ESSAYS ON THE QUALITY EDUCATION INVESTMENT ACT AND WEIGHTED QUANTILE REGRESSION

By

Paul Burkander

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Economics - Doctor of Philosophy

2014

ABSTRACT

ESSAYS ON THE QUALITY EDUCATION INVESTMENT ACT AND WEIGHTED QUANTILE REGRESSION

By

Paul Burkander

This dissertation contains three self-contained chapters. The first two are related in their analysis of California's Quality Education Investment Act (QEIA), with the former estimating its effect on student achievement and the latter exploiting an aspect of the law to estimate district preferences for resource allocation over low-performing schools. The final chapter considers the distributions of quantile regressors under complex random sampling.

Beginning in the 2007-08 school year, California's QEIA 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 4th grade increased by 0.32 SD in the population of California school-grade averages, and by the second fully-funded year 5th grade math scores improved by 0.37 SD. By the third fully-funded year of the program, math scores in 2nd grade were 0.30 SD higher in the distribution of California school-grade averages, and 0.29 SD higher in 3rd grade. Selected schools did not increase teacher experience, and had 4.4 to 4.8 fewer students in the first fully-funded year in 4th and 5th grade. In kindergarten through 3rd grade class sizes were reduced later and less dramatically, by 3 to 4.2 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.

This novel intervention also required school districts to rank their low-performing schools, the analysis of which constitutes my second chapter. Districts understood that higher ranked schools would be more likely to receive significant increases in funding to implement QEIA reforms. These rankings provide a unique revelation of district preferences for resource allocation across low-performing schools. Using a discrete-choice model, I estimate the school characteristics that districts ranked highly. I find descriptive evidence that districts were more likely to rank highly schools with a high percentage of students eligible for Free and Reduced Price Lunch, and which were repeatedly sanctioned under No Child Left Behind for failing to make Adequate Yearly Progress. I also find some evidence that districts ranked highly high schools that applied for an alternative program, in which they crafted their own reforms.

The final chapter, coauthored with Otávio Bartalotti, extends previous work that developed the asymptotic properties of quantile regression estimators under Standard Stratified sampling, to Variable Probability sampling. Formulas for the asymptotic variance and feasible estimators are provided. Simulation results are provided for both Standard Stratified and Variable Probability sampling. Simulation results confirm econometric theory by demonstrating that under exogenous stratification unweighted estimates preform well and are more efficient than weighted estimates. Under endogenous stratification and SS sampling, no estimate of standard errors performs best across coefficients, quantiles, and sample sizes. Under endogenous stratification and VP sampling, however, bootstrapped standard errors consistently perform well.

To Maggie and Ollie.

ACKNOWLEDGMENTS

The production of this research has benefited tremendously from faculty, friends, and family. Gary Solon pushed me to improve my analyses, without telling me exactly how. It is to this that I attribute much of my success as a researcher. In his labor and applied econometrics courses, Gary helped define for me the type of researcher I want to be. I appreciated stops by Todd Elder's office, the best of which ended with us both haphazardly scribbling various formulas on paper. In Todd's labor course it was typically when the final bell sounded that the class got most interesting, as he would hesitate to let us leave, insisting, as chalk dust flew and the board became increasingly filled, on showing us one more proof. Mike Conlin recognized one summer that I was not being productive enough, and insisted on meeting with me weekly, helping me find the pace of work that led to where I am.

The graduate student community at Michigan State owes much of its current strength to those who came before. Quentin Brummet, Otávio Bartalotti, Steve Dieterle, and others exemplified the type of graduate student and colleague that I wanted to be. I was also fortunate to be part of a tremendous cohort. Discussions with Michelle Maxfield, Brian Stacy, Paul Thompson, and arguments with Hassan Enayati were instrumental in my first year's success and beyond. Dan Litwok, Michael Bates, and Hassan are the type of people that, if they all stand up to go somewhere, I am very likely to follow, even against the weight of statistical evidence.

My family stood by me through my non-traditional path. Thank you, Nick, Janet, and John. To no one am I more grateful than Kri, whose endless support has carried me throughout.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

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

1.1 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

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

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 nonparametrically 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; 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 (API), 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 1.2 reviews the relevant literature; section 1.3 describes QEIA in greater detail; section 1.4 reviews the data used in this analysis; section 1.5 outlines the identification strategy; section 1.6 presents the results. Section 1.7 concludes.

1.2 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 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 et al. (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 et al. (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 et al. (2009), who note that, in using test scores from the following year, estimates may be attenuated. Jepsen et al. (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 et al. (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 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 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 et al. (2006), Jacob (2005)), shift resources towards marginal students (Neal et al. (2010)), and that teachers manipulate test results (Jacob et al. (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

[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.

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.

1.3 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 per class.[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

[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).

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 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.

[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.

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 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

[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 the California legislature (CDE (2010)), written 3 years after the selection process.

[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.

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 the federal 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 1.1, 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

[13] This restriction has two motivations: there are additional QEIA requirements for high schools, complicating the interpretation of the treatment, and beginning in 6th 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 from the author by request.

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 accel-

[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.

erated 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.

1.4 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

[16] Teachers for these excluded classes are included in the teacher experience category, in part so my measure 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.

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 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 2nd-5th grade math and English language arts tests.

[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.

Table 1.1 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 unfunded 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 did however tend to have nearly three years of experience.

From Table 1.1 it is apparent that the relative reduction in class size in funded schools in kindergarten through third grade is driven by increased class sizes in unfunded schools, while the relative 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.

[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.

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

[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.

[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 et al. (1952). See also Imbens et al. (2009) and Wooldridge (2010).

$$\tau_{\text{ATE}} = \mathrm{E}\left(\frac{T_i y_i}{p_i} - \frac{(1-T_i)y_i}{(1-p_i)}\right) = \mathrm{E}\left(\frac{(T_i - p_i)y_i}{(1-p_i)p_i}\right) = \mathrm{E}(y_{i1} - y_{i0}) \qquad (1.1)$$

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. $\mathrm{E}(y_{i1} - y_{i0})$ is the Average Treatment Effect: it captures the average change in outcome caused by QEIA. The parameter in equation (1.1) 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 1 million 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., $\mathrm{E}(y_j|T,p) = \mathrm{E}(y_j|p)$, $j \in \{0,1\}$). 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

[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] As a result, my estimates have SE of $\frac{\sqrt{p_i(1-p_i)}}{1000}$, which at its largest is 0.0005. Given this precision, I refer to my estimates as the true probabilities of selection. The actual random numbers assigned to districts are made publicly available by CDE. Using these, and the district rankings, my simulation of the selection process perfectly predicts funded schools.

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 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:

$$\frac{1}{\alpha(1-\alpha)} = 2 * \mathrm{E}\left[\frac{1}{p_i(1-p_i)} \middle| \frac{1}{p_i(1-p_i)} < \frac{1}{\alpha(1-\alpha)}\right]. \tag{1.2}$$

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 (1.2), 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 1.1 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 (1.1) 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 et al. (2004), Hitomi et al. (2008), Prokhorov et al. (2009), Han et al. (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 could 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 $d_i = \mathbf{X}_i(T_i - p_i)$. This is equivalent to what Qian et al. (1999) call an augmented GMM estimator, in which efficiency gains are achieved with moment conditions that are not a function of the parameter of interest. I provide results using the sample analog of (1.1), 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 1.2, the sample moment conditions implied by $\mathrm{E}(d_i) = 0$ are all quite close to zero.

1.6 Results

1.6.1 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 assumptions that might not be satisfied. Each regression has a full set of year dummies, excluding 2005, 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, 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.

[24] The full set of results is available from the author by request.

As Table 1.3 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 5th 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 math, and from 0.19 ($p < 0.01$) to 0.30 ($p < 0.001$) 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 same school from the beginning of the school year through the time assessments were administered. A causal effect of QEIA on the composition of students in a school could suggest that it did not benefit particular students, but rather affected the likelihood that better students remained in the school. However, the estimated effects are small relative to the baseline of about 90% (see Table 1.1), precise only in some specifications, and then only significant at the 5% level. More importantly, as indicated below, results from IPW estimates suggest no change in student characteristics.

1.6.2 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 1.4 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 4.35 students per class, an estimate that is

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

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 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 1.4. 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 1.5. Using two sets of moment conditions,[26] QEIA increased API scores in funded schools by 0.35 standard deviations ($p < 0.001$) by 2011, with respect to the population of all elementary school-level averages. The effect for Hispanic students is significantly larger than for all students by 2011 ($p = 0.068$), as is the effect for low-SES students ($p = 0.075$). From 2008 onward there is a clear pattern of funded schools improving over unfunded schools.

[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.

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 from the beginning of the school year through the time assessments were administered.

Table 1.6 displays the results for FRPL,[27] percent of students whose parents have a college 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 1.7 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

[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.

[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.

there is a valid score, and the number who take the regular standardized test. Table 1.8 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 1.9 and 1.10 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 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 0.92 years greater in funded schools in 2009 ($p < 0.1$).

Table 1.10 lists the change in proportion of teachers who are probationary, tenured,

[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.

and long-term substitutes. The differences between funded and unfunded schools in the proportion of probationary teachers before 2008, significant at the 10%, and 5% levels, 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 significant evidence of fewer tenured teachers in funded schools ($p < 0.05$), and more long-term substitutes ($p < 0.10$).

Tables 1.11 and 1.12 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 3.0 to 4.2 fewer students in those grades in funded schools. As mentioned above, class size data are not available in 2010, and 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 4.8 ($p < 0.001$) and 4.4 ($p < 0.01$) students per class in 2009, respectively, and by 5.5 ($p < 0.001$) to 6.1 ($p < 0.001$) 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 1.13 and 1.14 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 0.29 standard deviations higher in funded schools, with respect to the population of grade-level averages ($p < 0.001$, 0.13 student-level standard deviations). The 3rd grade math scores increase one year earlier, in 2010, by 0.18 standard deviations ($p < 0.05$, 0.08 student-level standard deviations), and by 0.29 standard deviations by 2011 ($p < 0.01$, 0.12 student-level standard deviations).

Math scores in 4th grade improve earlier; by 2009 they show an increase of 0.32 standard deviations($p < 0.001$, 0.15 student-level standard deviations), 0.40 ($p < 0.001$, 0.17 student-level standard deviations) in 2010, and level off in 2011 at 0.40 ($p < 0.001$, 0.17 student-level standard deviations). Interestingly, 5th grade math scores do not begin improving until 2010, at which time they were 0.37 standard deviations higher in funded schools ($p < 0.001$, 0.17 student-level standard deviations), and by 2011 they were 0.42 standard deviations higher ($p < 0.001$, 0.19 student-level standard deviations)

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 0.20 standard deviations higher in funded schools ($p < 0.01$, 0.09 student-level standard deviations), and in 5th grade they are 0.23 standard deviations higher ($p < 0.001$, 0.11 student-level standard deviations).

To better understand the effects of increased exposure to QEIA, Figure 1.2 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 3^{rd} grade class size in 2010 is the average of 3^{rd} grade class size in 2009 and 2011.

As the figure suggests, consecutive years of smaller classes do 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 2^{nd} and 3^{rd} grade are delayed in the same manner that the relative change in test scores is delayed, it seems likely that the effect is driven by class size. Otherwise, there would have to be some reason that professional training and accountability pressure were also delayed. Though professional training is not observed, test scores in each of the first three years counted towards the achievement target, and it therefore seems unlikely that schools would not respond to it until the third year of the program.

1.7 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), Hirano, Imbens, and Ridder (2003), and Qian et al. (1999)

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 4th grade increased by 0.32 standard deviations in the population of school-grade averages, and by the second fully-funded year 5th grade math scores improved by 0.36 standard deviations. The improvement in test scores in 2nd and 3rd 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 2nd grade were 0.28 standard deviations higher in the distribution of school-grade averages, and 0.27 standard deviations higher 3rd 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.

To estimate the cost effectiveness of QEIA, I consider the expense that would be incurred by replicating the intervention for each of three cohorts of students, those in 2nd, 3rd, and 4th grade in 2009. As an upper-bound, I consider the expenses that were incurred in addition to the per-pupil allocation as fixed costs. These include an annual expense of $2 million for county superintendents, \$1.177 million annually for CDE staff, and a one-time expense of \$5 million for regional support offices. The lower bound treats these expenses as variable costs. I use the OMB nominal interest rate on 3-year treasury bills from 2009-2011 to express the PDV of costs in 2008, the first year of the program. I average math and English language arts scores, and express all effects in student-level standard deviations, with respect to the population of all California students.

For the sake of comparison, I compare the cost-benefit estimates to the cost of achieving the same class reduction as in Project STAR. The average teacher salary in California in 2008 was \$64,424 (U.S. Department of Education (2009)). Following Podgursky (2006), I allow for benefits to account for 20% of compensation, and thus the cost of an additional teacher in 2008 is \$80,530. The one-year cost of the same class size reduction as in Project STAR is therefore $(\frac{1}{15}-\frac{1}{23})*80,530 \approx \$1,867$ per student. Comparing the change in student test scores caused by Project STAR to those caused by QEIA is complicated by a lack of common measure. Under the strong assumption that a standard deviation with respect to a select sample of Tennessee students in kindergarten through third grade is comparable to a standard deviation with respect to the population of all California students in 2$^{\text{nd}}$ through 4$^{\text{th}}$ grade, a class-size reduction of this magnitude would result in gains of 0.20 to 0.28 standard deviations (Krueger (1999)).

