funding for participation in California's existing class size reduction program was $1,071.

[8]The reduced amount in 2008 was intended to give schools a chance to prepare for full implementation of reforms by 2009.

[9]Very small schools, whose API scores were considered unreliable, were excluded.

in the selection process. Schools could choose to participate in the regular QEIA program, or an alternative program. Schools in the alternative program, which are excluded from this analysis, were able to design their own reform plans, which had to be approved as part of the application to participate in QEIA. Of the 1,455 schools eligible to participate in QEIA, 1,260 chose to do so, and 88 of these chose to participate in the alternative program.

Each district with more than one participating school was required to rank its schools. It was permissible to give multiple schools the same rank, and indeed several districts did so. Districts received as many random numbers as they had participating schools, and these random numbers were assigned to each district's schools based on the district's rankings. For example, if a district with two schools received random numbers 213 and 314, the highest ranked school was assigned 213 and the second was assigned 314. If a district assigned the same ranking to multiple schools, the order was determined randomly within that ranking, and was done so by the California Department of Education. The selection then proceeded in four stages.

First, schools for the alternative program were selected. High schools were given priority for this program, and the number of schools was chosen such that no more than 15% of the anticipated number of students in funded schools would be in the alternative program. The high schools with the lowest random numbers in the alternative program were funded until this target was reached.[10]

Second, to ensure geographic diversity, in each county without a funded school from the first stage, the school with the lowest random number was selected. Districts were told that after schools were selected for the alternative program and geographic diversity, all schools with the lowest random numbers would be funded until funds were exhausted.[11]

In fact, high schools were selected separately in the third stage: to ensure the legislatively

[10]Several middle and elementary schools applied for the alternative program, but given the number of high schools that applied they effectively had zero probability of being chosen.

[11]The actual selection differed somewhat, as described below. That districts were told this simplified version is evidenced in contemporaneous school board minutes (Santa Rosa City Schools, 2007), and CDE presentations (Balcom, 2007). This is also the depiction in the report to CA legislature (CDE, 2010), written 3 years after the selection process.

mandated fair representation of grade spans, a target number of high school students was selected so that the proportion of high school students in funded schools would be roughly equivalent to the proportion of high school students in all participating schools. The high schools with the lowest random numbers were selected until this target was reached. Any school with at least one high school student in 2007 was considered a high school for this purpose. Finally, the elementary and middle schools with the lowest random numbers were selected until QEIA funds were exhausted.[12]

At the conclusion of the selection process, 25 schools had been selected for the alternative program and 463 for the regular program. One school that was selected immediately withdrew from the program, and in subsequent years 13 schools were added. For the purpose of this analysis, I consider all schools initially selected to be treated, and all participating schools not selected to be the control group. Additionally, I restrict the sample to elementary schools,[13] which account for over 70% of schools participating in the regular QEIA program.

Funded schools were required to implement the following: reduce class size; align average teacher experience with their district average; ensure that all teachers in the school be considered Highly Qualified Teachers (HQT) under California's Elementary and Secondary Education Act; satisfy the requirements of the *Williams* settlement, which required schools to provide qualified teachers and safe, well-maintained facilities; provide professional training to teachers and paraprofessionals; and, for high schools, increase the counselor-student ratio.

According to CDE (2010), the vast majority of schools eligible to participate in QEIA were already required to meet the HQT standard and the requirements of the *Williams* settlement, regardless of whether they were selected to be funded. This claim is substantiated by Table

[12]As a result of this process, and unbeknownst to districts prior to selection, the funding results did not always follow district rankings. For instance, a highly ranked high school could go unfunded, and a lower ranked elementary school could be funded. Ranking a high school ahead of an elementary school could also lead to both not being funded, while if the elementary were ranked higher it would be funded, if the difference in random numbers is sufficiently large.

[13]This restriction has two motivations: there are additional QEIA requirements for high schools, complicating the interpretation of the treatment, and beginning in 6^{th} grade, students are sorted into various math examinations, thus compositional changes may be confounded with changes in achievement. Results that include middle and high schools are qualitatively quite similar, and are available in the online appendix.

1, which shows that the typical participating elementary school had 94% of its teachers classified as Highly Qualified, and 95% of participating schools were required to satisfy the terms of the *Williams* settlement.

The class size reduction requirement stipulated that funded schools reduce class size to 20 students per class in grades K-3.[14] In grades 4-12, class sizes had to be reduced from their baseline level[15] by 5 students, or to 25 students per class, whichever was lower. In each of the first three years of QEIA, schools were required to reduce the difference between the pre-QEIA average class size and QEIA target class size by 1/3. For some schools, the average in 2007 was quite low, which was particularly strenuous for small schools with a single classroom per grade. As such, many schools applied for and were granted waivers from this requirement, and instead met a higher minimum class size requirement.

Under QEIA, teacher experience is measured by the Teacher Experience Index (TEI). Teachers with more than 10 years of experience are assigned 10 years in calculating the average. Part-time teachers are given full weight in the calculation, and teachers teaching at multiple schools count towards each school's average. Funded schools are required to exceed the district average TEI.

Districts selected for QEIA are required to provide professional development opportunities for teachers, administrators, and paraprofessionals, e.g., teaching assistants. Funded schools are required to build and maintain a system for tracking participation in professional development programs, and districts are required to ensure that funded schools are in fact meeting the requirements. Participation requirements for teachers are clearly spelled out by QEIA, e.g., each year at least one third of teachers in a QEIA funded school must participate in training, but the specifics of the training program are largely left to the schools and districts.

In addition to these reforms, participation in QEIA was contingent on meeting accelerated

[14]This is precisely the original requirement of California's 1996 class size reduction policy, the maximum cap of which increased over time.

[15]The baseline was the grade-level average class size in 2006, unless that average was greater than 25, in which case 2007 was used

student achievement growth targets, as measured by California's API. The target API for all schools in California is 800; all California schools below that target have a growth target of 5% of the difference between their API and 800, or 1 point, whichever is greater. By the third year of QEIA, funded schools are required to have exceeded growth targets on average over those first three years. A school is permitted to fall short of its growth target in the first two years of full funding without repercussions, then after the third fully-funded year schools whose average growth did not exceed average growth targets lost QEIA funding.

4 Data

This analysis relies on several publicly available data sets produced by the California Department of Education. These include school-level data on API scores, a data set that includes a rich set of school demographics; teacher-level data; assignment-level data, including, e.g., the number of classes assigned to a teacher and the number of students in each of those classes; and subject-grade-level data on California standardized tests. Though these data sets are available for earlier years as well, I rely primarily on data from 2005-2011 with one exception: the assignment-level data were not collected in 2010 due to budget constraints. As a result, I am unable to calculate average class size, proportion of teachers classified as HQT, or TEI for 2010. In addition to these publicly available data, I have obtained from CDE the rankings of participating schools submitted by each district, which include a variable for whether the school applied for the regular or alternative program.

The teacher-level data are not linked from year to year. Instead, in each year teachers are assigned a new ID. The purpose of the ID is to facilitate linking teacher-level data to assignment-level data. The teacher-level data do however contain a number of teacher characteristics, such as years of teaching experience, years teaching in the same district, ethnicity, gender and education. I use these characteristics to link teachers across years within a school. If multiple teachers at a school are observationally similar, I randomly link

them across years.

As a result, I do not reliably observe duration of employment spells at a school. Similarly, if a teacher leaves, and in the following year a new observationally identical teacher enters the school, I do not observe a change in the composition of teachers. The data can however be used to reliably determine net changes in the characteristics of the teacher workforce at a school. I use these data to measure average teacher experience, the proportion of teachers new to a school, and the proportion of teachers new to a school who are either new to teaching, or experienced but new to the district.

For my measure of class size, I restrict the set of classes to math, English, science, and self-contained classes. Self-contained classes are those in which subjects such as math and English are taught by the same teacher, and are the most common class type in elementary schools. This analysis excludes special education courses, vocational courses, and other electives.[16]

It has become common in the education production function literature to use student performance on standardized tests as a measure of the output of this production process. Standardized tests surely fail to capture a number of cognitive and non-cognitive skills that an educational system is expected to impart on students. However, there is evidence that variation in school inputs that increase test scores also have a positive impact on a number of later-life outcomes, such as probability of attending college, selectivity of college, and income (Chetty et al., 2011a, Chetty et al., 2011b). Often a student's performance on standardized tests is the outcome in a regression including measures of scholastic inputs and the student's performance in previous years as controls. The use of California's API in this analysis is similar, but it differs from student level assessment scores in important ways.

Notably, the API is an average of performance not just across students, but across subjects and even test types. For instance, an elementary school in 2010 would have administered

[16]Teachers for these excluded classes are included in the teacher experience category, in part so my of experience is not dependent on data missing in 2010. The TEI is based on a subset of classes similar to that which I use to calculate average class size.

an English and language arts test in grades 2-5, a math test in grades 2-5, and a science exam in grade 5. Additionally, two alternative exams, the California Modified Assessment and the California Alternative Performance Assessment would have been administered to students with varying degrees of cognitive impairment. The API for that school is a weighted average across all these tests, subjects, and grades.[17] Nonetheless, the API is California's primary tool for assessing academic performance, and the goal of QEIA was to improve API scores, so I include it in my analysis. I standardize API scores within years with respect to the distribution of all elementary school APIs.

I supplement this measure of student achievement with grade-subject-level data on California's primary standardized test, the eponymous CST. California makes publicly available the mean scaled score,[18] and percent of students whose scores fall into particular bins, referred to as proficiency levels. I use these data for 2^{nd}-5^{th} grade math and English language arts tests.

Table 1 lists descriptive statistics for all funded and unfunded elementary schools in 2007, and those for which p_i, the probability a school is selected, lies between 0.10 and 0.90, as well as descriptive statistics for the restricted sample in 2011. The restricted sample is similar to the full sample with two notable exceptions: a much smaller proportion of schools in the restricted sample are in Los Angeles, and funded schools in the restricted sample have a higher TEI.