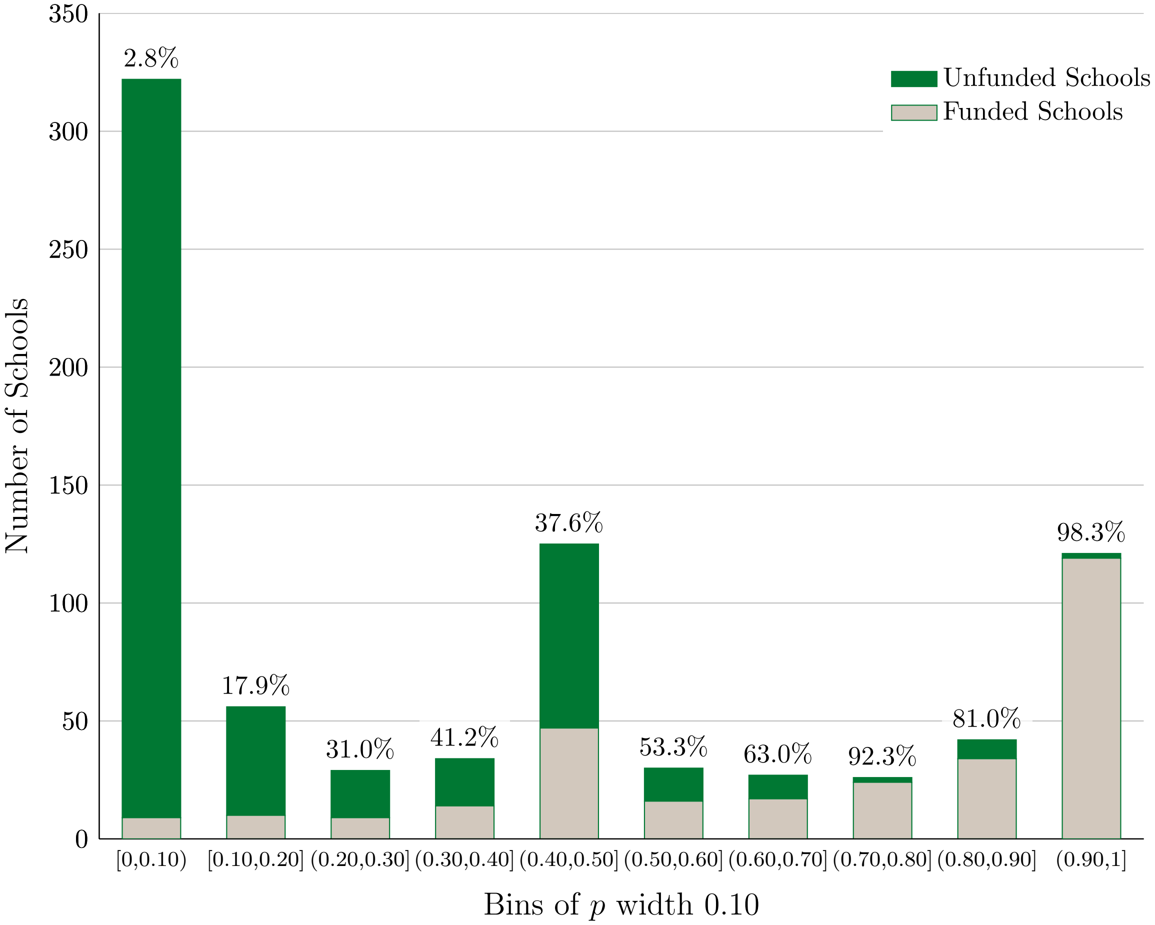
Table 1.15 shows the results of this exercise for each of three cohorts of QEIA students: those in 2$^{\text{nd}}$, 3$^{\text{rd}}$, and 4$^{\text{th}}$ grade in 2009. Where the class size requirements of QEIA duplicated the existing class size reduction program, QEIA had no effect, and was of course not cost effective. In other years and grades, the upper-bound cost per standard deviation gain in test scores is comparable to Project STAR in the first and third year of each program, while the second year of Project STAR lies closer to the lower bound. Project STAR's much more dramatic, and much more expensive, reduction in class sizes is estimated to have only achieved concomitant dramatic increase in student achievement in its second year.

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. Though potentially cost-effective relative to Project STAR in years in which it was effective, QEIA was hampered by overlap with existing policies that caused it to be completely ineffective in certain years and grades. The unique design of QEIA, which accommodated district preferences for resource allocation across schools, State budget constraints and preferences for reform design, also allows for non-parametric identification of its causal effect. Were more policies to follow this design, our understanding of the effectiveness of various reforms could be dramatically improved.

APPENDICES

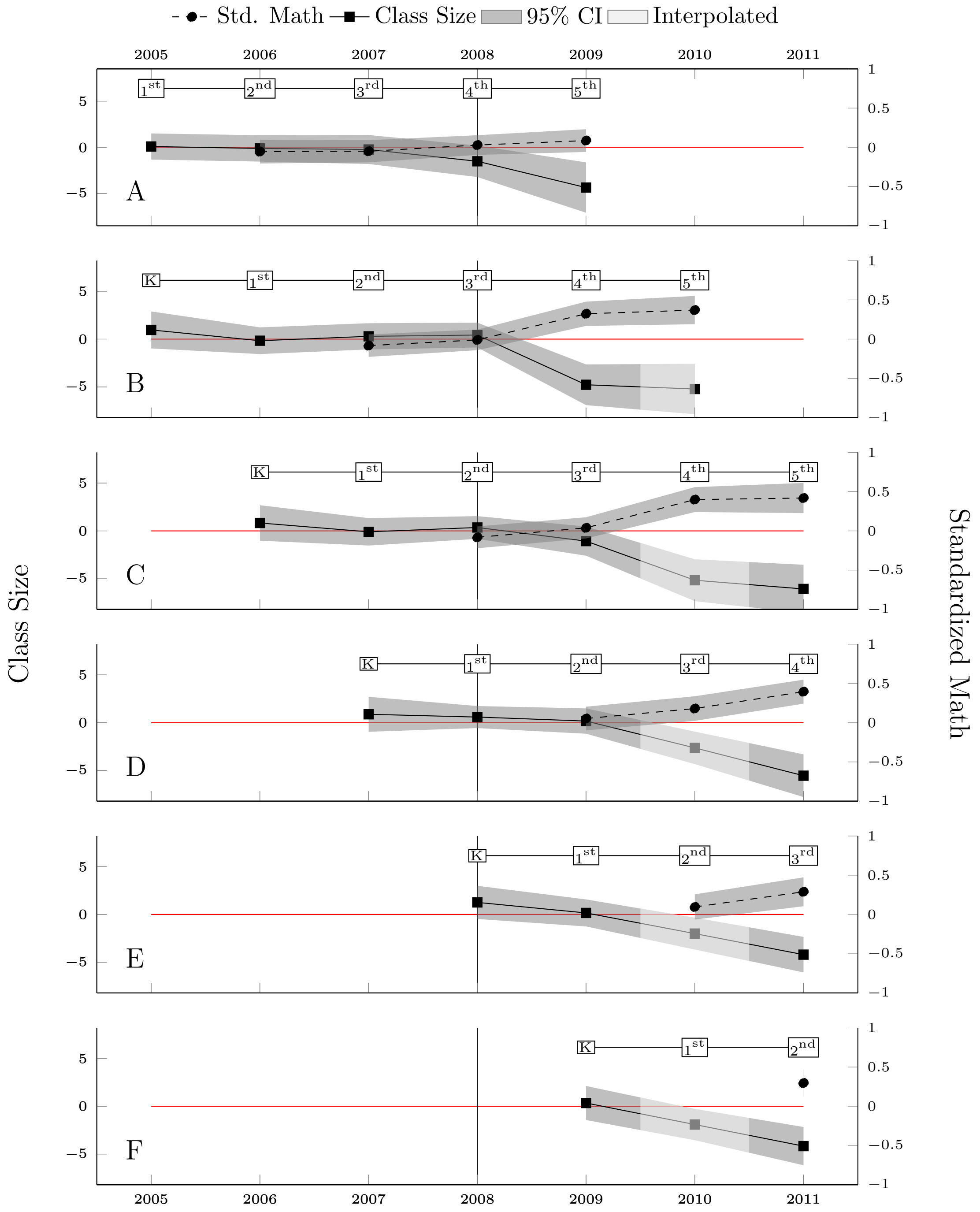
APPENDIX A - FIGURES

Figure 1.1: 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 1.2: 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 4th grade class size is average of 2009 4th grade class size and 2011 4th grade class size. Only grades 2 and above are tested.

APPENDIX B - TABLES

Table 1.1: Descriptives, Elementary Regular QEIA Schools 2007

	All Elem.	$p_i \in [0,1]$, 2007		$p_i \in [0.10, 0.90]$, 2007		$p_i \in [0.10, 0.90]$, 2011	
	2007	Unfunded	Funded	Unfunded	Funded	Unfunded	Funded
Avgerage Class Size	21.86	21.93	21.96	22.14	22.12	25.03	20.34
	(3.54)	(1.87)	(2.04)	(1.75)	(2.05)	(3.19)	(2.19)
Class Size Kindergarten	20.64	20.24	20.41	20.28	20.70	23.97	20.23
	(4.14)	(3.39)	(3.78)	(3.46)	(4.23)	(4.31)	(3.48)
Class Size 1^{st} Grade	19.32	19.29	19.13	19.33	19.21	23.78	19.54
	(1.85)	(1.40)	(1.31)	(1.47)	(1.29)	(4.07)	(2.27)
Class Size 2^{nd} Grade	19.14	18.84	18.97	18.76	19.00	23.85	19.36
	(1.87)	(1.52)	(1.40)	(1.58)	(1.34)	(4.24)	(2.54)
Class Size 3^{rd} Grade	19.89	19.42	19.60	19.76	19.60	23.95	19.54
	(3.17)	(2.71)	(3.01)	(3.18)	(3.08)	(4.49)	(2.63)
Class Size 4^{th} Grade	28.47	28.01	28.18	28.21	28.56	28.14	22.37
	(4.23)	(3.82)	(3.80)	(3.57)	(3.81)	(4.30)	(3.97)
Class Size 5^{th} Grade	28.89	28.28	28.51	28.70	28.62	28.36	22.25
	(4.15)	(3.72)	(3.77)	(3.57)	(3.95)	(4.53)	(3.35)
Average Experience	13.01	11.92	11.61	12.38	11.70	13.68	12.95
	(3.95)	(3.17)	(3.35)	(3.10)	(3.38)	(3.56)	(3.17)
TEI Relative	-0.04	-0.17	-0.28	-0.07	-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)	(0.22)	(0.22)	(0.22)	(0.22)
Std. API	0.00	-1.10	-1.25	-1.13	-1.15	-1.11	-0.77
	(1.00)	(0.48)	(0.48)	(0.44)	(0.47)	(0.56)	(0.61)
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)	(0.47)	(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	0.51	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	445.58	420.82	405.13	378.30
	(192.56)	(195.68)	(179.70)	(164.11)	(154.59)	(138.53)	(135.89)
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	0.26	0.25	0.47	0.43
	(0.24)	(0.23)	(0.23)	(0.25)	(0.23)	(0.31)	(0.29)
Average Experience New Teachers	3.21	2.76	3.13	3.17	3.13	6.44	5.87
	(3.64)	(3.18)	(3.12)	(3.47)	(3.07)	(4.88)	(4.48)
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	198	171	198	171

Note: Table lists means and standard deviations in parenthesis. Funded and unfunded includes all elementary schools participating in the regular QEIA program.

Table 1.2: Sample Moment Conditions

Variable	Mean	S.D.
(Funded$_i$-p_i)	-0.00	(0.45)
(2007 Met Growth Target)(Funded$_i$-p_i)	-0.02	(0.36)
(2007 Proportion FRPL)(Funded$_i$-p_i)	-0.01	(0.39)
(2007 Student Enrollment)(Funded$_i$-p_i)	-5.80	(203.35)
(2007 Std. API)(Funded$_i$-p_i)	0.00	(0.53)
(2007 Proportion Hispanic)(Funded$_i$-p_i)	-0.01	(0.36)
(2007 English Language Learners)(Funded$_i$-p_i)	0.00	(0.26)
(2007 Migrant)(Funded$_i$-p_i)	-0.01	(0.06)

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

Table 1.3: Select Regression Results

	(1)	(2)	(3)	(4)	(5)
			Full Sample		
Avg. Class Size	-4.45***	-4.76***	-4.54***	-4.62***	-4.58***
	(0.56)	(0.56)	(0.40)	(0.46)	(0.42)
Experience	-0.83**	-0.73†	-0.90**	-0.93**	-0.98***
	(0.26)	(0.43)	(0.27)	(0.28)	(0.25)
Std. API	0.14†	0.41***	0.08	0.26***	0.22***
	(0.08)	(0.07)	(0.07)	(0.06)	(0.05)
5^{th} Grade Math	0.27**	0.46***	0.25**	0.37***	0.34***
	(0.09)	(0.09)	(0.08)	(0.08)	(0.08)
5^{th} Grade ELA	0.12*	0.30***	0.06	0.17**	0.16**
	(0.06)	(0.07)	(0.06)	(0.05)	(0.05)
Enrolled Since Previous Year	0.00	0.01	0.01*	0.00	0.01*
	(0.01)	(0.01)	(0.00)	(0.01)	(0.00)
			$p_i \in [.10, .90]$		
Avg. Class Size	-4.69***	-4.68***	-4.58***	-4.90***	-4.80***
	(0.45)	(0.50)	(0.41)	(0.45)	(0.45)
Experience	-0.74†	-0.61	-0.73†	-0.59†	-0.67*
	(0.38)	(0.45)	(0.40)	(0.34)	(0.34)
Std. API	0.34***	0.40***	0.26***	0.37***	0.34***
	(0.06)	(0.07)	(0.06)	(0.06)	(0.06)
5^{th} Grade Math	0.43***	0.44***	0.41***	0.49***	0.47***
	(0.09)	(0.09)	(0.09)	(0.10)	(0.10)
5^{th} Grade ELA	0.26***	0.30***	0.19**	0.26***	0.24***
	(0.06)	(0.07)	(0.06)	(0.06)	(0.06)
Enrolled Since Previous Year	0.01	0.01	0.01*	0.01	0.01*
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
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 1.4: 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	0.10	0.22	0.01	0.00	-0.06	-0.20	-0.02	-0.00
	(3.93)	(0.84)	(0.18)	(0.04)	(2.16)	(0.58)	(0.10)	(0.10)
2006	0.17	0.25	0.01	0.00	-0.21	-0.29	-0.09	-0.05
	(3.97)	(0.81)	(0.17)	(0.04)	(2.24)	(0.64)	(0.11)	(0.10)
2007	0.05	0.12	-0.01	-0.02	-0.69	-0.76	-0.30*	-0.23†
	(3.90)	(0.81)	(0.17)	(0.04)	(2.16)	(0.69)	(0.13)	(0.12)
2008	-0.18	0.08	0.01	0.00	-0.41	-0.55	-0.28*	-0.23†
	(3.92)	(0.68)	(0.18)	(0.04)	(2.19)	(0.66)	(0.13)	(0.14)
2009	-1.61	-1.45†	0.01	0.00	-0.45	-0.45	-0.25†	-0.18
	(3.82)	(0.80)	(0.18)	(0.04)	(2.33)	(0.72)	(0.13)	(0.13)
2010					-0.64	-0.63		
					(2.47)	(0.72)		
2011	-4.65	-4.35***	-0.01	-0.00	-0.57	-0.58	-0.02	0.01
	(4.05)	(0.95)	(0.18)	(0.04)	(2.39)	(0.64)	(0.10)	(0.12)

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 1.5: 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	-0.01	-0.03	0.08	0.04	-0.08	-0.07
	(0.22)	(0.05)	(0.21)	(0.07)	(0.20)	(0.06)
2006	0.02	-0.03	0.08	0.02	-0.04	-0.07
	(0.22)	(0.05)	(0.20)	(0.06)	(0.19)	(0.05)
2007	0.01		0.05	0.04	-0.04	-0.01
	(0.21)		(0.19)	(0.06)	(0.19)	(0.04)
2008	0.07	0.03	0.15	0.09	0.05	0.03
	(0.19)	(0.04)	(0.17)	(0.07)	(0.16)	(0.06)
2009	0.16	$0.10^{\dagger}$	0.22	$0.15^{\dagger}$	0.18	0.12
	(0.20)	(0.06)	(0.18)	(0.08)	(0.17)	(0.08)
2010	0.31	0.23***	0.47*	0.35***	0.39*	0.31***
	(0.20)	(0.06)	(0.19)	(0.09)	(0.19)	(0.08)
2011	0.43*	0.35***	0.45*	0.41***	0.48**	0.41***
	(0.19)	(0.07)	(0.18)	(0.08)	(0.18)	(0.08)

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 1.6: Demographics

	FRPL		Parents College Degree		Enrolled Since Previous Year	
	ATE 1	ATE 2	ATE 1	ATE 2	ATE 1	ATE 2
2005	-0.02	0.01	0.01	0.01	0.01	0.01
	(0.16)	(0.03)	(0.02)	(0.01)	(0.16)	(0.03)
2006	-0.03	-0.01	0.01	0.01	0.02	0.02
	(0.15)	(0.03)	(0.02)	(0.01)	(0.16)	(0.03)
2007	-0.02		0.01	0.00	0.01	0.01
	(0.15)		(0.02)	(0.01)	(0.16)	(0.03)
2008	-0.02	0.01	0.01	-0.00	0.02	0.01
	(0.16)	(0.03)	(0.01)	(0.01)	(0.17)	(0.03)
2009	-0.03	-0.01	0.01	0.00	0.02	0.01
	(0.16)	(0.03)	(0.02)	(0.01)	(0.16)	(0.03)
2010	-0.03	-0.01	0.01	0.01	0.02	0.01
	(0.16)	(0.03)	(0.02)	(0.01)	(0.16)	(0.03)
2011	-0.05	-0.03	0.01	0.00	0.02	0.02
	(0.16)	(0.04)	(0.01)	(0.01)	(0.16)	(0.03)

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 1.7: Demographics Continued

	Enrollment		Black		Hispanic	
	ATE 1	ATE 2	ATE 1	ATE 2	ATE 1	ATE 2
2005	-32.38	-5.38	0.02	0.02	-0.03	-0.00
	(92.10)	(20.44)	(0.03)	(0.01)	(0.13)	(0.03)
2006	-25.03	-1.19	0.02	0.02	-0.03	-0.00
	(84.61)	(17.32)	(0.03)	(0.01)	(0.13)	(0.03)
2007	-24.21		0.02	0.01	-0.02	
	(79.61)		(0.03)	(0.01)	(0.13)	
2008	-34.45	-11.82	0.02	0.01	-0.02	0.00
	(83.42)	(17.90)	(0.02)	(0.01)	(0.14)	(0.03)
2009	-30.22	-7.30	0.02	0.01	-0.02	0.00
	(76.39)	(16.19)	(0.02)	(0.01)	(0.14)	(0.03)
2010	-25.21	-6.93	0.01	0.01	-0.01	0.00
	(78.41)	(17.30)	(0.02)	(0.01)	(0.13)	(0.03)
2011	-31.41	-7.36	0.01	0.01	-0.02	0.00
	(76.97)	(18.15)	(0.02)	(0.01)	(0.14)	(0.03)

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 1.8: Test Taking

	Valid Scores		Number Low SES Scores		Number Hispanic Scores		Prop. CST	
	ATE 1	ATE 2	ATE 1	ATE 2	ATE 1	ATE 2	ATE 1	ATE 2
2005	-25.25	-2.19	-28.88	-0.73	-21.88	2.85	0.03	0.01
	(79.32)	(18.25)	(74.56)	(17.27)	(59.94)	(14.73)	(0.19)	(0.04)
2006	-17.20	3.96	-21.36	2.47	-14.61	8.95	0.03	0.01
	(73.57)	(14.75)	(69.07)	(13.33)	(57.90)	(11.88)	(0.17)	(0.03)
2007	-22.40	-1.28	-25.92	-0.29	-20.11	3.99	0.02	0.01
	(65.98)	(13.21)	(62.82)	(11.86)	(55.65)	(10.46)	(0.18)	(0.03)
2008	-23.29	-9.40	-25.84	-5.47	-21.02	-3.41	0.01	0.01
	(68.51)	(15.06)	(62.13)	(14.03)	(54.45)	(12.31)	(0.18)	(0.03)
2009	-15.96	-2.69	-21.52	-3.30	-14.70	2.88	0.03	-0.00
	(67.14)	(14.40)	(61.05)	(13.87)	(53.18)	(12.12)	(0.18)	(0.03)
2010	-19.13	-2.09	-27.98	-4.49	-14.93	5.70	0.03	-0.00
	(69.82)	(15.35)	(65.13)	(14.96)	(56.92)	(12.86)	(0.17)	(0.03)
2011	-22.24	-0.65	-33.00	-7.19	-21.41	4.05	0.03	0.00
	(68.36)	(16.16)	(63.69)	(16.20)	(55.00)	(13.62)	(0.17)	(0.04)

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 1.9: Teacher Mobility

	New to school		New to School, Not New to Dist.		Avg. Experience New Teachers	
	ATE 1	ATE 2	ATE 1	ATE 2	ATE 1	ATE 2
2005	0.07	0.06	0.08	0.07	0.80	0.64
	(0.08)	(0.04)	(0.06)	(0.04)	(0.79)	(0.56)
2006	0.05	0.04	0.04	0.03	0.40	0.25
	(0.07)	(0.04)	(0.05)	(0.04)	(0.71)	(0.49)
2007	-0.01	-0.01	-0.04	-0.04	-0.47	-0.58
	(0.08)	(0.05)	(0.07)	(0.05)	(0.91)	(0.66)
2008	0.06	0.06	0.03	0.03	0.52	0.35
	(0.07)	(0.04)	(0.05)	(0.04)	(0.81)	(0.61)
2009	0.08	0.07†	0.08	0.07†	1.04	0.92†
	(0.07)	(0.04)	(0.06)	(0.04)	(0.80)	(0.55)
2010	0.03	0.04	0.03	0.05	0.10	0.37
	(0.09)	(0.04)	(0.08)	(0.04)	(1.04)	(0.58)
2011	-0.05	0.00	-0.06	-0.01	-0.65	-0.04
	(0.10)	(0.05)	(0.10)	(0.05)	(1.44)	(0.76)

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 1.10: Teacher Composition

	Probationary		Tenured		Long-term Substitute	
	ATE 1	ATE 2	ATE 1	ATE 2	ATE 1	ATE 2
2005	0.05	0.05*	-0.04	-0.04	0.00	0.00
	(0.04)	(0.02)	(0.13)	(0.05)	(0.01)	(0.01)
2006	0.04	0.03	-0.02	-0.03	0.00	-0.00
	(0.04)	(0.02)	(0.14)	(0.05)	(0.01)	(0.01)
2007	0.04	0.04†	-0.02	-0.02	0.00	0.01
	(0.03)	(0.02)	(0.13)	(0.05)	(0.01)	(0.01)
2008	0.04	0.04*	-0.05	-0.05	0.01	0.01
	(0.03)	(0.02)	(0.14)	(0.04)	(0.02)	(0.02)
2009	0.03	0.03*	-0.07	-0.07*	0.03†	0.03†
	(0.03)	(0.02)	(0.14)	(0.04)	(0.02)	(0.02)
2010	0.03†	0.04**	-0.04	-0.04	0.00	-0.00
	(0.02)	(0.01)	(0.15)	(0.04)	(0.01)	(0.01)
2011	0.03	0.03*	-0.05	-0.05	0.01	0.00
	(0.02)	(0.01)	(0.17)	(0.04)	(0.01)	(0.01)

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 1.11: 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	0.33	0.96	-0.18	0.09	0.08	0.24
	(3.97)	(0.99)	(3.53)	(0.72)	(3.62)	(0.70)
2006	0.28	0.83	-0.93	-0.17	-0.27	-0.12
	(3.84)	(0.95)	(3.58)	(0.71)	(3.61)	(0.73)
2007	0.24	0.88	-0.41	-0.10	-0.30	0.29
	(3.83)	(0.93)	(3.40)	(0.73)	(3.49)	(0.70)
2008	0.50	1.26	-0.01	0.59	-0.21	0.36
	(3.88)	(0.88)	(3.47)	(0.59)	(3.42)	(0.61)
2009	0.15	0.34	-0.50	0.16	-0.16	0.17
	(3.77)	(0.91)	(3.62)	(0.72)	(3.47)	(0.68)
2010						
2011	-3.26	-3.04**	-4.70	-3.98***	-4.39	-4.17***
	(4.36)	(1.10)	(4.20)	(0.94)	(4.06)	(1.02)

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 1.12: 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	0.21	-0.15	0.86	0.30	1.84	-0.45
	(3.70)	(0.74)	(5.21)	(1.21)	(5.70)	(1.36)
2006	-0.27	0.11	0.26	-0.25	0.42	-0.68
	(3.64)	(0.71)	(5.10)	(1.08)	(5.44)	(1.20)
2007	-0.16	-0.23	-0.40	0.50	0.26	-0.14
	(3.62)	(0.80)	(4.94)	(1.23)	(5.55)	(1.29)
2008	-0.08	0.42	-1.68	-1.51†	-1.89	-1.34
	(3.61)	(0.66)	(4.73)	(0.87)	(4.79)	(0.98)
2009	-1.33	-1.07	-4.63	-4.77***	-3.26	-4.36**
	(3.65)	(0.78)	(4.69)	(1.09)	(5.09)	(1.39)
2010						
2011	-4.12	-4.20***	-4.79	-5.54***	-5.48	-6.07***
	(4.04)	(0.95)	(4.80)	(1.14)	(4.85)	(1.29)