Both funded and unfunded schools in QEIA typically had high proportions of students who were Hispanic, English language learners, eligible for free and reduced price lunch, and whose parents did not have a college degree. In 2007, a typical school in my sample had at least 1/3 of its teacher work force that was not in the school the prior year.[19] New teachers

[17]The average is weighted by the proportion of students for whom there is a valid score, and each subject and test receives an additional weight.

[18]The scaling of scores takes into account changes in difficulty of tests across years, and therefore makes yearly comparisons more meaningful.

[19]Recall that my teacher-level data set can only distinguish net changes in teacher characteristics. New teachers who are observationally identical to departing teachers from the previous year are not recorded as new, and thus the one third estimate is a lower bound.

did however tend to have nearly three years of experience.

From Table 1 it is apparent that the reduction in class size in grades kindergarten through third is driven by increased class sizes in unfunded schools, while the reduction in class size in fourth and fifth grade is driven by smaller class sizes in funded schools. The QEIA requirement for class sizes in kindergarten through third grade replicated that of California's prior class size reduction policy, the incentives for which were drastically weakened in the first year of QEIA. This weakened incentive led many unfunded schools to gradually increase class sizes in lower grades.

5 Identification

Estimating the causal effect of QEIA is complicated by two facts: districts ranked schools according to unobserved objectives, and districts with more participating schools were more likely to be chosen at least once. A simple comparison of funded and unfunded schools within a district would surely be biased, though the direction of bias would depend on the district's objective functions. A comparison across even just the highest ranked schools in each district would also likely be biased, since larger districts, e.g., Los Angeles Unified, were almost certain to have their highest ranked schools funded, and the size of a district could be correlated with potential outcomes. Even within school over time, potential outcomes might be correlated with treatment if districts gave higher rankings to schools poised to improve.[20]

Instead, my estimation strategy relies on the following intuition: if we were to compare only schools that had an equal probability of being funded, e.g., 50%, then within that group treatment is random, and an OLS estimate would be consistent and unbiased. For each probability we could repeat this exercise, yielding treatment effects conditional on each probability. By the Law of Iterated Expectations, the unconditional average treatment effect could then be recovered. As Wooldridge (2004) shows, the result of an exercise like this is

[20]There is anecdotal evidence that this did in fact happen: in personal communication with a CDE employee who was on Sacramento's school board when schools were ranked, I was told that one school ranked highly because, with a new and talented principal soon starting there, it was poised to improve.

equivalent to the following population specification for τ_{ATE}, the average treatment effect:[21]

$$\tau_{\text{ATE}} = \text{E}\left(\frac{T_i y_i}{p_i} - \frac{(1-T_i)y_i}{(1-p_i)}\right) = \text{E}\left(\frac{(T_i - p_i)y_i}{(1-p_i)p_i}\right) = \text{E}(y_{i1} - y_{i0}) \tag{1}$$

where T_i is an indicator for treatment, in this case being funded by QEIA; y_i is an outcome measure; y_{i0} is the outcome for school i if it is not selected, and y_{i1} is the outcome for school i if it is selected; and p_i is the probability of selection, i.e., the propensity score. $\text{E}(y_{i1} - y_{i0})$ is the Average Treatment Effect: it captures the average change in outcome caused by QEIA. The parameter in equation (1) can be estimated using its sample analog.

Given the selection mechanisms, determining the functional form of p_i is complex.[22] However, given the rules of selection and districts' rankings, I am able to determine the true propensity score (up to an arbitrarily small error) by simulation; I do so by randomly assigning the numbers 1-1,260 to districts and replicating the selection process 10,000 times.[23]

This method allows for the causal effect of QEIA to be non-parametrically identified if two assumptions are satisfied. First, treatment must be mean independent of the potential outcomes conditional on the propensity score (i.e., $\text{E}(y_j|T,p) = \text{E}(y_j|p)$, $j \in \{0,1\}$). This requirement is satisfied by the nature of the selection process. The second requirement is that there can be no schools for which $p_i = 1$ or $p_i = 0$. The intuition for this requirement is straightforward. Among the schools for which $p_i = 0$ or $p_i = 1$, there is no variation in treatment, and so these schools contribute nothing to identification. Among schools participating in QEIA, some were in counties with only one participating school, and that

[21] Wooldridge (2004) shows that the coefficient in a "conditional linear projection" of outcome on treatment, where the conditioning is on probability of selection, can be averaged across probabilities to yield this form of the average treatment effect. He also notes several alternative and asymptotically equivalent forms of the estimator. The estimator is similar to that used in Horvitz and Thompson (1952). See also Imbens and Wooldridge (2009) and Wooldridge (2010).

[22] Were the total number of high schools and elementary or middle schools predetermined, the problem would be considerably simpler, and p_i would be based on a summation of hypergeometric functions, weighted by the probability that the district has a school selected for geographic diversity. Since the number of schools selected depended ultimately on the number of students in each grade level in each school, the problem is considerably more complicated.

[23] The actual random numbers assigned to districts are also made publicly available by CDE. Using these, and the district rankings, my simulation of the selection process perfectly predicts funded schools.

school was therefore selected with probability one. Conversely, the middle and elementary schools that applied for the alternative program had zero probability of being selected. There were also many schools, e.g., Los Angeles Unified's highest ranked schools, whose probability of selection was near one, and many, e.g., Los Angeles Unified's lowest ranked schools, whose probability of selection was very near zero.

In practice, researchers drop observations with probability of treatment "close" to zero or one. Crump et al. (2009) suggest discarding observations less than α away from zero or one, where α satisfies the following:

$$\frac{1}{\alpha(1-\alpha)} = 2 * \mathrm{E}\left[\frac{1}{p_i(1-p_i)} \middle| \frac{1}{p_i(1-p_i)} < \frac{1}{\alpha(1-\alpha)}\right]. \tag{2}$$

As a general rule of thumb, Crump et al. (2009) suggest using $\alpha = .10$. After dropping schools for which p_i is identically zero or one, I am able to calculate (2), and $\alpha = .10$ nearly satisfies this requirement exactly. I therefore restrict the sample to schools for which $p_i \in [0.10, 0.90]$.

To examine whether funded and unfunded schools share a common support across p_i, Figure 1 graphs the number of elementary schools that are funded and unfunded by bins of p_i. Since schools are not uniformly distributed within each bin, we should not necessarily expect the proportion of schools funded to be the midpoint in each bin even in the population.

Though consistent, the sample analog to (1) is not efficient: as Hahn (1998) shows, it fails to achieve the semiparametric efficiency bound. Hirano, Imbens and Ridder (2003) show that a two-step estimator, in which the first step estimates the probability of treatment using a logit series estimator, does achieve the semiparametric efficiency bound, even when the true probability is known. This puzzle is well known in the econometric literature (Henmi and Eguchi, 2004, Hitomi et al., 2008, Prokhorov and Schmidt, 2009, Han and Kim, 2011), though as far as I know the result has never been applied empirically, presumably because probability of treatment is rarely known, as it is in this case.

Though seemingly counter-intuitive, this result rests on a well-known fact: even under

exogenous treatment, if variation in the outcome can be explained by variation in other observables, partialling out this variation results in more efficient estimation. This same principle leads to the inclusion of covariates in an OLS estimate with random and dichotomous treatment. An OLS estimate of the causal effect is consistent and unbiased without covariates, but is more precise when covariates that explain variation in the outcome are included.

Wooldridge (2010) makes explicit the application of this intuition. Consider $k_i = [(T_i - p_i)y_i]/[p_i(1-p_i)]$, where $\mathrm{E}(k_i) = \tau$, my population parameter of interest. We would of course estimate τ using the sample average of k_i, but doing so treats variation in k_i that is explained by variation in covariates as noise, leading to inefficient estimation.

If instead we were to estimate p_i in a first stage using a logit model, as Hirano et al. (2003) suggest, this would be equivalent to regressing $\hat{k}_i = [(T_i - \hat{p}_i)y_i]/[\hat{p}_i(1-\hat{p}_i)]$, where $\hat{p}_i$ is the predicted probability from the first stage, on a constant and $\hat{d}_i = \mathbf{X}_i(T_i - \hat{p}_i)$: the constant would be an estimate of τ, and the residuals can be used to estimate the variance of $\hat{\tau}$. To the degree that $\hat{d}_i$ explains variation in $\hat{k}_i$, $\hat{\tau}$ will gain efficiency. Another way to reach the same conclusion is to note that $\mathrm{E}(k_i - \tau) = 0$ and $E(d_i) = 0$ are moment conditions. Estimating τ treating p_i as known disregards the second moment condition, which, so long as it is correlated with the first moment condition, contains useful information that we incorporate in estimation by treating p_i as unknown.

With known p_i, the gains in efficiency can be achieved by regressing k_i on $d_i = \mathbf{X}_i(T_i - p_i)$. I provide results using the sample analog of (1), which I refer to as those with one moment condition, and results that regress k_i on d_i, where $\mathbf{X}_i$ includes an indicator for having met the growth target, proportion of students eligible for Free and Reduced Price Lunches, enrollment, Standardized API, percent of students who are Hispanic, English language learners, and migrant. I use the value of these variables in 2007. As illustrated in Table 2, the sample moment conditions implied by $\mathrm{E}(d_i) = 0$ are all quite close to zero.

6 Results

6.1 Regression Results

For comparison, I first present results based on various regression specifications with the main QEIA requirements as outcome variables. It's important to note that, unlike the IPW estimator, consistent estimation of average treatment effects by the regression models depends on untestable assumptions. Each regression has a full set of year dummies and interactions between a treatment indicator and the year dummies. For expositional purposes the table includes only the coefficient on the interaction between the treatment indicator and the dummy for 2011.[24] I present results from regressions on the full sample as well as regressions on the restricted sample such that $p_i \in [.10, .90]$.

For each main QEIA requirement, I present five regression models. Model 1 includes only the year dummies and interactions between a treatment indicator and the year dummies. This model consistently estimates the effect of QEIA on outcomes only if treatment is uncorrelated with potential outcomes. Since the probability of treatment is dependent on district rankings, as well as on the size of the district, we wouldn't expect this assumption to be satisfied.

The second regression model adds to Model 1 an interaction between the year dummies and the probability of treatment, p_i. If there is no heterogeneity in the treatment effect, Model 2 will consistently estimate it. If there is heterogeneity, then consistent estimation of the average treatment effect requires $\text{Var}(T|p)$ to be uncorrelated with potential outcomes (Wooldridge, 2004). There is of course no way to know whether this condition is satisfied.

Model 3 also nests Model 1, and includes as covariates an indicator for whether the school met its growth target in 2007, whether the school is in Los Angeles Unified, the percent of students eligible for free and reduced price lunch in 2007, and the enrollment in 2007. If conditional on these covariates treatment is uncorrelated with potential outcomes,

[24] The full set of results is available in an online appendix

the average treatment effect will be consistently estimated.

Models 4 and 5 build on Model 1 by including fixed effects in the former, and the above mentioned covariates with fixed effects in the latter. Fixed effects estimation requires treatment to be uncorrelated with trends in potential outcomes. This assumption would be violated if districts ranked highly those schools that were primed for improvement.

As Table 3 illustrates, the estimated effect of QEIA on class size is robust to a broad range of specifications and to the sample restriction. Average class size is estimated to have decreased in selected schools by about 4.5 students per class. In the full sample, estimates of the effect of QEIA on teacher experience are similarly robust to a broad range of specifications. Average experience appears to have decreased by 0.73 to 0.98 years, suggesting that funded schools were not able to reduce class size by hiring more experienced teachers. In the restricted sample, the standard errors are generally larger and the effects are smaller in each model, suggesting at most a 0.74 reduction in average teacher experience, significant at the 10% level.

The regression estimates of the effect of QEIA on schools' API vary across models. Controlling for the probability of treatment, the API in funded schools is estimated to have increased by 0.41 standard deviations ($p < 0.001$) in the distribution of APIs across all elementary schools in California. At the other extreme, controlling for covariates suggests QEIA had no effect on schools' API. Estimates from the restricted sample are precise and less widely dispersed, with a maximum of 0.4 standard deviations and a minimum of 0.26 standard deviations.

Results for 5^{th} grade assessments in math and English language arts vary across specifications, with effects for math being larger across specifications in the full and restricted sample. In the full sample, the effect of QEIA on math scores varies from 0.25 ($p < 0.01$) to 0.46 ($p < 0.001$) standard deviations in the population of all school-level averages in California, and ELA scores range from an insignificant 0.06 to 0.30 ($p < 0.001$). The estimates are more uniform in the restricted sample, varying from 0.41 ($p < 0.001$) to 0.50 standard

deviations ($p < 0.001$) in 9math, and from 0.31 ($p < 0.001$) to 0.19 ($p < 0.01$) in ELA.

There is some evidence from the regression results of an increase in persistence in funded schools, measured by the percent of students who were in the school the previous year. A causal effect of QEIA on the composition of students in a school could suggest that it did not benefit particular students, but rather better students were attracted to the QEIA schools. However, the estimated effects are small relative to the baseline of about 90% (see Table 1), precise only in some specifications, and then only precise at the 5% level. More importantly, as indicated below, results from IPW estimates suggest no change in student characteristics.

6.2 Main Results

Unlike the regression results above, IPW estimates depend only on the assumption that the randomization was carried out correctly, and by all accounts it was. The remainder of the paper therefore focuses on these estimates, presenting those that depend on one and two sets of moment conditions.[25]

Table 4 shows the causal effect of QEIA on average class size, the percent of teachers classified as highly qualified, average teacher experience, and the TEI. The point estimate on class size in 2009, the first full year of QEIA funding, suggests that QEIA reduced class size, but the effect is imprecisely measured. The standard errors are much smaller using two sets of moment conditions, though they are still larger than those from the regression. Using estimates based on two moment conditions, in the final year for which class size data are available, QEIA reduced class size by 3.95 students per class, an estimate that is significant at the 0.001 level.

Consistent with the claim that both funded and unfunded schools were required to have high proportions of HQT teachers, being funded had no causal impact on the proportion of HQT teachers. The estimates that rely on two sets of moment conditions are all practically

[25]Average treatment effects on the treated are available in the online appendix. The point estimates are quite similar to the average treatment effects, and are too noisy to distinguish from the average treatment effects.

small, precisely measured, and not statistically discernible from zero.

Similarly, teacher experience does not appear to have been affected by QEIA. From 2009 through 2011 the point estimates using both one and two sets of moment conditions are neither positive nor statistically discernible from zero, as evidenced by Table 4. The point estimates based on two moment conditions are smaller in absolute value than the regression estimates, both of which are small relative to the 2007 baseline of 11.81. The marginally significant effects on TEI in 2007 are presumably spurious, and cast doubt on the significant differences in 2008 and 2009, using one moment condition, and 2008, using two moment conditions. Even taking the point estimates at face value, QEIA appears to have reduced teacher experience, measured by years or by the TEI.

The estimated effect of QEIA on student achievement, as measured by California's API, is quite similar to the regression estimates based on the restricted sample, as shown in Table 5. Using two sets of moment conditions,[26] QEIA increased API scores in funded schools by 0.33 standard deviations ($p < 0.001$) by 2011, with respect to the population of all elementary schools. The effect for Hispanic students is significantly larger than for all students by 2011 ($p = 0.046$), as is the effect for low-SES students ($p = 0.051$). From 2008 onward there is a clear pattern of funded schools improving over unfunded schools.

These estimates capture the causal effect of QEIA at the school level. However, it is possible that these results are driven partly by changes in the composition of students in response to QEIA. For instance, it may be that especially savvy parents, whose children are more likely to receive extra support, will be aware of QEIA and select into a QEIA school. Although I cannot currently observe student-level characteristics, I do observe school-level averages of such things as Free and Reduced Price Lunch eligibility, whether parents have a college degree, their race, and whether the student was enrolled in the school the prior year.

Table 6 displays the results for FRPL,[27] percent of students whose parents have a college

[26]Note that effects on the 2007 API scores are not calculated using both moment conditions. This is because I use 2007 API scores in the second set of moment conditions.

[27]Estimates relying on two sets of moment conditions for FRPL in 2007 are not calculated, since FRPL in 2007 is used in that set of moment conditions.

degree, and percent of students who were in the school the prior year. Focusing on the estimates based on two sets of moment conditions, no coefficient is significant, and the magnitudes of the point estimates are quite small. Similarly, Table 7 shows no discernible impact on student enrollment, proportion black, or proportion Hispanic. This is consistent with the student population not changing in response to QEIA.

However, it is still possible that the population of test takers at schools may have changed in response to QEIA. This is particularly concerning since schools were required to improve API scores in order to remain in the program, increasing the stakes of the tests. Schools could manipulate their API scores by either encouraging more students to take alternative tests,[28] discouraging low-performing students from taking any test, or manipulating answer sheets.

Of these possibilities, I currently am able to observe the number of students for whom there is a valid score, and the number who take the regular standardized test. Table 8 displays the results of this analysis. There is no evidence that the number of valid scores differed between funded and unfunded schools, either before or after QEIA. Similarly, there is no evidence of a difference in the number of valid scores for low-SES students or for Hispanic students. Neither is there evidence of a change in the proportion of students taking the regular standardized test. Though this is not definitive evidence, it is at least consistent with the population of test takers not changing in response to QEIA.

Two key policy levers of QEIA, decreased class size and increased teacher experience, require changes to the teacher workforce at a school. Using the teacher-level data, I am able to observe the net changes in teacher characteristics at a school. Tables 9 and 10 list the results from this analysis. I examine differences caused by QEIA in the proportion of teachers new to the school, new to the school but not new to the district,[29] average

[28] Alternative tests are included in the calculation of the API, but presumably the marginal student would find the regular California Standardized Test challenging, and the California Alternative Performance Assessment or the California Modified Assessment less so.

[29] A teacher is new to the school but not the district if no teacher with the same characteristics is observed in the school the prior year, and the teacher has more than one year of experience in the district.

experience conditional on being new to the school, and proportion of probationary, tenured, and temporary teachers.

In 2009 QEIA appears to have caused an increase in the proportion of teachers new to the school in funded schools relative to unfunded schools. In 2009 there were 7 percentage points more ($p < 0.1$) new teachers in funded schools. The similar estimates for the change in new teachers with experience in the district suggests that nearly all teachers new to the school had experience in the district. Comparing the set of teachers new to a school in funded and unfunded schools, average experience is 0.94 years greater in funded schools in 2009 ($p < 0.1$).

Table 10 lists the change in proportion of teachers who are probationary, tenured, and long-term substitutes. The differences between funded and unfunded schools in the proportion of probationary teachers before 2008, significant at the 10% level, are presumably spurious, casting some doubt on the results in later years. Assuming that the more precisely estimated differences in proportion probationary after 2008 are not spurious, there is evidence of an increase in probationary teachers caused by QEIA. There is also marginally significant evidence of fewer tenured teachers in funded schools ($p < 0.10$).

Tables 11 and 12 show the effect of QEIA on class sizes at the grade level. Estimates based on one set of moment conditions are noisy, and are at no point statistically different from zero. Class sizes in kindergarten through 3^{rd} grade are not affected by QEIA until 2011, at which time there are 2.7 to 3.7 fewer students in those grades in funded schools. As mentioned above, class size data are not available in 2010, and that the difference in class sizes in these earlier grades is driven by unfunded schools exiting the previous class size reduction program. Estimates based on two sets of moment conditions suggest that grades 4 and 5 decreased class sizes by about 4.4 ($p < 0.001$) and 4 ($p < 0.01$) students per class in 2009, respectively, and by 5.1 ($p < 0.001$) to 5.7 ($p < 0.001$) students in 2011.