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 1.13: 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	-0.01	-0.02	-0.09	-0.08	-0.08	-0.07	-0.10	-0.09
	(0.22)	(0.08)	(0.19)	(0.06)	(0.20)	(0.06)	(0.20)	(0.08)
2006	-0.04	-0.06	0.02	0.00	-0.05	-0.07	-0.06	-0.05
	(0.23)	(0.08)	(0.18)	(0.06)	(0.19)	(0.07)	(0.21)	(0.06)
2007	-0.08	-0.08	-0.07	-0.05	-0.02	-0.02	-0.01	0.00
	(0.19)	(0.07)	(0.21)	(0.07)	(0.17)	(0.06)	(0.18)	(0.06)
2008	-0.06	-0.08	-0.01	-0.01	0.04	0.03	0.16	0.11
	(0.18)	(0.07)	(0.19)	(0.07)	(0.17)	(0.06)	(0.17)	(0.07)
2009	0.10	0.05	0.07	0.04	0.30*	0.32***	0.14	0.09
	(0.17)	(0.08)	(0.18)	(0.07)	(0.15)	(0.08)	(0.16)	(0.07)
2010	0.16	0.10	0.23	0.18*	0.40**	0.40***	0.45*	0.37***
	(0.19)	(0.08)	(0.18)	(0.08)	(0.16)	(0.08)	(0.18)	(0.09)
2011	0.36*	0.30***	$0.32^{\dagger}$	0.29**	0.45**	0.40***	0.47**	0.42***
	(0.15)	(0.09)	(0.18)	(0.09)	(0.15)	(0.08)	(0.16)	(0.10)

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 1.14: 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	0.10	0.09	-0.01	-0.02	0.03	0.01	-0.01	0.00
	(0.21)	(0.06)	(0.20)	(0.04)	(0.20)	(0.05)	(0.21)	(0.07)
2006	0.00	-0.02	0.05	0.05	0.00	-0.03	-0.01	-0.01
	(0.21)	(0.06)	(0.20)	(0.06)	(0.20)	(0.05)	(0.21)	(0.06)
2007	-0.06	-0.06	0.01	0.01	0.09	0.06	-0.02	-0.01
	(0.19)	(0.05)	(0.22)	(0.07)	(0.20)	(0.05)	(0.19)	(0.04)
2008	-0.03	-0.06	0.00	-0.02	0.07	0.03	0.14	0.08
	(0.18)	(0.07)	(0.20)	(0.06)	(0.19)	(0.05)	(0.19)	(0.05)
2009	0.09	0.03	0.05	-0.01	0.17	0.16*	0.07	0.03
	(0.18)	(0.06)	(0.21)	(0.07)	(0.19)	(0.06)	(0.19)	(0.05)
2010	0.10	0.07	0.19	0.10	0.24	0.20**	0.26	0.20**
	(0.17)	(0.07)	(0.19)	(0.08)	(0.19)	(0.06)	(0.19)	(0.06)
2011	0.26	0.20**	0.23	0.18*	0.33*	0.24***	$0.32^{\dagger}$	0.23***
	(0.16)	(0.07)	(0.19)	(0.09)	(0.17)	(0.06)	(0.19)	(0.07)

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 1.15: Cost-Benefit Analysis

Grade	PDV	SD gain			Cost per SD		
in 2009		2009	2010	2011	2009	2010	2011
	Upper Bound						
4rd	\$2,975	0.11	0.13	NA	\$27,047	\$22,885	NA
3th	\$3,609	0	0.13	0.15	∞	\$27,764	\$24,062
2th	\$3,235	0	0	0.14	∞	∞	\$23,107
	Lower Bound						
4rd	\$2,059	0.11	0.13	NA	\$18,720	\$15,838	NA
3th	\$2,506	0	0.13	0.15	∞	\$19,277	\$16,707
2th	\$2,132	0	0	0.14	∞	∞	\$15,228
	Project STAR						
Pre-K	\$5,247	0.20	0.28	0.22	\$26,239	\$18,742	\$23,853

Note: Upper bound of QEIA costs assumes administrative expenses are all fixed costs, while the lower bound assumes they are variable costs. Estimates of cost of implementing Project STAR class size reduction assume cost of additional teacher is \$80,530. Test score gains from Project STAR class size reduction are from Krueger (1999).

BIBLIOGRAPHY

BIBLIOGRAPHY

Angrist, J.D. and V. Lavy (1999). Using Maimonides' rule to estimate the effect of class size on scholastic achievement. *The Quarterly Journal of Economics* 114.2, pp. 533–575.

Balcom, Fred (Feb. 2007). *Quality Education Investment Act (QEIA) of 2006*. http://www.cde.ca.gov/fg/fo/r16/documents/qeia07present.ppt. California Department of Education.

Barrett, Nathan, JS Butler, and Eugenia F Toma (2012). Do less effective teachers choose professional development? Does it matter? *Evaluation Review* 36.5, pp. 346–374.

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

CDE (Jan. 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. California Department of Education.

Chetty, Raj, John N Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan (2011a). How does your kindergarten classroom affect your earnings? Evidence from Project STAR. *The Quarterly Journal of Economics* 126.4, pp. 1593–1660.

Chetty, Raj, John N Friedman, and Jonah E Rockoff (2011b). The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood.

Chiang, Hanley (2009). How accountability pressure on failing schools affects student achievement. *Journal of Public Economics* 93.9, pp. 1045–1057.

Chingos, Matthew M (2012). The impact of a universal class-size reduction policy: Evidence from Florida's statewide mandate. *Economics of Education Review* 31.5, pp. 543–562.

Cohodes, Sarah, David Deming, Jennifer Jennings, and Christophe Jencks (2013). School Accountability, Postsecondary Attainment and Earnings. *NBER Working Paper*.

Crump, Richard K, V Joseph Hotz, Guido W Imbens, and Oscar A Mitnik (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika* 96.1, pp. 187–199.

Dieterle, Steven (2013). *Class-size Reduction Policies and the Quality of Entering Teachers.*

Figlio, David N and Lawrence S Getzler (2006). Accountability, ability and disability: Gaming the system? *Advances in Applied Microeconomics* 14, pp. 35–49.

Greenwald, R., L.V. Hedges, and R.D. Laine (1996). The effect of school resources on student achievement. *Review of Educational Research* 66.3, pp. 361–396.

Hahn, Jinyong (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, pp. 315–331.

Han, Chirok and Beomsoo Kim (2011). A GMM interpretation of the paradox in the inverse probability weighting estimation of the average treatment effect on the treated. *Economics Letters* 110.2, pp. 163–165.

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

Hanushek, E.A. (1986). The economics of schooling: Production and efficiency in public schools. *Journal of Economic Literature* 24.3, pp. 1141–1177.

Hanushek, E.A. (1996). A more complete picture of school resource policies. *Review of Educational Research* 66.3, pp. 397–409.

Hanushek, E.A. (1997). Assessing the effects of school resources on student performance: An update. *Educational Evaluation and Policy Analysis* 19.2, p. 141.

Hanushek, E.A. (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* 21.2, p. 143.

Henmi, Masayuki and Shinto Eguchi (2004). A paradox concerning nuisance parameters and projected estimating functions. *Biometrika* 91.4, pp. 929–941.

Hirano, Keisuke, Guido W Imbens, and Geert Ridder (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* 71.4, pp. 1161–1189.

Hitomi, Kohtaro, Yoshihiko Nishiyama, and Ryo Okui (2008). A puzzling phenomenon in semiparametric estimation problems with infinite-dimensional nuisance parameters. *Econometric Theory* 24.06, pp. 1717–1728.

Horvitz, Daniel G and Donovan J Thompson (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* 47.260, pp. 663–685.

Hoxby, C.M. (2000). The effects of class size on student achievement: New evidence from population variation. *Quarterly Journal of Economics* 115.4, pp. 1239–1285.

Imbens, Guido W and Jeffrey M Wooldridge (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature* 47.1, pp. 5–86.

Jacob, Brian A (2005). Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago Public Schools. *Journal of Public Economics* 89.5, pp. 761–796.

Jacob, Brian A and Steven D Levitt (2003). Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *The Quarterly Journal of Economics* 118.3, pp. 843–877.

Jepsen, Christopher and Steven Rivkin (2009). Class Size Reduction and Student Achievement: The Potential Tradeoff between Teacher Quality and Class Size. *Journal of Human Resources* 44.1, pp. 223–250.

Krueger, A.B. (1999). Experimental estimates of education production functions. *The Quarterly Journal of Economics* 114.2, pp. 497–532.

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

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. *The Economic Journal* 111.468, pp. 1–28.

Neal, Derek and Diane Whitmore Schanzenbach (2010). Left behind by design: Proficiency counts and test-based accountability. *The Review of Economics and Statistics* 92.2, pp. 263–283.

Nye, B., L.V. Hedges, and S. Konstantopoulos (1999). The long-term effects of small classes: A five-year follow-up of the Tennessee class size experiment. *Educational Evaluation and Policy Analysis* 21.2, pp. 127–142.

Podgursky, Michael (2006). Is Teacher Pay "Adequate?" *Education Working Paper Archive.*

Prokhorov, Artem and Peter Schmidt (2009). GMM redundancy results for general missing data problems. *Journal of Econometrics* 151.1, pp. 47–55.

Qian, Hailong and Peter Schmidt (1999). Improved instrumental variables and generalized method of moments estimators. *Journal of Econometrics* 91.1, pp. 145–169.

Rice, J.K. and A.E. Schwartz (2008). "Toward an understanding of productivity in education." *Handbook of Research in Education Finance and Policy.* Routledge New York. Chap. 8, pp. 131–165.

Rockoff, Jonah (2009). Field experiments in class size from the early twentieth century. *The Journal of Economic Perspectives* 23.4, pp. 211–230.

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

Todd, P.E. and K.I. Wolpin (2003). On the Specification and Estimation of the Production Function for Cognitive Achievement. *The Economic Journal* 113.485, F3–F33.

U.S. Department of Education (2009). *U.S. Department of Education.* `http://nces.ed.gov/programs/digest/d09/tables/dt09_079.asp`. U.S. Department of Education. Institute of Education Sciences, National Center for Education Statistics.

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

Wooldridge, Jeffrey M (2010). *Econometric analysis of cross section and panel data.* Second Edition. MIT Press.

Yoon, Kwang Suk, Teresa Duncan, Silvia Wen-Yu Lee, Beth Scarloss, and Kathy L Shapley (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.

Chapter 2

School Districts' Revealed Preference for Resource Allocation: Evidence from California's Quality Education Investment Act

2.1 Introduction

Despite a growing concern since the 1970s over financial disparities across school districts, with the least resources going to those districts least able to raise local revenue,[1] relatively little attention has been paid to disparities in resource allocation within district. Due primarily to a paucity of school-level financial data, few analyses have explored the causes and consequences of intra-district disparities. Those that do analyze disparities within district are typically restricted to post-hoc descriptive analyses. They observe, often imperfectly, the outcome of the process that allocates resources across schools, a process driven by district preferences and constrained by institutional factors such as union contracts. California's Quality Education Investment Act (QEIA), on the other hand, provides a unique

[1] For a review, see Corcoran et al. (2008), and Springer et al. (2008).

opportunity to directly examine district preferences over low-performing schools, which may be a driving force of intra-district disparities.

Implemented in 2007, the QEIA allocated approximately $2.7 billion to low-performing schools, and required that funded schools implement a bundle of reforms. Rather than spread the money across the approximately 1,200 schools that were eligible and chose to participate, it was decided that funds would be distributed to a small number of semi-randomly chosen schools, to enable financing of ambitious reforms in each school.

As part of the process of choosing funded schools, each district with more than one participating school was required to rank its participating schools. If selected for funding, the school was required to reduce class size, increase their counselor-student ratio,[2] provide training opportunities to staff, align their average teacher experience with the district average, and meet accelerated academic performance targets. Districts believed that the probability of a school receiving funding was higher for higher ranked schools.[3] District rankings therefore provide a window into district preferences for resource allocation.

Using a discrete choice mode that leverages districts' full rankings, this paper seeks to address the question of how districts choose to allocate resources across schools. I find some evidence that districts preferred schools that applied to, and would receive priority for, an alternative program, in which schools crafted their own reforms. There is also some evidence that districts ranked highly schools with a high percentage of students eligible for Free and Reduced Price Lunch (FRPL).

There is clear evidence that districts ranked highly those schools that had been repeatedly sanctioned under No Child Left Behind for failing to make Adequate Yearly Progress.

[2] This requirement applied only to high schools.

[3] The actual selection process was misrepresented to schools, as described below. In fact, the probability of selection was not monotonic in district rankings.

These sanctions imposed costs on districts, and districts apparently preferred using QEIA resources to help mitigate those costs. This has important implications for policy makers, particularly where policies from various levels of government overlap: schools for which the federal government imposed a cost for failing to meet achievement targets were more likely to receive support from their districts. The federal government did therefore effectively incentivize districts to shift resources toward under-performing schools.

Tests of the assumption that all districts share common preferences reject, perhaps due solely to the preferences of one large district, LA Unified. Additionally, a test that districts weight characteristics in selecting the highest ranked school in the same way that they rank characteristics for all schools rejects. Nonetheless, the general pattern in coefficients persists, particularly those pertaining to NCLB sanctions. An important limitation of QEIA is that it does not reveal district preferences over all schools, only over those that are low-performing and chose to participate in QEIA.

The remaining paper proceeds as follows: section 2.2 relates this paper to existing literature; section 2.3 lays out the institutional details of QEIA; section 2.4 describes the data; section 2.5 outlines the model and identification; section 2.6 presents results. Section 2.7 concludes.