The effect of QEIA on API scores is important, since the primary goal of the policy was to improve API scores. However, given that the API is an average across students, grades,

subjects and even test types, changes in API scores are hard to interpret or compare to other findings in the literature. Tables 13 and 14 therefore list the estimated effect of QEIA on mean scaled scores from California's Standardized Test for math and English language arts.

The effects of QEIA on math scores is greater in later years of the program and in higher grades. There's no discernible effect on 2nd grade math scores until 2011, when they are 0.28 standard deviations higher in funded schools, with respect to the population of grade-level averages ($p < 0.001$, 0.12 student-level standard deviations). The 3rd grade math scores increase one year earlier, in 2010, by 0.16 standard deviations (0.07 student-level standard deviations, $p < 0.05$), and by 0.27 standard deviations by 2011 (0.12 student-level standard deviations, $p < 0.01$).

Math scores in 4th grade improve earlier; by 2009 they show an increase of 0.32 standard deviations(0.15 student-level standard deviations, $p < 0.001$), 0.40 (0.17 student-level standard deviations, $p < 0.001$) in 2010, and level off in 2011 at 0.38 (0.16 student-level standard deviations, $p < 0.001$). Interestingly, 5th grade math scores to not begin improving until 2010, at which time they were 0.36 standard deviations higher in funded schools (0.16 student-level standard deviations, $p < 0.001$), and by 2011 they were 0.44 standard deviations higher (0.20 student-level standard deviations, $p < 0.001$)

Consistent with results from the vast majority of education reforms, the effects are smaller for English language arts. Still, the previous pattern persists: effects are larger in later years, in later grades, and there is an effect on 4th grade test scores in 2009 but not on 5th grade test scores. By 2011, ELA scores in 2nd grade are 0.18 standard deviations higher in funded schools (0.08 student-level standard deviations, $p < 0.01$), and in 5th grade they are 0.21 standard deviations higher (0.10 student-level standard deviations, $p < 0.001$).

To better understand the effects of increased exposure to QEIA, Figure 2 replicates the information in the tables, displaying the average treatment effect on class size and achievement at the grade level, but by cohort exposure. Each panel in the figure displays the change in class size and achievement that a group of students with a normal grade progression would

face. For instance, students in panel A enter kindergarten in 2005, and those who progress one grade each year are exposed to QEIA for one year, in 2009. Since class size data aren't available in 2010, I instead use the same-grade average from 2009 and 2011, e.g., the 3rd grade class size in 2010 is the average of 3rd grade class size in 2009 and 2011.

As the figure suggests, consecutive years of smaller classes does not lead to a widening of the achievement gains. Additionally, though it is not possible to empirically separate the effects of teacher training, high-stakes testing, and reduced class size, it must nonetheless be the case that if teacher training and high-stakes testing explain the improved scores, the timing would have to be correlated with changes in class size. Since both the reduced class sizes in 2nd and 3rd grade are delayed, it seems likely that the effect is driven by class size. Otherwise, there would have to be some reason that professional training was delayed, and that schools did not respond to high-stakes testing until 2011,

7 Conclusion

California's QEIA provides a unique opportunity to study the causal effects of school reform. Using district rankings and the details of the selection process, the probability of any school being selected is known. Between any two schools with the same probability of selection, being funded is uncorrelated with potential outcomes. Using this, and relying on methods described in Wooldridge (2004) and Hirano, Imbens and Ridder (2003), I am able to estimate the causal impact of QEIA by inverse probability weighting.

Doing so, I find that QEIA caused a decrease in class size, and had no discernible effect on teacher experience. Two of the other QEIA requirements applied to all QEIA eligible schools, and are therefore not considered part of the treatment here. The remaining components of treatment, professional training for teachers, which is unobserved, and added incentive to increase achievement to maintain funding, may also contribute to the improvement in test scores.

Test scores improved significantly, albeit unevenly across grades and years. Grades 4 and 5, in which class sizes were first reduced, experienced the largest and earliest increase in test scores. In the first fully-funded year of the program, math scores in 4^{th} grade increased by 0.32 standard deviations in the population of school-grade averages, and by the second fully-funded year 5^{th} grade math scores improved by 0.36 standard deviations. The improvement in test scores in 2^{nd} and 3^{rd} grade, like the reduction in class sizes in those grades, occurred later, and was more modest. By the third fully-funded year of the program, math scores in 2^{nd} grade were 0.28 standard deviations higher in the distribution of school-grade averages, and 0.27 standard deviations higher 3^{rd} grade. For teacher professional training and added test pressure to explain the improvement, they would have to exhibit a similar pattern of implementation across grades and years. Gains in English language arts were modest, but exhibit the same pattern across grades and years.

Though the design of QEIA precludes separate identification of effects of its constituent reforms, it is nonetheless a remarkable policy, unprecedented in education for being a large-scale policy intervention with random assignment. In terms of students affected, it dwarfs Tennessee's Project STAR by an order of magnitude. Though certainly imperfect, more policies like QEIA could vastly improve our understanding of the effectiveness of school reform.

References

Angrist, J., Lavy, V., 1999. Using Maimonides' rule to estimate the effect of class size on scholastic achievement. The Quarterly Journal of Economics 114, 533–575.

Balcom, F., 2007. Quality Education Investment Act (QEIA) of 2006. `http://www.cde.ca.gov/fg/fo/r16/documents/qeia07present.ppt`.

Barrett, N., Butler, J., Toma, E.F., 2012. Do less effective teachers choose professional development does it matter? Evaluation Review 36, 346–374.

Bluth, A.H., 2005. Lawsuit seeking cash for schools; Governor broke his word, say teachers and schools chief. Sacramento Bee `http://www.mikemcmahon.info/ctasuit.htm`.

CDE, 2010. Report to the Legislature and the Governor; Quality Education Investment Act First Progress Report. `http://www.cde.ca.gov/ta/lp/qe/documents/qeialegrpt.doc`.

Chetty, R., Friedman, J.N., Hilger, N., Saez, E., Schanzenbach, D.W., Yagan, D., 2011a. How does your kindergarten classroom affect your earnings? Evidence from Project STAR. The Quarterly Journal of Economics 126, 1593–1660.

Chetty, R., Friedman, J.N., Rockoff, J.E., 2011b. The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood. Technical Report. National Bureau of Economic Research.

Chiang, H., 2009. How accountability pressure on failing schools affects student achievement. Journal of Public Economics 93, 1045–1057.

Chingos, M.M., 2012. The impact of a universal class-size reduction policy: Evidence from Florida's statewide mandate. Economics of Education Review 31, 543–562.

Cohodes, S., Deming, D., Jennings, J., Jencks, C., 2013. School accountability, postsecondary attainment and earnings. NBER Working Paper .

Crump, R.K., Hotz, V.J., Imbens, G.W., Mitnik, O.A., 2009. Dealing with limited overlap in estimation of average treatment effects. Biometrika 96, 187–199.

Dieterle, S., 2013. Class-size reduction policies and the quality of entering teachers.

Figlio, D.N., Getzler, L.S., 2006. Accountability, ability and disability: Gaming the system? Advances in applied microeconomics 14, 35–49.

Greenwald, R., Hedges, L., Laine, R., 1996. The effect of school resources on student achievement. Review of Educational Research 66, 361–396.

Hahn, J., 1998. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. Econometrica , 315–331.

Han, C., Kim, B., 2011. A GMM interpretation of the paradox in the inverse probability weighting estimation of the average treatment effect on the treated. Economics Letters 110, 163–165.

Hanushek, E., 1979. Conceptual and empirical issues in the estimation of educational production functions. Journal of Human Resources , 351–388.

Hanushek, E., 1986. The economics of schooling: Production and efficiency in public schools. Journal of Economic Literature 24, 1141–1177.

Hanushek, E., 1996. A more complete picture of school resource policies. Review of Educational Research 66, 397–409.

Hanushek, E., 1997. Assessing the effects of school resources on student performance: An update. Educational Evaluation and Policy Analysis 19, 141.

Hanushek, E., 1999. Some findings from an independent investigation of the Tennessee STAR experiment and from other investigations of class size effects. Educational Evaluation and Policy Analysis 21, 143.

Henmi, M., Eguchi, S., 2004. A paradox concerning nuisance parameters and projected estimating functions. Biometrika 91, 929–941.

Hirano, K., Imbens, G.W., Ridder, G., 2003. Efficient estimation of average treatment effects using the estimated propensity score. Econometrica 71, 1161–1189.

Hitomi, K., Nishiyama, Y., Okui, R., 2008. A puzzling phenomenon in semiparametric estimation problems with infinite-dimensional nuisance parameters. Econometric Theory 24, 1717–1728.

Horvitz, D.G., Thompson, D.J., 1952. A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association 47, 663–685.

Hoxby, C., 2000. The effects of class size on student achievement: New evidence from population variation. Quarterly Journal of Economics 115, 1239–1285.

Imbens, G.W., Wooldridge, J.M., 2009. Recent developments in the econometrics of program evaluation. Journal of Economic Literature 47, 5–86.

Jacob, B.A., 2005. Accountability, incentives and behavior: The impact of high-stakes testing in the chicago public schools. Journal of Public Economics 89, 761–796.

Jacob, B.A., Levitt, S.D., 2003. Rotten apples: An investigation of the prevalence and predictors of teacher cheating. The Quarterly Journal of Economics 118, 843–877.

Jepsen, C., Rivkin, S., 2009. Class size reduction and student achievement: The potential tradeoff between teacher quality and class size. Journal of Human Resources 44, 223–250.

Krueger, A., 1999. Experimental estimates of education production functions. The Quarterly Journal of Economics 114, 497–532.

Krueger, A., 2002. Understanding the magnitude and effect of class size on student achievement. The Class Size Debate , 7–35.

Krueger, A., Whitmore, D., 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. The Economic Journal 111, 1–28.

Neal, D., Schanzenbach, D.W., 2010. Left behind by design: Proficiency counts and test-based accountability. The Review of Economics and Statistics 92, 263–283.

Nye, B., Hedges, L., Konstantopoulos, S., 1999. The long-term effects of small classes: A five-year follow-up of the Tennessee class size experiment. Educational Evaluation and Policy Analysis 21, 127–142.

Prokhorov, A., Schmidt, P., 2009. GMM redundancy results for general missing data problems. Journal of Econometrics 151, 47–55.

Rice, J., Schwartz, A., 2008. Handbook of Research in Education Finance and Policy. Routledge New York. chapter 8. pp. 131–165.

Rockoff, J., 2009. Field experiments in class size from the early twentieth century. The Journal of Economic Perspectives 23, 211–230.

Santa Rosa City Schools, 2007. School board minutes, Quality Education Investment Act. `http://www.srcs.k12.ca.us/board/agendas/attachments/032807-BR-F7.pdf`.

Todd, P., Wolpin, K., 2003. On the specification and estimation of the production function for cognitive achievement. The Economic Journal 113, F3–F33.

Wooldridge, J.M., 2004. Estimating average partial effects under conditional moment independence assumptions. CeMMAP Working Paper Number CWP03/04 .

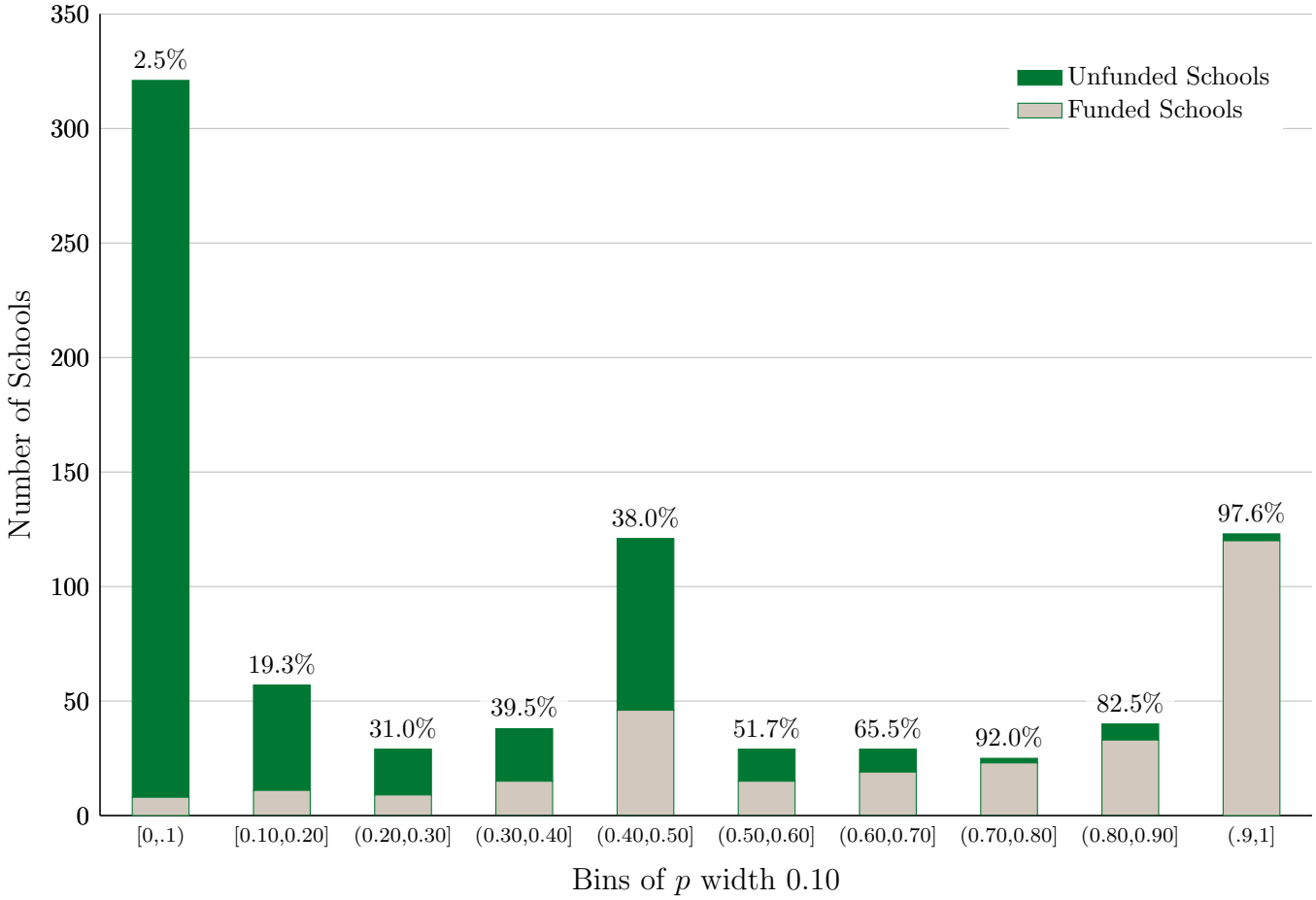
Wooldridge, J.M., 2010. Econometric analysis of cross section and panel data. Second ed., MIT Press.

Yoon, K.S., Duncan, T., Lee, S.W.Y., Scarloss, B., Shapley, K.L., 2007. Reviewing the evidence on how teacher professional development affects student achievement. National

Center for Educational Evaluation and Regional Assistance, Institute of Education Sciences, US Department of Education.

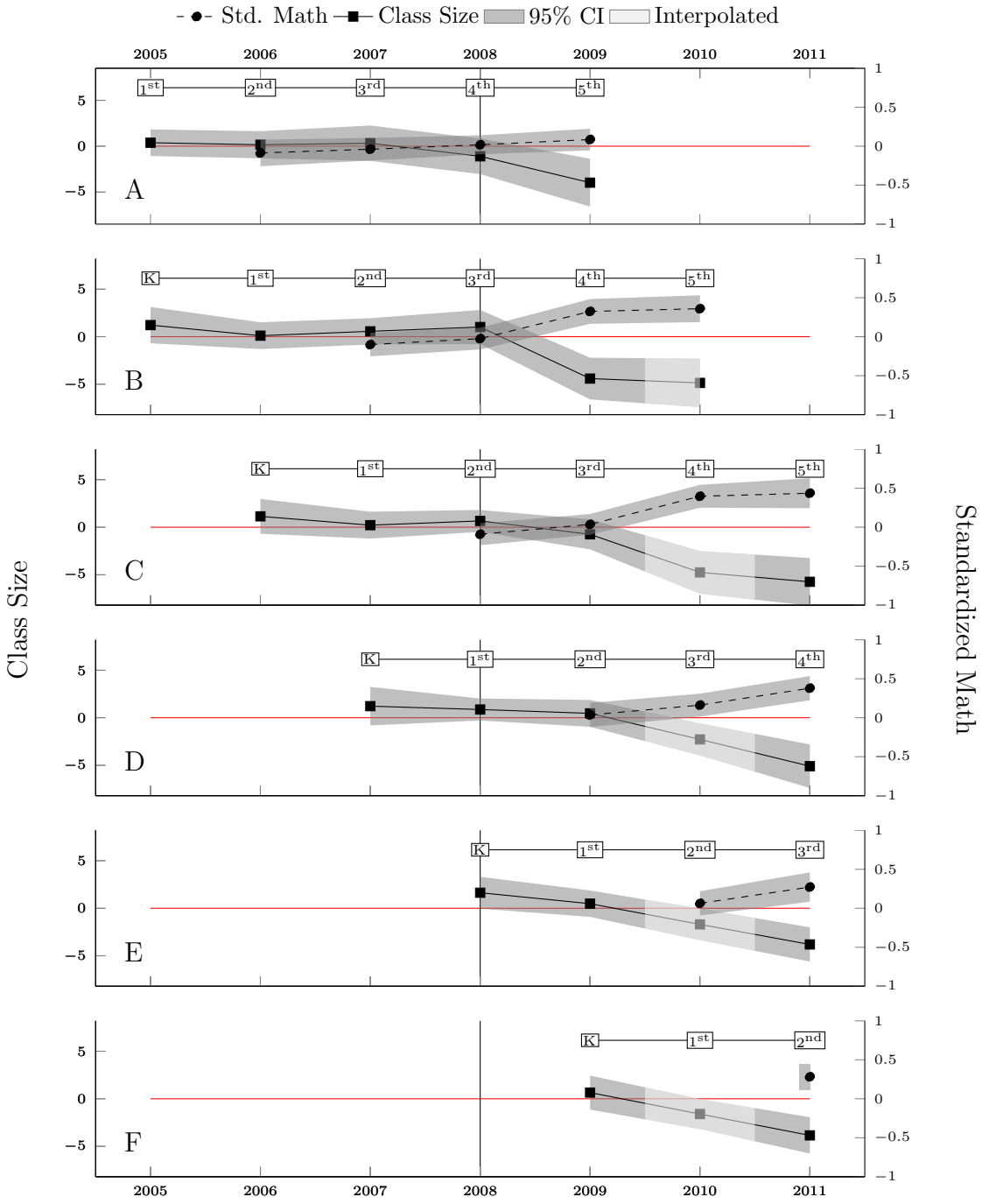
A Figures

Figure 1: Support Over p, All Elementary Schools



Note: Numbers above bars refer to the percentage of schools funded in that bin. Sample includes all elementary schools participating in the regular QEIA program.

Figure 2: Cohort-Level Class Size and Math Achievement Comparison



Note: Estimates are of average treatment effect using two moment conditions. Left axis refers to class size, right axis refers to standardized math scores on California's CST. Class size data are missing for 2010. Shaded regions indicate use of average of 2009 and 2011 same-grade class size. For example 2010 4^{th} grade class size is average of 2009 4^{th} grade class size and 2011 4^{th} grade class size. Only grades 2 and above are tested.