2.2 Literature

This paper contributes to the still-nascent literature on intra-district resource allocation, which by necessity focuses on large school districts for which school-level financial data are available. For example, Iatarola et al. (2003) examine resource allocation across middle and elementary schools in New York City. They examine distributions across schools for similar students, distributions within schools for different students, e.g., special education and regular education students, and the association between educational outcomes and equity of resources. In their analysis of resource allocation, they use measures of inequality such as the range, Gini coefficient, and coefficient of variation of resource allocation. They find that there is a trade-off between teacher salaries and certification on one hand, and lower pupil-teacher ratios on the other. They hypothesize that districts try to channel more teachers toward lower-performing schools, but that union contracts allow experienced teachers to choose higher-performing schools.

Roza et al. (2004) find that the practice of using average salary, common among researchers and public officials, creates important intra-distract disparities that can be exacerbated by "budget layering." They note, for example, that Title I is intended to supplement schools after districts equate spending across schools. However, since the legislation permits calculation of teacher salaries using the average in the district, disparities persist. They note that the vast majority of school districts are blind to these disparities because they focus on average teacher salaries, ignoring the possibility that more experienced, and thus higher paid teachers might be channeled toward particular types of schools.

In another study, Klein (2008) analyzes school-level financial data from the Metropolitan Nashville-Davidson County School District in Tennessee, and finds that when enrollment is controlled for, districts actually allocate more resources to schools with high percentages of students eligible for Free and Reduced Price Lunches. No evidence is found to suggest that preferences are determined by academic performance or percent of minority students.

Unlike the above studies, which rely on observed allocations of funding that are the result of district preferences and institutional constraints, the QEIA allows for a direct analysis of district preferences over low-performing schools. The following section details the background and implementation of the QEIA.

2.3 Institutional Details

The QEIA was the consequence of litigation against then California Governor Arnold Schwarzenegger. The plaintiffs in the case successfully argued that the state underfunded public schools in the 2004-2005 and 2005-2006 school years.[4] As a result of the settlement, the state was required to pay back approximately $2.7 billion to K-12 schools.

Recognizing that allocating the money equally across all schools, or even across all low-performing schools, would have a small impact on per-pupil funding in those schools, legislators decided instead to focus on a subset of low-performing schools. The subset was chosen using a lottery mechanism, and the number of schools was chosen such that funding would increase by $500 per student in kindergarten through 3rd grades, $900 in grades 4-8, and $1,000 in high school from 2008-2014, and by half as much in 2007.

Schools were eligible to participate in the lottery if they were in the bottom two deciles of the state's academic performance distribution, as determined by California's Academic Performance Index (API). Eligible schools had to commit to meeting the requirements of QEIA before they could participate in the selection process, though they could instead apply to an alternative program, in which they would design their own reforms. Each district with more than one participating school was required to rank each of its participating schools. It was permissible to give multiple schools the same rank, and indeed several districts did so.

[4] Henceforth, academic years are referred to by the year in which the Spring semester occurs. Thus the 2005-2006 school year is referred to as 2006.

Figure 2.1 provides an excerpt of the ranking submitted by the San Diego School District. Districts had to rank participating schools, list the type of each school (e.g., middle or elementary), whether the school was applying to the alternative program, and if it should receive priority consideration for the alternative program. Priority consideration for the alternative program was given to all high schools, and not to any elementary or middle schools. Each district was then assigned as many random numbers as it had participating schools, and the numbers were allocated to schools based on the rankings. For example, if a district received the numbers 1 and 201, 1 was assigned to the highest ranked school and 201 to the next school. If districts assigned tied rankings, the California Department of Education randomly chose the ordering. Within each stage of the selection, these random numbers determined the order.

Districts were told that the selection would occur in three stages: first, high schools that applied to the alternative program with the lowest random numbers would be chosen until 15% of funds were allocated;[5] second, to ensure geographic diversity, the school with the lowest random number in each county would be selected, if that county did not have a school funded in the first stage; finally, schools with the lowest random numbers applying to the regular program would be selected, until all funds were exhausted. In fact, the last stage was divided in two, with high schools randomly selected separately from and prior to middle and elementary schools. That districts were told high schools and elementary schools would be treated equally is evidenced in contemporaneous school board minutes, e.g., (Santa Rosa City Schools (2007)), and CDE presentations (Balcom (2007)). This is also the depiction in the report to the California legislature (CDE (2010)), written 3 years

[5] If all high schools were selected and the 15% of funds not exhausted, those elementary and middle schools with the lowest random numbers applying to the alternative program would have been selected. However, given the number of high schools that applied, this outcome was not possible.

after the selection process occurred. Indeed, to my knowledge there is no public source that correctly describes the separation of high schools in the selection process.[6]

Once a school was selected, it was required to implement the following reforms: reduce class sizes; align average teacher experience, as measured by the Teacher Experience Index (TEI), with the district average; provide professional development for teachers and staff; ensure that all teachers at the school met requirements to be Highly-Qualified Teachers (HQT); high schools were required to increase counselor-student ratios; and all selected schools were required to exceed their average achievement growth target over the first three years.[7] Schools in the alternative program were exempt from the bulk of these requirements.

The QEIA stipulated that funded schools should have no more than 20 students per class in kindergarten through 3rd grades, and no more than 25 per class in grades 4-12 - or 5 fewer than the baseline average class size, which ever was lower.[8] In the first three years of QEIA, schools were required to reduce the difference between the pre-QEIA average class size and the 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 class room per grade. As such, many schools applied for and were granted waivers from this requirement, and instead met a higher minimum class size requirement. There is no evidence that schools were aware of the possibility of a waiver when they ranked schools.

Under QEIA, teacher experience is measured by the Teacher Experience Index (TEI).

[6] I discovered that the common description was incorrect only after attempting to reconcile the district rankings, actual random number allocation, and funding results. After extensive conversation with CDE employees I learned the actual method.

[7] Student achievement growth targets in California are formulaic: each school is required to improve by 5% of the difference between their API and 800, or by 1 point, which ever is greater, until they reach 800.

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

In calculating the 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 at multiple schools count towards each school's average. Funded schools are required to exceed the district average Teacher Experience Index.

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

All teachers in QEIA funded schools had to meet the requirements of the federal Elementary and Secondary Education Act (ESEA) for Highly Qualified Teachers. According to CDE (2010), and as corroborated in the data (see Table 2.1), the vast majority of schools eligible to participate in QEIA were already required to meet the HQT standard. Schools were also required to meet the stipulations of the *Williams* settlement, which was the result of a 2004 court case, Williams v. California State, and already applied to most QEIA eligible schools prior to QEIA (again, see Table 2.1). The *Williams* settlement required low-performing schools to have qualified teachers and safe, well-maintained facilities.

High schools that received funding under QEIA that were not in the alternative program were required to increase their counselor-student ratio to 1:300. As with the class size reduction, schools were required to reduce the difference between their initial counselor-student ratios and the target level by 1/3 in each of the first three years. Since so few schools in QEIA are high schools, and since half of the high schools in QEIA are in the alternative program, this requirement has not been monitored as extensively as the others (CDE (2010)).

All funded schools were required to have a growth in API over the first three years of funding that exceeded the average target growth over those three years. The API growth target is determined formulaically for all schools in California. After the first three years of funding, regular QEIA schools are required to meet target growth rates for each subsequent year, and QEIA schools participating in the alternative program are required to continue exceeding the target.

QEIA went into effect against the backdrop of the federal No Child Left Behind Act. Under NCLB, schools are required to make Adequate Yearly Progress (AYP), or enter "Program Improvement" (PI) status. Each year that a school remains in PI it faces increasing sanctions. If a school meets AYP for one year, it's PI status remains the same, i.e., it does not advance to the next year of PI. If the school meets AYP for two consecutive years, it exits PI. In the first year of PI, districts must notify parents and provide them the option to choose another school in the district that is not in PI, while schools must divert Title I funding toward professional development. In the second year, the requirements of the first year persist, and districts must also provide supplemental services to students.

In the third year, the district is required to take more severe corrective action, which can include replacing the entire school staff, replacing the curriculum, extending the school year or day, or appointing an outside expert. In the fourth year, a district must plan for changing the governance structure of the school; it can for instance reopen the school as a charter school, replace the staff, or allow the state to take over. In the fifth year the district must implement this plan.

2.4 Data

This analysis draws on a number of publicly available data sets collected by the California Department of Education (CDE). School-level demographic data are made publicly available in California's (API) Data Files.[9] These data include demographics such as enrollment, percent Free and Reduced Price Lunch, and percent of parents with a high school degree. Also included is each school's API, which is a weighted average across subjects of that school's performance on state standardized tests.

Yearly teacher-level data are also publicly available. Unfortunately, each year teachers are reassigned unique identifiers, so the data are not linked across years. I therefore create a synthetic panel, which links teachers across years and within schools on the basis of teacher characteristics, notably teaching experience and experience within the district. For example, if in a particular school there's a teacher in 2008 with four years of experience teaching and 2 years experience teaching in the district, and in the following year that same school has a teacher with five years of teaching experience and three years experience in the district, and both are equivalent in gender and race, I link those observations. If two teachers are observationally equivalent, I randomly link them across years. An important shortcoming of this synthetic panel is that I cannot accurately determine duration spells. These data also include administrators and employees who interact with students but are not teachers, e.g., guidance counselors, librarians.

[9] Available at http://www.cde.ca.gov/ta/ac/ap/apidatafiles.asp

Class size data are available at the employee-assignment level, where an assignment is a class taught by a particular teacher. Teachers teaching multiple math classes at a school appear multiple times in this data, as do classes with multiple teachers, e.g., a teacher and teacher's aide. I use these data to construct average class sizes and the class size targets for schools were they to participate in QEIA.

Finally, I use the rankings submitted by districts to the California Department of Education. A portion of the form submitted by San Diego County is displayed in figure 2.1. Districts were required to note whether the school was participating; the type of the school; whether the school was applying to the alternative program; and, for alternative schools, whether they met the requirement for priority funding.[10] Most importantly, districts had to assign a rank to each participating school. Districts could give all schools a rank of one, or rank a subset of schools equally.

Table 2.1 provides descriptive statistics for all California schools, those eligible to participate which chose not to, those that applied for the alternative program, and those that applied to the regular program. High schools were more likely to apply to the alternative program, for which they received priority. As mentioned above, the vast majority of teachers in schools eligible and participating in QEIA were already classified as HQT, and indeed this is true of all schools in California. The *Williams* Settlement applied to nearly a quarter of all schools, and nearly all schools eligible and or participating in QEIA. Schools that were eligible to participate but chose not to do so were more likely to have met their growth target in 2006. Not surprisingly, schools that were ineligible were less likely to be in PI in 2007.

Table 2.2 provides descriptive statistics for my analytic sample. From the set of all

[10] Priority funding for the alternative program was given to all high schools that applied for that program.

schools participating in QEIA, my analytic sample drops those that are the only school in their district, and those for which all schools in the district were given the same ranking, since these schools provide no information on district preferences. The table provides summary statistics by the number of schools in the district. The data contain 30 districts that ranked 2 schools, 29 that ranked 3 schools, 51 districts with 4-10 schools, and one district, LA Unified, with 234 schools.

For comparison with later results, table 2.3 presents unconditional differences in covariates between schools ranked above the median, and schools ranked below the median, by district size. If a district has an odd number of schools, the median school is randomly set to be above or below. In larger districts, including LA Unified, charter schools were less likely to be ranked above the median.[11] Larger districts are also less likely to rank highly schools that have just entered PI, but more likely to rank highly schools that are in the 5th stage of PI. Across districts of all sizes, schools that bring in more revenue are more likely to be ranked highly, and even more so if they met their growth target in 2006. In larger districts middle schools were more likely to be found ranked above the median.

2.5 Model

Each district with more than one participating school was required to rank each of its participating schools, with the understanding that the highest ranked school in each district had the highest probability of being chosen as a QEIA school.

Let N_d denote the number of schools in district d. Schools are indexed by r, which is also their ranking, and are denoted as S_r, where $S_r \succ S_i \quad \forall i > r$, and $\succ$ denotes preference.

[11] That several coefficients are significant only for larger districts is in part an artifact of the greater statistical power given the larger samples.

If there exists a function $F(\boldsymbol{Z}_r) = F^*(\boldsymbol{Z}_r) + \epsilon_r$, where ϵ_r is iid type I extreme value, such that

$$S_1 \succ S_2 \succ \cdots \succ S_{N_d} \iff F(\boldsymbol{Z}_1) > F(\boldsymbol{Z}_2) > \cdots > F(\boldsymbol{Z}_{N_d}) \tag{2.1}$$

then we can use the following result, known as a rank order logit or exploding logit, and first introduced into the economics literature by Beggs et al. (1981):

$$\Pr(S_1 \succ S_2 \succ \cdots \succ S_N) = \prod_{j=1}^{N} \frac{\mathrm{e}^{F_j^*}}{\sum\limits_{m=j}^{N} \mathrm{e}^{F_m^*}} \tag{2.2}$$

This model does not explicitly allow for tied rankings, which would occur with probability zero. In the QEIA rankings, most ties occur where districts gave the same ranking to all their participating schools, and these districts therefore provide no information about district preferences. I drop these schools, and am left with five districts that give the same rank to two schools, and different ranks to other schools, and one that ranks three schools first, followed by others.

Tied rankings are analogous to tied exit times in a proportional hazard model. There, ties can be thought of as the consequence of low-frequency data. For example, if data are collected yearly, multiple observations may exit throughout the year at different points, but they are observed as exiting simultaneously. If districts are insensitive to small differences in F, then tied rankings would in fact obscure an underlying ordering.