B Tables

Table 1: Descriptives, Elementary Regular QEIA Schools 2007

	All Elem.	$p_i \in [0,1]$, 2007		$p_i \in [.1,.9]$, 2007		$p_i \in [.1,.9]$, 2011	
	2007	Unfunded	Funded	Unfunded	Funded	Unfunded	Funded
Avgerage Class Size	21.86	21.93	21.96	22.14	22.12	25.02	20.36
	(3.54)	(1.87)	(2.04)	(1.75)	(2.05)	(3.17)	(2.20)
Class Size Kindergarten	20.64	20.24	20.41	20.30	20.74	23.97	20.23
	(4.14)	(3.39)	(3.78)	(3.47)	(4.21)	(4.29)	(3.48)
Class Size 1^{st} Grade	19.32	19.29	19.13	19.33	19.21	23.84	19.54
	(1.85)	(1.40)	(1.31)	(1.47)	(1.29)	(4.00)	(2.27)
Class Size 2^{nd} Grade	19.14	18.84	18.97	18.79	19.00	23.88	19.35
	(1.87)	(1.52)	(1.40)	(1.56)	(1.34)	(4.18)	(2.53)
Class Size 3^{rd} Grade	19.89	19.42	19.60	19.76	19.68	23.92	19.56
	(3.17)	(2.71)	(3.01)	(3.19)	(3.22)	(4.48)	(2.64)
Class Size 4^{th} Grade	28.47	28.01	28.18	28.27	28.58	28.15	22.40
	(4.23)	(3.82)	(3.80)	(3.56)	(3.80)	(4.31)	(3.99)
Class Size 5^{th} Grade	28.89	28.28	28.51	28.68	28.66	28.42	22.24
	(4.15)	(3.72)	(3.77)	(3.56)	(3.97)	(4.51)	(3.35)
Average Experience	13.01	11.92	11.61	12.37	11.68	13.68	12.94
	(3.95)	(3.17)	(3.35)	(3.12)	(3.38)	(3.57)	(3.17)
TEI Relative	-0.04	-0.17	-0.28	-0.06	-0.26	-0.12	-0.15
	(1.05)	(1.02)	(1.00)	(0.89)	(0.98)	(0.80)	(0.69)
Highly Qualified Teachers	0.96	0.94	0.94	0.96	0.94	0.99	0.99
	(0.11)	(0.10)	(0.14)	(0.09)	(0.15)	(0.04)	(0.11)
Williams Settlement Applies	0.24	0.94	0.96	0.95	0.95	0.95	0.95
	(0.43)	(0.24)	(0.20)	(0.21)	(0.22)	(0.21)	(0.22)
Standardized API	0.00	-1.10	-1.25	-1.14	-1.15	-1.10	-0.77
	(1.00)	(0.48)	(0.48)	(0.44)	(0.47)	(0.56)	(0.60)
API Percentile Rank	0.50	0.16	0.12	0.15	0.15	0.17	0.26
	(0.29)	(0.11)	(0.10)	(0.11)	(0.11)	(0.14)	(0.17)
Met Growth Target	0.70	0.68	0.65	0.66	0.60	0.60	0.71
	(0.46)	(0.47)	(0.48)	(0.48)	(0.49)	(0.49)	(0.46)
Proportion Black	0.08	0.09	0.11	0.07	0.08	0.06	0.08
	(0.12)	(0.14)	(0.16)	(0.12)	(0.13)	(0.11)	(0.12)
Proportion Hispanic	0.46	0.79	0.74	0.78	0.76	0.80	0.78
	(0.30)	(0.21)	(0.25)	(0.21)	(0.25)	(0.20)	(0.24)
Proportion White	0.33	0.06	0.07	0.09	0.07	0.08	0.06
	(0.28)	(0.10)	(0.10)	(0.13)	(0.10)	(0.12)	(0.10)
English Language Learners	0.29	0.55	0.55	0.55	0.56	0.52	0.53
	(0.23)	(0.18)	(0.20)	(0.18)	(0.20)	(0.18)	(0.20)
Proportion FRPL	0.55	0.89	0.88	0.87	0.86	0.89	0.85
	(0.31)	(0.11)	(0.11)	(0.12)	(0.12)	(0.14)	(0.17)
Parent College Grad	0.18	0.06	0.06	0.06	0.07	0.07	0.07
	(0.14)	(0.06)	(0.07)	(0.06)	(0.09)	(0.04)	(0.05)
Student Enrollment	376.60	472.24	434.61	446.60	418.85	404.83	377.09
	(192.56)	(195.68)	(179.70)	(164.72)	(151.15)	(139.09)	(133.58)
Proportion Same School	0.92	0.91	0.91	0.91	0.91	0.92	0.93
	(0.08)	(0.04)	(0.04)	(0.04)	(0.04)	(0.05)	(0.08)
Los Angeles	0.09	0.23	0.09	0.06	0.03	0.06	0.03
	(0.29)	(0.42)	(0.29)	(0.24)	(0.17)	(0.24)	(0.17)
Proportion Teachers New to School	0.33	0.30	0.35	0.34	0.36	0.50	0.48
	(0.27)	(0.26)	(0.25)	(0.26)	(0.26)	(0.31)	(0.29)
New to School, Not District	0.24	0.22	0.25	0.25	0.25	0.47	0.43
	(0.24)	(0.23)	(0.23)	(0.24)	(0.23)	(0.30)	(0.29)
Average Experience New Teachers	3.21	2.76	3.13	3.16	3.13	6.43	5.83
	(3.64)	(3.18)	(3.12)	(3.47)	(3.07)	(4.88)	(4.47)
Proportion Temp Teachers	0.06	0.05	0.05	0.06	0.06	0.04	0.04
	(0.10)	(0.08)	(0.10)	(0.09)	(0.11)	(0.08)	(0.07)
Proportion Probationary	0.14	0.13	0.16	0.15	0.18	0.06	0.09
	(0.18)	(0.14)	(0.16)	(0.15)	(0.17)	(0.10)	(0.13)
N	6476	546	307	197	171	197	171

Note: Funded and unfunded includes all elementary schools participating in the regular QEIA program

Table 2: Sample Moment Conditions

Variable	Mean	S.D.
(Funded$_i$-p_i)	0.0032	(0.4540)
(2007 Met Growth Target)(Funded$_i$-p_i)	-0.0195	(0.3607)
(2007 Proportion FRPL)(Funded$_i$-p_i)	-0.0029	(0.3950)
(2007 Student Enrollment)(Funded$_i$-p_i)	-3.4513	(203.2044)
(2007 Std. API)(Funded$_i$-p_i)	0.0009	(0.5296)
(2007 Proportion Hispanic)(Funded$_i$-p_i)	-0.0021	(0.3596)
(2007 English Language Learners)(Funded$_i$-p_i)	0.0029	(0.2611)
(2007 Migrant)(Funded$_i$-p_i)	-0.0074	(0.0623)

Note: Sample analogs of moments in condition $\mathrm{E}[\mathbf{X}(\mathrm{Funded}_i - p_i)] = \mathbf{0}$.

Table 3: Select Regression Results

	(1)	(2)	(3)	(4)	(5)
	Full Sample				
Avg. Class Size	-4.454***	-4.761***	-4.538***	-4.616***	-4.578***
	(0.559)	(0.564)	(0.401)	(0.462)	(0.419)
Experience	-0.831**	-0.730†	-0.903**	-0.932**	-0.977***
	(0.256)	(0.431)	(0.271)	(0.280)	(0.250)
Std. API	0.143†	0.407***	0.079	0.260***	0.221***
	(0.075)	(0.071)	(0.069)	(0.063)	(0.053)
5^{th} Grade Math	0.273**	0.462***	0.248**	0.365***	0.341***
	(0.086)	(0.094)	(0.078)	(0.083)	(0.078)
5^{th} Grade ELA	0.116*	0.301***	0.056	0.173**	0.157**
	(0.059)	(0.069)	(0.056)	(0.054)	(0.048)
Prop. Enrolled Since Previous Year	0.0008	0.010	0.006*	0.004	0.007*
	(0.005)	(0.008)	(0.003)	(0.005)	(0.004)
	$p_i \in [.10, .90]$				
Avg. Class Size	-4.661***	-4.587***	-4.544***	-4.839***	-4.749***
	(0.439)	(0.471)	(0.400)	(0.429)	(0.437)
Experience	-0.738†	-0.659	-0.741†	-0.573†	-0.649†
	(0.381)	(0.450)	(0.398)	(0.345)	(0.340)
Std. API	0.334***	0.400***	0.256***	0.365***	0.333***
	(0.063)	(0.070)	(0.060)	(0.063)	(0.062)
5^{th} Grade Math	0.439***	0.453***	0.408***	0.502***	0.482***
	(0.086)	(0.095)	(0.086)	(0.097)	(0.100)
5^{th} Grade ELA	0.256***	0.305***	0.192**	0.262***	0.244***
	(0.061)	(0.070)	(0.057)	(0.060)	(0.061)
Prop. Enrolled Since Previous Year	0.006	0.009	0.012*	0.006	0.011*
	(0.007)	(0.009)	(0.005)	(0.007)	(0.005)
Covariates	No	No	Yes	No	Yes
Propensity Score	No	Yes	No	No	No
Fixed Effects	No	No	No	Yes	Yes

Note: † indicates $p < 0.10$, * indicates $p < 0.05$, ** indicates $p < 0.01$, *** indicates $p < 0.001$. Standard errors robust and clustered at district level. Omitted variable is 2005 unfunded. Covariates include dummy for whether school met growth target in 2007, percent of students eligible for Free and Reduced Price Lunches in 2007, an indicator for being in LA, and total enrollment in 2007. Results are from regression with time dummies, and time dummies interacted with treatment indicator. When propensity score is included, it is interacted with time dummies. Reported coefficients are from interaction of dummy for 2011 and treatment indicator.