If there is in fact an underlying order, tied rankings can be accommodated by modifying equation (2.2) such that a set of tied schools contribute to the likelihood through the sum of all possible orderings. For example, if a district had 3 schools, and ranked two first and the other second, their contribution to the likelihood would be

$$\mathrm{P}(S_1 \succ S_2 \succ S_3 \cup S_2 \succ S_1 \succ S_3) = \\ \left[\left(\frac{\mathrm{e}^{F_1^*}}{\mathrm{e}^{F_1^*} + \mathrm{e}^{F_2^*} + \mathrm{e}^{F_3^*}}\right)\left(\frac{\mathrm{e}^{F_2^*}}{\mathrm{e}^{F_2^*} + \mathrm{e}^{F_3^*}}\right) + \left(\frac{\mathrm{e}^{F_2^*}}{\mathrm{e}^{F_1^*} + \mathrm{e}^{F_2^*} + \mathrm{e}^{F_3^*}}\right)\left(\frac{\mathrm{e}^{F_1^*}}{\mathrm{e}^{F_1^*} + \mathrm{e}^{F_3^*}}\right)\right]$$

This can quickly complicate the likelihood, since for T_j schools given tied ranking j, there are $T_j!$ terms in the summand. As such, Stata provides various methods for approximating the exact likelihood. For my results, I use Efron's approximation Efron (1977), which for the above example would use

$$\mathrm{P}(S_1 \succ S_2 \succ S_3 \cup S_2 \succ S_1 \succ S_3) = \\ \left(\frac{\mathrm{e}^{F_1^*}}{\mathrm{e}^{F_1^*} + \mathrm{e}^{F_2^*} + \mathrm{e}^{F_3^*}}\right)\left(\frac{\mathrm{e}^{F_2^*}}{0.5 * (\mathrm{e}^{F_1^*} + \mathrm{e}^{F_2^*}) + \mathrm{e}^{F_3^*}}\right)$$

Stata's "exactm" option, which uses a Gauss-Laguerre quadrature approximation of the exact likelihood, yields similar point estimates,[12] but does not allow for the calculation of robust standard errors.

The question then is how to model F. I assume $F^* = F^*(\mathrm{E}(\text{Revenue}), \mathrm{E}(\text{Cost}), \mathbf{X})$, and my interest lies in estimating $\partial F^*/\partial X$. That is, holding constant the expected revenue and

[12] Results are available from the author by request.

costs of participating in QEIA, what school characteristics influence the rankings?

The revenue from a school is a function of its enrollment and the probability that the school remains funded in each subsequent year, while the expected cost is a function of how many teachers must be hired, the required average experience of those teachers, and the total number of teachers, administrators, and para-professionals for whom professional development must be provided. Conditional on meeting the class size, teacher experience, and professional training requirements, whether a school remains in the program is a function of whether its average growth over the first three years exceeds its average growth target over those years. I proxy for this using an indicator for whether the school met its growth requirement in 2006, which I interact with measures of revenue and costs.

Revenue is included as the sum across grades of the per-pupil increase in funding times the number of students in each grade. I model the cost of meeting the class size reduction requirement as being proportional to the number of new teachers that must be hired, i.e., the change in the teacher-pupil ratio times the number of students:

$$\Delta T = \text{Number new teachers} = \left(\frac{1}{CS_{\text{target}}} - \frac{1}{CS_{2007}} \right) * \text{Enrollment}_{2007} \tag{2.3}$$

A school may be able to satisfy its teacher experience requirement and the class size reduction requirement by ensuring that all newly hired teachers have at least 10 years of experience. However, it may also be the case that even after having satisfied the class size reduction requirement, additional changes to the teacher workforce will be required to meet the experience requirement. To capture this, I include an indicator for whether TEI binds after meeting the class size reduction requirement, i.e.,

$$\text{TEI Binds} = 1\left[\frac{1}{T_{2007} + \Delta T}(\text{TEI} * T_{2007} + \Delta T * 10) < \text{TEI}_{\text{target}}\right] \tag{2.4}$$

where T_t is the number of teachers in year t, and $1[\cdot]$ is the indicator function. To capture the cost of providing professional development, I include T_{2007}, and I include the number of paraprofessionals. This last variable is also interacted with an indicator for being a high school to capture the need for high schools to meet the counselor-to-student ratio.[13]

Also included in the model are demographic variables, including indicators for what if any year of PI the schools is in, whether the school applied to the alternative program, indicators for being high schools or middle schools, an interaction between high school and alternative, an indicator for whether the school is a charter school, the percent of students eligible for Free and Reduced Price Lunch, and the percent of students who are Hispanic.

To summarize the model, I estimate the following:

$$F = \xi_1 \text{Revenue} + \xi_2 \text{Revenue} * \text{Met}_{2006} + \xi_3 * \text{Met}_{2006} + \mathbf{c}\boldsymbol{\gamma} + \mathbf{c}\boldsymbol{\delta} * \text{Met}_{2006} + \boldsymbol{\beta} X + \varepsilon \tag{2.5}$$

where Met_{2006} is an indicator for having met the growth target in 2006, $\mathbf{c}$ is a row vector containing number of teachers, number of non-teaching employees working with students, the gap between a school's TEI and its target, required number of new teachers, and indicators for high school and middle school.

Though this "kitchen sink" approach might fully control for the expected cost and revenue of participating in QEIA, care must be taken in interpreting the coefficients on measures of revenue and cost. Consider, for example, a middle school with an average

[13] Guidance counselors are included in this variable, but cannot be identified separately from other positions, e.g., school nurses.

class size of 35 students and a target of 25 students per class. Revenue, as a function of enrollment, cannot increase while holding constant the required number of new teachers, since for every 87.5 additional students the school receives an additional $78,780, and must hire one more teacher to fulfill the class size requirement.[14] While I attempt to control for expected revenue and costs, I don't tease apart their effects.

2.6 Results

Coefficients on demographic variables are presented in Table 2.4, with the baseline results in the first column. Standard errors in brackets are robust to misspecification. That is, even if ε does not follow the type I extreme value distribution, the robust standard errors are correct for estimates of the parameters that minimize the misspecified log likelihood. Standard errors in parentheses are not robust to this misspecification. Though no more likely to rank highly middle and elementary schools that applied for the alternative program, the non-robust standard error and point estimate suggests districts were more likely to rank highly high schools applying to this program. This suggests that districts understood that high schools would be given priority in this program, and they valued the flexibility of the program. Districts were no more likely to rank charter schools higher than regular schools, though they were more likely to rank highly schools with a higher percentage of FRPL students, with a 43 percentage point increase in FRPL having an effect of the same magnitude as being a high school applying for the alternative program. There is marginal evidence that districts preferred schools with a high percentage of black students.

The pattern in the point estimates on year in PI suggests the more years a school was in

[14] For comparison, average teacher salary in California in 2008 was $64,424 U.S. Department of Education (2009), excluding benefits, which Podgursky (2006) estimates to account for 20% of total compensation.

PI the more likely the district was to rank the school highly. The financial and reputation costs of PI increased each year, and districts appear to have seen QEIA as a way to limit these costs. The effect of being in the 5th year of PI is almost four times as large as the effect for being in the 1st year, and is comparable to the effect of a school going from no students eligible for FRPL to all students being eligible.

LA schools make up 21.1% of my sample, and my results may partially be driven by the ranking of LA Unified. Column 2 presents estimation results excluding LA from the sample. In an interaction of all variables with an indicator for being an LA school, a Wald test rejects the null of no difference in coefficients with $p < 0.001$. Nonetheless, as the second column indicates, the pattern on demographic coefficients is quite similar to the baseline model. One exception is the effect of being a charter school, which diminished a school's ranking more in districts other than LA Unified. Districts give more favorable rankings to high schools participating in the alternative program, and to schools with a high percentage of students eligible for Free and Reduced Price Lunch. The pattern persists of districts giving higher rankings to schools the longer they are in PI, though the effect is smaller.

The distinct preferences of LA Unified are one example of how the assumption of same coefficients across districts could be violated. Other tests are presented in columns 3 and 4, both of which include an interaction with a count variable of the number of participating schools, and the latter of which also excludes LA schools. The count variable is the number of participating schools minus the average number of participating schools in districts other than LA Unified. A Wald test of the null hypothesis that the coefficients on terms interacted with the count variable are all zero rejects when LA is included ($p < .001$), and at the 10%

level when LA is excluded ($p = 0.08$). For a typical district excluding LA, schools with high percentages of FRPL lunch students, and schools in latter years of PI, are more likely to be ranked highly. There is some evidence that the typical district, excluding LA, was less likely to rank charter schools highly.

Another assumption of the rank order logit model is that districts weight characteristics equally whether they are ranking the first school or the last. One way to test this assumption is to estimate a conditional logit model, in which districts choose only the highest ranked school. The results from this exercise are presented in the final column. A Hausman test of the null hypothesis of no misspecification rejects ($p = 0.0026$), suggesting the rank order logit assumptions are violated. Nonetheless, one finding remains true across all specifications: districts were more likely to rank highly schools that were in the fifth, and most severe year of PI. Given that the rank order logit model fails several specification tests, the results should be interpreted as descriptive.

Nonetheless, across specifications a clear pattern emerges: districts preferred to rank highly those schools that faced sanctions under NCLB, and the more severe those sanctions, the more highly ranked the school became. NCLB sanctions were intended to improve student achievement, but they imposed a cost on districts. For instance, in the first year of Program Improvement, schools had to provide additional professional training, but this effort was to be funded using existing Title I allocations, which were therefore diverted from elsewhere. The descriptive evidence suggests that districts thought sanctioned schools most deserved to participate in QEIA, in an effort to help those schools exit NCLB sanctioning. The federal government was therefore able to influence intra-district allocations, by providing incentives for districts to shift resources to sanctioned schools.

2.7 Conclusion

The landmark court case *Serrano v. Priest* of 1971 ushered in an era of awareness of disparities in educational resources across districts, with students from families with the least resources attending districts that likewise were under-resourced. Due primarily to a lack of within-district financial data, few studies have been able to address the question of whether disparities exist within district as well. Resource allocation within a district is determined by district preference and institutional constraints. Studies of within-district disparities have by necessity focused on the outcome of this process in a handful of districts.

This paper seeks to understand determinants of intra-district resource allocation by observing directly district preferences over low-performing schools. Due to a requirement of California's QEIA, districts were essentially required to answer the question, "Were you to receive funding for one school to implement mandated reforms, which would you choose? Conditional on that school being funded, which would you choose next?" Using districts' responses, in the form of rankings, I model district preferences using a discrete choice model.

Doing so, I find consistent evidence that districts preferred to fund schools that were in the 5th year of PI. Under No Child Left Behind, schools that fail to meet Adequate Yearly Progress are forced into increasingly strict sanctions, referred to in California as PI. By the fifth year of PI, schools are required to implement plans that dramatically change their organizational structure, by for instance reopening as a charter school, replacing the entire staff, or allowing the state to take over. Districts seemed to have preferred giving these schools the opportunity to participate in QEIA. This has important implications for policy makers, particularly where policies from various levels of government overlap: schools for which the federal government imposes a cost for failing to meet achievement targets are more likely to receive support from their districts. The federal government can therefore effectively incentivize districts to shift resources toward under-performing schools.

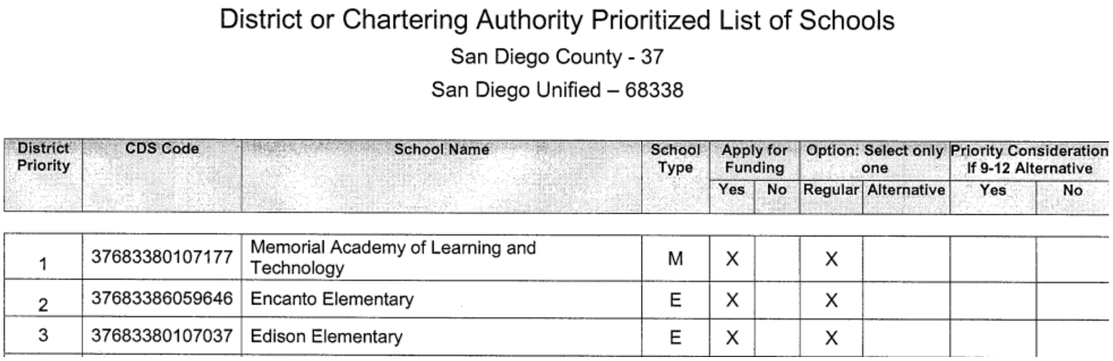
The rank order logit model that I employ has strong assumptions, such as constant coefficients across districts and choices. That is, each district is assumed to weight characteristics equally, whether they are choosing their highest ranked or second-to-last ranked schools. Tests for the validity of these assumptions fail in the case of QEIA, and the results are therefore best viewed as descriptive, rather than as estimates of underlying parameters of districts' utility functions.

Another shortcoming of this study is that it is only capable of describing preferences over low-performing schools. QEIA required districts to rank only those schools eligible to participate, and eligibility was determined by an academic achievement cut-off. While district preferences for resource allocation across low-performing schools is important, this undoubtedly misses important dynamics in the allocation of resources across all schools within a district. The study of district preferences across all schools is therefore left to future research.