Table 4: Class Size and HQT

	Avg. Class Size		HQT		Experience		TEI	
	ATE 1	ATE 2	ATE 1	ATE 2	ATE 1	ATE 2	ATE 1	ATE 2
2005	1.143	0.528	0.055	0.019	0.347	-0.135	-0.044	-0.013
	(4.119)	(0.880)	(0.187)	(0.033)	(2.083)	(0.576)	(0.102)	(0.102)
2006	1.237	0.567	0.054	0.015	0.223	-0.207	-0.102	-0.053
	(3.963)	(0.854)	(0.165)	(0.040)	(2.094)	(0.636)	(0.106)	(0.092)
2007	1.191	0.493	0.036	-0.004	-0.229	-0.655	-0.304*	-0.231†
	(3.813)	(0.838)	(0.171)	(0.038)	(2.193)	(0.674)	(0.130)	(0.121)
2008	0.977	0.484	0.058	0.016	0.060	-0.431	-0.289*	-0.229†
	(3.861)	(0.739)	(0.177)	(0.040)	(2.207)	(0.641)	(0.135)	(0.131)
2009	-0.537	-1.116	0.052	0.016	-0.0008	-0.355	-0.269*	-0.182
	(3.863)	(0.817)	(0.173)	(0.036)	(2.244)	(0.755)	(0.131)	(0.126)
2010					-0.158	-0.532		
					(2.423)	(0.706)		
2011	-3.356	-3.952***	0.043	0.010	-0.037	-0.455	-0.017	0.008
	(3.984)	(0.923)	(0.183)	(0.037)	(2.480)	(0.653)	(0.100)	(0.118)

Note: † indicates $p < 0.10$, * indicates $p < 0.05$, ** indicates $p < 0.01$, *** indicates $p < 0.001$. ATE1 is estimate of average treatment effect using one moment condition; ATE2 is estimate of average treatment effect using two moment conditions. Standard errors bootstrapped and clustered at district level. Class size data, which is used to calculate TEI and includes HQT data, is not available in 2010.

Table 5: Average Performance Index

	Std. API		Std. API Hispanic		Std. API Low SES	
	ATE 1	ATE 2	ATE 1	ATE 2	ATE 1	ATE 2
2005	-0.069	-0.050	0.035	0.020	-0.120	-0.088
	(0.219)	(0.059)	(0.213)	(0.075)	(0.206)	(0.062)
2006	-0.039	-0.047	0.039	0.010	-0.077	-0.080
	(0.219)	(0.050)	(0.205)	(0.069)	(0.192)	(0.052)
2007	-0.017		0.040	0.048	-0.046	-0.006
	(0.200)		(0.182)	(0.058)	(0.183)	(0.039)
2008	0.027	0.012	0.119	0.085	0.034	0.028
	(0.194)	(0.043)	(0.161)	(0.060)	(0.165)	(0.055)
2009	0.123	0.091†	0.201	0.149†	0.161	0.122
	(0.179)	(0.054)	(0.165)	(0.083)	(0.159)	(0.078)
2010	0.248	0.208***	0.414*	0.325***	0.332*	0.290***
	(0.182)	(0.061)	(0.168)	(0.085)	(0.159)	(0.078)
2011	0.383*	0.330***	0.414**	0.394***	0.445**	0.404***
	(0.169)	(0.065)	(0.152)	(0.082)	(0.165)	(0.082)

Note: † indicates $p < 0.10$, * indicates $p < 0.05$, ** indicates $p < 0.01$, *** indicates $p < 0.001$. ATE1 is estimate of average treatment effect using one moment condition; ATE2 is estimate of average treatment effect using two moment conditions. Standard errors bootstrapped and clustered at district level. 2007 API effect not calculated since that covariate is used in the second moment condition.

Table 6: Demographics

	FRPL		Parents College Degree		Enrolled Since Previous Year	
	ATE 1	ATE 2	ATE 1	ATE 2	ATE 1	ATE 2
2005	0.030	0.023	0.014	0.005	0.056	0.021
	(0.162)	(0.034)	(0.015)	(0.006)	(0.162)	(0.031)
2006	0.018	0.008	0.014	0.005	0.060	0.027
	(0.152)	(0.034)	(0.014)	(0.007)	(0.157)	(0.030)
2007	0.017		0.013	0.005	0.052	0.017
	(0.154)		(0.017)	(0.008)	(0.160)	(0.030)
2008	0.024	0.021	0.006	-0.001	0.059	0.023
	(0.153)	(0.030)	(0.014)	(0.006)	(0.157)	(0.029)
2009	0.013	0.005	0.009	0.0001	0.061	0.026
	(0.158)	(0.034)	(0.014)	(0.006)	(0.158)	(0.032)
2010	0.013	0.002	0.015	0.005	0.067	0.025
	(0.162)	(0.032)	(0.015)	(0.005)	(0.168)	(0.031)
2011	-0.006	-0.018	0.010	-0.0002	0.064	0.029
	(0.155)	(0.039)	(0.016)	(0.005)	(0.170)	(0.031)

Note: † indicates $p < 0.10$, * indicates $p < 0.05$, ** indicates $p < 0.01$, *** indicates $p < 0.001$. ATE1 is estimate of average treatment effect using one moment condition; ATE2 is estimate of average treatment effect using two moment conditions. Standard errors bootstrapped and clustered at district level. FRPL, parents with college degree, and enrolled previous year are expressed in proportions.

Table 7: Demographics

	Enrollment		Black		Hispanic	
	ATE 1	ATE 2	ATE 1	ATE 2	ATE 1	ATE 2
2005	-6.541	1.826	0.026	0.018	0.007	0.006
	(89.279)	(22.459)	(0.027)	(0.014)	(0.130)	(0.028)
2006	-1.332	5.158	0.025	0.016	0.011	0.008
	(82.240)	(18.640)	(0.027)	(0.013)	(0.138)	(0.027)
2007	0.522		0.020	0.014	0.018	
	(82.737)		(0.025)	(0.013)	(0.137)	
2008	-9.580	-4.517	0.018	0.012	0.020	0.015
	(78.651)	(18.204)	(0.024)	(0.013)	(0.140)	(0.028)
2009	-5.770	0.289	0.020	0.014	0.017	0.013
	(74.625)	(16.940)	(0.025)	(0.012)	(0.131)	(0.029)
2010	1.656	2.082	0.015	0.008	0.027	0.015
	(76.704)	(18.336)	(0.024)	(0.012)	(0.138)	(0.029)
2011	-2.720	2.720	0.015	0.009	0.019	0.015
	(72.262)	(18.119)	(0.023)	(0.011)	(0.145)	(0.029)

Note: † indicates $p < 0.10$, * indicates $p < 0.05$, ** indicates $p < 0.01$, *** indicates $p < 0.001$. ATE1 is estimate of average treatment effect using one moment condition; ATE2 is estimate of average treatment effect using two moment conditions. Standard errors bootstrapped and clustered at district level. Black and Hispanic are expressed in proportions.

Table 8: Test Taking

	Valid Scores		Number Low SES Scores		Number Hispanic Scores		CST	
	ATE 1	ATE 2	ATE 1	ATE 2	ATE 1	ATE 2	ATE 1	ATE 2
2005	-2.275	4.076	-5.931	6.224	-5.118	6.959	0.075	0.022
	(77.676)	(19.553)	(72.290)	(18.422)	(64.237)	(16.237)	(0.184)	(0.037)
2006	3.588	9.148	-1.484	8.239	0.820	12.317	0.080	0.027
	(75.051)	(15.011)	(67.519)	(15.449)	(59.916)	(12.972)	(0.176)	(0.035)
2007	-1.175	4.731	-7.220	4.915	-2.778	8.814	0.072	0.029
	(69.466)	(13.761)	(61.021)	(12.123)	(53.774)	(11.992)	(0.187)	(0.035)
2008	-0.971	-2.421	-5.290	0.826	-2.617	2.248	0.056	0.021
	(69.978)	(15.299)	(58.471)	(13.492)	(54.816)	(13.093)	(0.174)	(0.033)
2009	6.147	4.356	-1.070	3.495	3.515	8.631	0.079	0.013
	(66.625)	(14.462)	(62.504)	(14.762)	(54.005)	(13.615)	(0.176)	(0.032)
2010	5.079	6.011	-4.223	3.950	5.240	12.440	0.083	0.012
	(68.400)	(15.908)	(64.110)	(15.185)	(56.355)	(13.371)	(0.172)	(0.034)
2011	4.168	8.715	-7.012	2.553	1.128	12.162	0.076	0.016
	(68.329)	(15.966)	(60.272)	(17.070)	(56.573)	(14.543)	(0.161)	(0.035)

Note: † indicates $p < 0.10$, * indicates $p < 0.05$, ** indicates $p < 0.01$, *** indicates $p < 0.001$. ATE1 is estimate of average treatment effect using one moment condition; ATE2 is estimate of average treatment effect using two moment conditions. Standard errors bootstrapped and clustered at district level. Valid, low SES, and Hispanic scores refers to all standardized tests used in API. CST is proportion of students taking the California Standardized Test, a subset of the API.

Table 9: Teacher Mobility

	New to School		New to School, Not New to District		Average Experiene of New Teachers	
	ATE 1	ATE 2	ATE 1	ATE 2	ATE 1	ATE 2
2005	0.080	0.061	0.086	0.068	0.853	0.632
	(0.076)	(0.045)	(0.066)	(0.043)	(0.800)	(0.534)
2006	0.058	0.042	0.046	0.036	0.485	0.274
	(0.068)	(0.041)	(0.055)	(0.037)	(0.713)	(0.486)
2007	0.023	0.005	-0.008	-0.024	-0.221	-0.490
	(0.073)	(0.041)	(0.064)	(0.040)	(0.826)	(0.650)
2008	0.081	0.066	0.050	0.042	0.663	0.405
	(0.073)	(0.042)	(0.060)	(0.040)	(0.841)	(0.642)
2009	0.101	0.077†	0.089	0.073*	1.141	0.942†
	(0.073)	(0.040)	(0.056)	(0.036)	(0.755)	(0.552)
2010	0.060	0.056	0.062	0.060	0.434	0.469
	(0.077)	(0.040)	(0.071)	(0.039)	(0.894)	(0.524)
2011	-0.013	0.010	-0.026	-0.002	-0.333	0.011
	(0.094)	(0.047)	(0.087)	(0.046)	(1.386)	(0.777)

Note: † indicates $p < 0.10$, * indicates $p < 0.05$, ** indicates $p < 0.01$, *** indicates $p < 0.001$. ATE1 is estimate of average treatment effect using one moment condition; ATE2 is estimate of average treatment effect using two moment conditions. Standard errors bootstrapped and clustered at district level.

Table 10: Teacher Composition

	Probationary		Tenured		Long-term Substitute	
	ATE 1	ATE 2	ATE 1	ATE 2	ATE 1	ATE 2
2005	0.048	0.041†	-0.015	-0.044	0.004	0.003
	(0.039)	(0.023)	(0.124)	(0.053)	(0.014)	(0.013)
2006	0.035	0.029	0.0004	-0.028	0.0003	-0.001
	(0.036)	(0.023)	(0.130)	(0.052)	(0.012)	(0.013)
2007	0.044	0.037†	0.0006	-0.021	0.005	0.004
	(0.034)	(0.020)	(0.128)	(0.056)	(0.015)	(0.014)
2008	0.045	0.041*	-0.006	-0.038	0.013	0.012
	(0.031)	(0.018)	(0.136)	(0.037)	(0.018)	(0.015)
2009	0.033	0.034*	-0.035	-0.064†	0.030†	0.025
	(0.027)	(0.015)	(0.140)	(0.038)	(0.017)	(0.017)
2010	0.038*	0.040**	-0.003	-0.033	0.004	-0.003
	(0.019)	(0.013)	(0.156)	(0.037)	(0.015)	(0.014)
2011	0.028	0.027*	-0.0002	-0.032	0.008	0.004
	(0.019)	(0.013)	(0.163)	(0.039)	(0.013)	(0.011)

Note: † indicates $p < 0.10$, * indicates $p < 0.05$, ** indicates $p < 0.01$, *** indicates $p < 0.001$. ATE1 is estimate of average treatment effect using one moment condition; ATE2 is estimate of average treatment effect using two moment conditions. Standard errors bootstrapped and clustered at district level. Probationary, tenured, and long-term substitute are expressed in proportions.

Table 11: Class Size Grades K-2

	Class Size Kindergarten		Class Size 1^{st} grade		Class Size 2^{nd} grade	
	ATE 1	ATE 2	ATE 1	ATE 2	ATE 1	ATE 2
2005	1.330	1.218	0.792	0.377	0.978	0.513
	(3.716)	(0.975)	(3.603)	(0.740)	(3.623)	(0.733)
2006	1.328	1.143	0.018	0.107	0.674	0.158
	(3.890)	(0.934)	(3.542)	(0.718)	(3.483)	(0.755)
2007	1.306	1.218	0.587	0.216	0.600	0.563
	(3.930)	(1.035)	(3.584)	(0.722)	(3.457)	(0.706)
2008	1.578	1.633†	0.929	0.865	0.769	0.662
	(3.808)	(0.855)	(3.422)	(0.591)	(3.366)	(0.587)
2009	1.154	0.651	0.480	0.482	0.798	0.459
	(3.778)	(0.913)	(3.620)	(0.709)	(3.537)	(0.722)
2010						
2011	-1.988	-2.692*	-3.457	-3.693***	-3.128	-3.851***
	(4.151)	(1.059)	(4.187)	(0.909)	(4.001)	(0.975)

Note: † indicates $p < 0.10$, * indicates $p < 0.05$, ** indicates $p < 0.01$, *** indicates $p < 0.001$. ATE1 is estimate of average treatment effect using one moment condition; ATE2 is estimate of average treatment effect using two moment conditions. Standard errors bootstrapped and clustered at district level. Class size data are not available in 2010.

Table 12: Class Size Grades 3-5

	Class Size 3^{rd} grade		Class Size 4^{th} grade		Class Size 5^{th} grade	
	ATE 1	ATE 2	ATE 1	ATE 2	ATE 1	ATE 2
2005	1.340	0.282	2.138	0.635	3.222	-0.177
	(3.563)	(0.937)	(5.325)	(1.166)	(5.609)	(1.252)
2006	0.844	0.533	1.586	0.174	1.890	-0.292
	(3.644)	(0.896)	(5.248)	(1.142)	(5.237)	(1.176)
2007	1.130	0.322	0.815	0.852	1.874	0.425
	(3.689)	(0.985)	(5.082)	(1.216)	(5.450)	(1.283)
2008	1.265	1.017	-0.435	-1.103	-0.383	-0.870
	(3.705)	(0.917)	(4.668)	(0.991)	(4.989)	(1.007)
2009	-0.294	-0.748	-3.451	-4.408***	-1.964	-3.979**
	(3.574)	(0.807)	(4.538)	(1.120)	(4.875)	(1.333)
2010						
2011	-2.795	-3.807***	-3.404	-5.098***	-4.173	-5.744***
	(4.045)	(0.918)	(4.673)	(1.170)	(4.752)	(1.275)

Note: † indicates $p < 0.10$, * indicates $p < 0.05$, ** indicates $p < 0.01$, *** indicates $p < 0.001$. ATE1 is estimate of average treatment effect using one moment condition; ATE2 is estimate of average treatment effect using two moment conditions. Standard errors bootstrapped and clustered at district level. Class size data are not available in 2010.

Table 13: Math Standardized Test

	2^{nd} Grade		3^{rd} Grade		4^{th} Grade		5^{th} Grade	
	ATE 1	ATE 2	ATE 1	ATE 2	ATE 1	ATE 2	ATE 1	ATE 2
2005	-0.069	-0.045	-0.114	-0.089	-0.088	-0.060	-0.154	-0.114
	(0.222)	(0.081)	(0.196)	(0.064)	(0.201)	(0.062)	(0.205)	(0.076)
2006	-0.108	-0.087	-0.033	-0.025	-0.090	-0.082	-0.076	-0.039
	(0.213)	(0.087)	(0.209)	(0.070)	(0.190)	(0.068)	(0.192)	(0.061)
2007	-0.122	-0.101	-0.078	-0.039	-0.029	-0.018	-0.024	0.003
	(0.189)	(0.077)	(0.192)	(0.074)	(0.178)	(0.053)	(0.168)	(0.055)
2008	-0.086	-0.090	-0.051	-0.025	0.015	0.018	0.131	0.096
	(0.172)	(0.072)	(0.187)	(0.071)	(0.179)	(0.062)	(0.160)	(0.070)
2009	0.062	0.036	0.038	0.036	0.286*	0.324***	0.117	0.085
	(0.168)	(0.079)	(0.174)	(0.068)	(0.145)	(0.081)	(0.153)	(0.071)
2010	0.091	0.063	0.177	0.162*	0.383**	0.397***	0.409**	0.359***
	(0.163)	(0.079)	(0.171)	(0.075)	(0.146)	(0.075)	(0.154)	(0.088)
2011	0.323*	0.282***	0.277†	0.272**	0.414**	0.379***	0.472**	0.436***
	(0.143)	(0.084)	(0.167)	(0.096)	(0.132)	(0.079)	(0.148)	(0.098)

Note: † indicates $p < 0.10$, * indicates $p < 0.05$, ** indicates $p < 0.01$, *** indicates $p < 0.001$. ATE1 is estimate of average treatment effect using one moment condition; ATE2 is estimate of average treatment effect using two moment conditions. Standard errors bootstrapped and clustered at district level.

Table 14: ELL Standardized Test

	2^{nd} Grade		3^{rd} Grade		4^{th} Grade		5^{th} Grade	
	ATE 1	ATE 2	ATE 1	ATE 2	ATE 1	ATE 2	ATE 1	ATE 2
2005	0.035	0.068	-0.060	-0.036	-0.019	-0.007	-0.073	-0.021
	(0.207)	(0.066)	(0.214)	(0.050)	(0.203)	(0.052)	(0.222)	(0.075)
2006	-0.060	-0.044	-0.003	0.021	-0.041	-0.040	-0.051	-0.019
	(0.209)	(0.062)	(0.213)	(0.062)	(0.199)	(0.054)	(0.200)	(0.057)
2007	-0.103	-0.068	-0.027	0.006	0.044	0.047	-0.047	-0.013
	(0.190)	(0.051)	(0.209)	(0.065)	(0.196)	(0.055)	(0.197)	(0.044)
2008	-0.060	-0.062	-0.046	-0.033	0.031	0.012	0.099	0.064
	(0.189)	(0.066)	(0.186)	(0.059)	(0.200)	(0.053)	(0.182)	(0.048)
2009	0.053	0.024	-0.0010	-0.026	0.135	0.153*	0.038	0.025
	(0.175)	(0.058)	(0.193)	(0.066)	(0.183)	(0.062)	(0.189)	(0.055)
2010	0.037	0.042	0.125	0.074	0.195	0.186**	0.207	0.186**
	(0.169)	(0.071)	(0.187)	(0.076)	(0.182)	(0.062)	(0.189)	(0.064)
2011	0.227	0.188**	0.161	0.147	0.278†	0.219***	0.276	0.217***
	(0.151)	(0.066)	(0.184)	(0.090)	(0.163)	(0.056)	(0.173)	(0.066)

Note: † indicates $p < 0.10$, * indicates $p < 0.05$, ** indicates $p < 0.01$, *** indicates $p < 0.001$. ATE1 is estimate of average treatment effect using one moment condition; ATE2 is estimate of average treatment effect using two moment conditions. Standard errors bootstrapped and clustered at district level.