APPENDICES

APPENDIX A - FIGURES

Figure 2.1: Portion of Form Submitted by San Diego



District or Chartering Authority Prioritized List of Schools

San Diego County - 37

San Diego Unified – 68338

District Priority	CDS Code	School Name	School Type	Apply for Funding		Option: Select only one		Priority Consideration If 9-12 Alternative	
				Yes	No	Regular	Alternative	Yes	No
1	37683380107177	Memorial Academy of Learning and Technology	M	X		X			
2	37683386059646	Encanto Elementary	E	X		X			
3	37683380107037	Edison Elementary	E	X		X			

APPENDIX B - TABLES

Table 2.1: Descriptive Statistics, All Schools

	All Schools	Eligible, Not Participating	Participating in Alternative	Participating in Regular
2007 class size	22.53	23.33	25.36	23.18
	(6.13)	(4.49)	(4.29)	(3.55)
Middle school	0.15	0.14	0.19	0.17
	(0.36)	(0.35)	(0.40)	(0.38)
High school	0.24	0.22	0.45	0.11
	(0.43)	(0.41)	(0.50)	(0.31)
HQT 2007	0.92	0.91	0.87	0.90
	(0.17)	(0.17)	(0.13)	(0.15)
Williams applies	0.23	0.89	0.95	0.95
	(0.42)	(0.32)	(0.21)	(0.23)
Met target 2006	0.61	0.71	0.60	0.65
	(0.49)	(0.46)	(0.49)	(0.48)
2007 LA	0.08	0.01	0.24	0.18
	(0.27)	(0.10)	(0.43)	(0.39)
Year 1 of PI	0.07	0.10	0.19	0.17
	(0.26)	(0.30)	(0.40)	(0.37)
Year 2 of PI	0.03	0.10	0.07	0.10
	(0.18)	(0.30)	(0.25)	(0.30)
Year 3 of PI	0.05	0.15	0.10	0.21
	(0.22)	(0.36)	(0.30)	(0.41)
Year 4 of PI	0.03	0.18	0.18	0.16
	(0.18)	(0.39)	(0.39)	(0.37)
Year 5 of PI	0.04	0.14	0.26	0.19
	(0.19)	(0.35)	(0.44)	(0.39)
N	9714	195	88	1172

Note: Table lists sample means and standard deviations in parentheses. With the exception of 2007 class size, all variables are dichotomous, and thus means are the proportion of schools falling into that category. PI refers to “Program Improvement.” All schools is the universe of public schools in California in 2007.

Table 2.2: Descriptive Statistics, by Ranking

	Participating Schools in District						
	2	3	4	5	6-10	11-55	234
Number of Districts	30	29	16	16	19	18	1
Alternative Program	0.033	0.080	0.063	0	0.144	0.055	0.090
	(0.181)	(0.274)	(0.244)	(0)	(0.352)	(0.228)	(0.286)
HS*(Alternative)	0	0	0.016	0	0.068	0.027	0.051
	(0)	(0)	(0.125)	(0)	(0.253)	(0.163)	(0.221)
Charter	0.033	0.023	0.016	0.025	0.014	0.037	0.026
	(0.181)	(0.151)	(0.125)	(0.157)	(0.117)	(0.188)	(0.158)
Prop. FRPL	0.851	0.836	0.821	0.750	0.803	0.867	0.896
	(0.125)	(0.126)	(0.175)	(0.145)	(0.139)	(0.136)	(0.096)
Prop. Hispanic	0.732	0.746	0.828	0.741	0.790	0.704	0.832
	(0.221)	(0.227)	(0.185)	(0.216)	(0.173)	(0.231)	(0.195)
Prop. Black	0.055	0.057	0.031	0.113	0.079	0.151	0.132
	(0.087)	(0.079)	(0.065)	(0.148)	(0.104)	(0.167)	(0.193)
Year 1 of PI	0.167	0.138	0.188	0.138	0.164	0.105	0.299
	(0.376)	(0.347)	(0.393)	(0.347)	(0.372)	(0.307)	(0.459)
Year 2 of PI	0.150	0.092	0.047	0.125	0.137	0.084	0.047
	(0.360)	(0.291)	(0.213)	(0.333)	(0.345)	(0.278)	(0.212)
Year 3 of PI	0.233	0.184	0.234	0.175	0.199	0.256	0.192
	(0.427)	(0.390)	(0.427)	(0.382)	(0.400)	(0.437)	(0.395)
Year 4 of PI	0.150	0.322	0.188	0.250	0.144	0.194	0.047
	(0.360)	(0.470)	(0.393)	(0.436)	(0.352)	(0.396)	(0.212)
Year 5 of PI	0.133	0.149	0.156	0.138	0.123	0.224	0.299
	(0.343)	(0.359)	(0.366)	(0.347)	(0.330)	(0.417)	(0.459)
Elementary	0.750	0.713	0.641	0.700	0.699	0.692	0.675
	(0.437)	(0.455)	(0.484)	(0.461)	(0.460)	(0.462)	(0.469)
N	60	87	64	80	146	438	234

Note: Table lists sample means and standard deviations in parentheses. With the exception of 2007 class size, all variables are dichotomous, and thus means are the proportion of schools falling into that category. PI is "Program Improvement." Required new teachers is the change in the mandated change in the teacher/student ratio times student enrollment.

Table 2.3: Difference Above Median Ranking-Below Median Ranking

	Participating Schools in District						
	2	3	4	5	6-10	11-55	234
Number of Districts	30	29	16	16	19	18	1
Alt. program	0.000	-0.017	-0.063	0.000	0.068	0.028	0.043
	(0.047)	(0.059)	(0.061)	(0.000)	(0.058)	(0.022)	(0.037)
HS*(alt.)	0.000	0.000	-0.031	0.000	0.027	0.019	0.051*
	(0.000)	(0.000)	(0.031)	(0.000)	(0.042)	(0.016)	(0.029)
Charter	0.000	-0.044	0.031	-0.048	-0.027	-0.054***	-0.034*
	(0.047)	(0.031)	(0.031)	(0.033)	(0.019)	(0.018)	(0.021)
Prop. FRPL	-0.011	-0.015	0.029	-0.034	-0.003	0.018	-0.031**
	(0.032)	(0.027)	(0.044)	(0.033)	(0.023)	(0.013)	(0.012)
Prop Hispanic	0.001	0.037	-0.007	0.016	0.029	-0.023	-0.029
	(0.058)	(0.049)	(0.047)	(0.048)	(0.029)	(0.022)	(0.026)
Prop black	0.006	-0.010	0.004	-0.006	-0.010	0.027*	0.030
	(0.023)	(0.017)	(0.016)	(0.033)	(0.017)	(0.016)	(0.025)
Year 1 of PI	-0.000	0.102	-0.063	-0.061	-0.137**	-0.081***	-0.376***
	(0.098)	(0.075)	(0.099)	(0.077)	(0.061)	(0.029)	(0.055)
Year 2 of PI	0.167*	-0.040	-0.094*	-0.188***	-0.055	-0.022	-0.043
	(0.091)	(0.062)	(0.052)	(0.069)	(0.057)	(0.027)	(0.028)
Year 3 of PI	-0.067	-0.033	0.031	0.068	0.178***	-0.034	0.060
	(0.111)	(0.084)	(0.108)	(0.086)	(0.065)	(0.042)	(0.052)
Year 4 of PI	-0.100	0.022	0.125	-0.025	0.041	0.070*	0.060**
	(0.093)	(0.101)	(0.098)	(0.098)	(0.058)	(0.038)	(0.028)
Year 5 of PI	0.067	0.079	0.125	0.189**	0.055	0.212***	0.496***
	(0.089)	(0.077)	(0.091)	(0.077)	(0.055)	(0.039)	(0.051)
Elementary	-0.167	-0.273***	0.094	-0.281***	-0.301***	-0.163***	-0.530***
	(0.112)	(0.095)	(0.121)	(0.100)	(0.072)	(0.044)	(0.051)
N	60	87	64	80	146	438	234

Note: Robust standard errors in parentheses for the coefficient from a regression of the characteristic on an indicator for whether the school is above or below district's median ranking. In districts with an odd number of schools, the median school is randomly assigned to be above or below the median. PI is "Program Improvement."

Table 2.4: Rank Order Logit Results, Efron's Approximation for Ties

	Baseline	LA Excluded	Interact Count	Exclude LA	C. Logit
				Interact Count	
Alt. prog.	-0.09	0.09	0.09	0.17	0.53
	(0.33)	(0.40)	(0.41)	(0.39)	(0.94)
	[0.39]	[0.51]	[0.52]	[0.52]	[0.87]
HS*(Alt.)	0.73	0.91	0.37	0.79	0.10
	(0.37)**	(0.49)*	(0.59)	(0.50)	(1.18)
	[0.54]	[0.84]	[0.94]	[0.81]	[1.10]
Charter	-0.38	-0.65	-0.91	-0.68	-0.55
	(0.29)	(0.34)*	(0.40)**	(0.35)*	(0.95)
	[0.34]	[0.26]**	[0.26]***	[0.27]**	[0.65]
Prop. FRPL	1.68	1.57	1.97	1.32	2.07
	(0.54)***	(0.60)***	(0.65)***	(0.61)**	(1.82)
	[0.82]**	[0.85]*	[0.93]**	[0.95]	[1.82]
Prop. Hispanic	0.49	0.45	-0.14	0.31	-0.55
	(0.57)	(0.63)	(0.70)	(0.65)	(1.96)
	[0.84]	[0.79]	[0.76]	[0.81]	[1.69]
Prop. black	1.07	1.36	0.25	1.19	1.08
	(0.60)*	(0.74)*	(0.91)	(0.77)	(2.70)
	[0.86]	[0.98]	[1.10]	[0.97]	[1.83]
Year 1 of PI	0.42	0.10	0.10	0.06	-0.25
	(0.14)***	(0.18)	(0.20)	(0.19)	(0.53)
	[0.31]	[0.23]	[0.25]	[0.25]	[0.54]
Year 2 of PI	0.85	0.39	0.51	0.44	0.13
	(0.18)***	(0.19)**	(0.21)**	(0.20)**	(0.57)
	[0.46]*	[0.21]*	[0.25]**	[0.24]*	[0.56]
Year 3 of PI	1.19	0.67	0.74	0.67	0.33
	(0.15)***	(0.17)***	(0.18)***	(0.17)***	(0.50)
	[0.52]**	[0.25]***	[0.25]***	[0.25]***	[0.48]
Year 4 of PI	1.10	0.66	0.73	0.67	0.59
	(0.17)***	(0.18)***	(0.19)***	(0.18)***	(0.48)
	[0.54]**	[0.30]**	[0.29]**	[0.29]**	[0.48]
Year 5 of PI	1.65	1.03	1.03	0.98	0.98
	(0.16)***	(0.18)***	(0.19)***	(0.19)***	(0.49)**
	[0.67]**	[0.40]**	[0.36]***	[0.39]**	[0.47]**
N	1097	863	863	1097	1093

Note: * indicates $p < 0.10$, ** indicates $p < 0.05$, *** indicates $p < 0.01$. Coefficients are from rank order logit regression of district rankings on student outcomes, except the final column, which is conditional logit where dependent variable is 1 if district ranked the school highest. Coefficients for measures of expect revenue and expected cost and included but not shown. Standard errors in parentheses, and standard errors robust to misspecification in brackets. Count refers to measure of number of participating schools in the district, minus the average number of participating schools in districts other than LA.

BIBLIOGRAPHY

BIBLIOGRAPHY

Balcom, Fred (Feb. 2007). *Quality Education Investment Act (QEIA) of 2006.* http://www.cde.ca.gov/fg/fo/r16/documents/qeia07present.ppt. California Department of Education.

Beggs, S., S. Cardell, and J. Hausman (1981). Assessing the potential demand for electric cars. *Journal of Econometrics* 17.1, pp. 1–19.

CDE (Jan. 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. California Department of Education.

Corcoran, S.P. and W.N. Evans (2008). Equity, adequacy, and the evolving state role in education finance. *Handbook of Research in Education Finance and Policy.*

Efron, Bradley (1977). The efficiency of Cox's likelihood function for censored data. *Journal of the American Statistical Association* 72.359, pp. 557–565.

Iatarola, P. and L. Stiefel (2003). Intradistrict equity of public education resources and performance. *Economics of Education Review* 22.1, pp. 69–78.

Klein, C.C. (2008). Intradistrict public school funding equity, community resources, and performance in Nashville, Tennessee. *Journal of Education Finance*, pp. 1–14.

Podgursky, Michael (2006). Is Teacher Pay "Adequate?" *Education Working Paper Archive.*

Roza, M., P.T. Hill, S. Sclafani, and S. Speakman (2004). How within-district spending inequities help some schools to fail. *Brookings Papers on Education Policy*, pp. 201–227.

Santa Rosa City Schools (Mar. 2007). *School Board Minutes, Quality Education Investment Act.* http://www.srcs.k12.ca.us/board/agendas/attachments/032807-BR-F7.pdf.

Springer, Matthew G, Eric A Houck, and James W Guthrie (2008). History and scholarship regarding United States education finance and policy. *Handbook of Research in Education Finance and Policy*, pp. 3–22.

U.S. Department of Education (2009). *U.S. Department of Education.* http://nces.ed.gov/programs/digest/d09/tables/dt09_079.asp. U.S. Department of Education. Institute of Education Sciences, National Center for Education Statistics.

Chapter 3

Asymptotic Properties of Quantile Regression for Standard Stratified and Variable Probability Sampling[1]

3.1 Introduction

Quantile Regression has been widely used in the social sciences in recent decades, in part due to its ability to estimate changes throughout the conditional distribution of an outcome of interest. Ordinary Least Squares models the effect on an outcome of interest as a location shift in the conditional distribution of the outcome variable. Yet causal effects may manifest as greater variance, skewness, or density in the tails of the conditional distribution, all of which may be obscured by focusing exclusively on location shifts. As exemplified by Koenker (2005), changes in independent variables may even induce a bimodal conditional distribution. Quantile Regression can reveal these effects. A natural use of Quantile Regression has been to analyze the wage structure and potential differences in the determinants of wages observed at different points of the wage distribution, e.g., Albrecht et al. (2003); Buchinsky (1998); Buchinsky (2001); Machado et al. (2005); Martins et al. (2004);

[1] This chapter coauthored with Otávio Bartalotti.

and Melly (2005). Given a sample in which observations are selected with equal probability, well-established methods are available for estimating a Quantile Regression model (Koenker (2005), Wooldridge (2010)).

Frequently, however, samples are not drawn with equal probability. Commonly used data sets such as the Current Population Survey, Panel Study of Income Dynamics, National Longitudinal Survey of Youth, and the Health and Retirement Study sample with unequal probability. In order to more precisely estimate characteristics of subpopulations of interest, these subpopulations are often oversampled. Ignoring the sampling design of such data sets may lead to inconsistent estimation, in which case consistent estimation can proceed by weighting observations.[2]

Two types of sampling schemes that are prevalent in a wide range of surveys and datasets in social sciences are Standard Stratified (SS) and Variable Probability (VP) sampling. With SS sampling, the population is divided into J mutually exclusive, exhaustive strata, and a random sample of size N_j is taken from stratum j. Alternatively, in the VP sampling case an observation is first drawn at random from the population, and if the observation falls into stratum j, it is kept with probability p_j. In either case, when stratification is exogenous, i.e., the probability of selection is independent of the outcome conditional on covariates, estimation can proceed without regard to the stratification; the usual estimators that ignore stratification are consistent, efficient, and asymptotically normal, and the usual variance estimators are valid (Wooldridge (1999); Wooldridge (2001)).

When the probability of selection is not independent of the outcome conditional on covariates, stratification is said to be endogenous, and the standard estimators are generally inconsistent. The asymptotic properties of M-estimators with smooth objective functions

[2] For a general discussion and guidance on the appropriateness of weighting, see Solon et al. (2013)

under VP and SS sampling have been analyzed in Wooldridge (1999) and Wooldridge (2001), respectively. However, these results are not directly applicable to the Quantile Regression case due to the nonsmoothness in the objective function that provides the QR estimates.

Bartalotti (2012) partially fills this gap, by developing the asymptotic properties of quantile regressors under SS sampling. This paper extends the analysis to Quantile Regression under VP sampling. Additionally, we present evidence from simulations, which demonstrate that Stata's weighted standard errors are quite inaccurate, particularly under VP sampling. Bootstrapped standard errors outperform analytic standard errors under VP sampling across coefficients, quantiles, and sample sizes. Under SS sampling no method of estimating standard errors performs consistently well. In what follows, section 3.2 reviews the standard Quantile Regression estimator. Section 3.3 reviews the asymptotic properties of Quantile Regression under SS sampling, and develops those of VP sampling. Section 3.4 provides results from Monte Carlo simulations. Section 3.5 concludes.

3.2 The Quantile Regression Population Problem

We are interested in estimating the Conditional Quantile Function (CQF) of a random variable y conditional on a vector of q explanatory variables $\mathbf{x}$. This is defined by,

$$Q_\tau(y \mid \mathbf{x}) = \inf\{y : F(y \mid \mathbf{x}) \geq \tau\}$$

where $\tau \in (0, 1)$ indexes the τ^{th} quantile of the conditional distribution of y. Let the CQF be described by a known function $g(\cdot)$ of the parameters and the explanatory variables

$$Q_\tau(y \mid \mathbf{x}) = g\left(\mathbf{x}, \boldsymbol{\beta}_{o,\tau}\right) \tag{3.1}$$

β is subscripted with "o" to denote the true population parameter, and with τ to indicate that the parameters typically vary with τ.

A special case of interest is given by the linear model:[3]

$$y = \mathbf{x}'\boldsymbol{\beta}_{o,\tau} + \varepsilon \tag{3.2}$$

with $Q_\tau(\varepsilon \mid \mathbf{x}) = 0$. Throughout this paper we concentrate on the linear CQF, for ease of exposition and since it is the most widely used by practitioners. Nevertheless, the results presented are valid for a nonlinear, correctly specified CQF, $g(\cdot)$. In the population, $\boldsymbol{\beta}_{o,\tau}$ solves the following problem:

$$\min_{\boldsymbol{\beta}_\tau \in \mathbf{B}} E\left[\rho_\tau\left(y - \mathbf{x}'\boldsymbol{\beta}_\tau\right)\right] \tag{3.3}$$

where, $\rho_\tau(u) = (\tau - 1\,[u \leq 0])u$ and $\mathbf{B} \in \mathbb{R}^K$ is the parameter space.

Given a random sample from the population of size N, it is possible to obtain consistent estimates of $\boldsymbol{\beta}_{o,\tau}$ by a standard Quantile Regression estimator, which solves the following:

$$\min_{\boldsymbol{\beta}_\tau \in \mathbf{B}} N^{-1} \sum_{i=1}^{N} \rho_\tau(y_i - \mathbf{x}_i'\boldsymbol{\beta}_\tau) \tag{3.4}$$

Note that the minimization problem has the following first order conditions and sample

[3] This formulation assumes the error term is additive and, hence, separable. For a treatment of the more general formulation with (potentially) non-separable ε see Powell (1991).

analogue (Buchinsky (1998)):[4]

$$
\begin{aligned}
E\left[\left(\tau - 1\left[y - \mathbf{x}'\boldsymbol{\beta}_{o,\tau} \leq 0\right]\right)\mathbf{x}\right] &= \mathbf{0} && (3.5)\\
N^{-1}\sum_{i=1}^{N}\left(\tau - 1\left[y_i - \mathbf{x}_i'\breve{\boldsymbol{\beta}}_{\tau} \leq 0\right]\right)\mathbf{x}_i &= \mathbf{0} && (3.6)
\end{aligned}
$$

where $1[\cdot]$ is the indicator function.

We can therefore frame this problem as a GMM estimator that uses as moment conditions the first order conditions of the Quantile Regression problem that identify $\boldsymbol{\beta}_{o,\tau}$. Under random sampling, the standard Quantile Regression procedures can be used to estimate $\boldsymbol{\beta}_{o,\tau}$ and to perform inference.

3.3 Quantile Regression under SS and VP Sampling

3.3.1 SS Sampling

We review here the SS sampling case explicated in Bartalotti (2012), and extend the analysis to VP sampling. Under SS sampling, the population is divided into J strata, $W_1, W_2, ..., W_J$. A sample of N_j observations is drawn randomly from each stratum j, and is denoted by $\{\mathbf{w}_{ij} = (y_{ij}, \mathbf{x}_{ij}) : i = 1, \ldots, N_j\}$.

The strata sample sizes N_j are nonrandom. Therefore, the total sample size, $N = N_1 + \cdots + N_J$, is nonrandom. The density of a characteristic y in the jth stratum is denoted by $dF(Y|j)$ with $F(a|j)$ denoting the population proportion of households in stratum W_j with $y < a$. Crucially, this density can differ across strata so even though the observations are

[4] In general the FOC does not hold exactly, but the left-hand side of equation 3.6 is $o_p(N^{1/2})$. See Buchinsky (1998).

i.i.d. within strata, observations from different strata are independent but not necessarily identically distributed.

Bartalotti (2012) shows that a consistent estimator of $\boldsymbol{\beta}_{o,\tau}$ uses the following sample moment condition:

$$\frac{1}{N}\sum_{i=1}^{N}\frac{Q_j}{H_j}\left(\tau - 1\left[y_{ij} - \mathbf{x}'_{ij}\widehat{\boldsymbol{\beta}}_{\tau} \leq 0\right]\right)\mathbf{x}_i = \mathbf{0} \tag{3.7}$$

where $Q_j = P(\mathbf{w} \in W_j)$ is assumed known, and $H_j = \frac{N_j}{N}$. If Q_j is unknown, it can readily be estimated from large survey data.

This is the empirical moment condition that is used to estimate the parameters of interest, defining the weighted Quantile Regression estimator under SS sampling. This estimator is consistent for the parameters of interest under Standard Stratified sampling (Wooldridge (2001)'s theorem 3.1).[5]

As Bartalotti (2012) shows, under standard regularity conditions, $\sqrt{N}(\hat{\boldsymbol{\beta}}_{\tau} - \boldsymbol{\beta}_{o,\tau}) \overset{a}{\sim} \mathcal{N}\left(0, A_1^{-1}B_1A_1^{-1}\right)$, where

$$A_1 = E\left[f_{y|x}\left(\mathbf{x}'\boldsymbol{\beta}_{o,\tau}\right)\mathbf{x}\mathbf{x}'\right]$$

and

$$B_1 = \sum_{j=1}^{J}\frac{Q_j^2}{\overline{H}_j}\text{Var}\left[\mathbf{q}|w \in W_j\right]$$

[5] As a minor point note that if one wants to implement the weighted estimator by applying a standard Quantile Regression to weighted data, the weights for each observation will be given by $\frac{Q_{j_i}}{H_{j_i}}$ instead of the $\left(\frac{Q_{j_i}}{H_{j_i}}\right)^{\frac{1}{2}}$ usually required when implementing least squares estimation and its variants.

and $\mathbf{q}_{ij} = \left(\tau - 1[y_{ij} - \mathbf{x}'_{ij}\boldsymbol{\beta}_{o,\tau} \leq 0]\right)\mathbf{x}_{ij}$.

Two main points regarding B_1 are worth mentioning. The first, which is general to the Standard Stratification literature, is that we cannot replace Var $\left[\mathbf{q}|w \in W_j\right]$ by the outer product of the score as in the random sampling case because in general

$$E\left[\left(\tau - 1\left[y - \mathbf{x}'\boldsymbol{\beta}_{\tau o} \leq 0\right]\right)\mathbf{x}|w \in W_j\right] \neq 0 \tag{3.8}$$

as pointed out by Wooldridge (2001). Without further assumptions, the population moment condition does not necessarily hold in each stratum. Second, it is interesting to note that, distinct from the standard results in Quantile Regression for random sampling, B_1 does not simplify to $\tau(1-\tau)E[\mathbf{x}\mathbf{x}']$ in this case, since the variance of the binary variable $1\left[y_{ij} - \mathbf{x}'_{ij}\boldsymbol{\beta}_{o,\tau} \leq 0\right]$ is not necessarily the same for each stratum. That is, $\mathbf{x}'_{ij}\boldsymbol{\beta}_{o,\tau}$ will not represent the τ^{th} quantile in every stratum.

A feasible estimator requires knowledge of $f_{y|x}$. Koenker (2005) suggests using the fact that $1/f_{y|x} = \mathrm{d}Q_\tau(Y|\mathbf{X})/\mathrm{d}t$. $f_{y|x}$ can therefore be estimated using the inverse of a difference quotient:

$$\hat{f}_{y|x} = \frac{2h}{\mathbf{X}\hat{\boldsymbol{\beta}}_{\tau+h} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\tau-h}} \tag{3.9}$$

We thus use the following estimate of A_1:

$$\hat{A}_1 = N^{-1}\sum_{i=1}^{N}\frac{Q_j}{H_j}\hat{f}_{i,y|x}(\mathbf{x}_{ij}\hat{\boldsymbol{\beta}}_\tau)\mathbf{x}_{ij}\mathbf{x}'_{ij} \tag{3.10}$$

A natural estimate of B_1 is

$$\hat{B}_1 = N^{-1} \sum_{i=1}^{N} \frac{Q_j^2}{H_j^2} \left(\hat{\mathbf{q}}_{ij} - \bar{\hat{\mathbf{q}}}_j \right) \tag{3.11}$$

where $\hat{\mathbf{q}}_{ij} = \left(\tau - 1[y_{ij} - \mathbf{x}_{ij}' \hat{\boldsymbol{\beta}}_\tau \leq 0] \right) \mathbf{x}_{ij}$, and $\bar{\hat{\mathbf{q}}}_j = N_j^{-1} \sum_{i \in W_j} \hat{\mathbf{q}}_{ij}$.

3.3.2 VP Sampling

Under VP sampling, N observations are first drawn at random from the population, and the sample is denoted by $\{\mathbf{w}_i = (y_i, \mathbf{x}_i)\ i = 1, \ldots, N\}$. If an observation falls into stratum j, it is kept with probability p_j. Following Wooldridge (1999), for each individual i we define J indicator variables $s_{ij} = 1[\mathbf{w}_i \in W_j]$. Likewise, we define for each individual i J binary variables h_{ij}, where $\mathrm{P}(h_{ij} = 1) = p_j$. If observation i is in stratum j it is kept if $h_{ij} = 1$. Finally, define $r_{ij} = s_{ij} h_{ij}$, which indicates whether random draw i is kept, and if so what stratum it belongs to. Note that under VP sampling, the number of observations kept from stratum j, N_j, is random, and so therefore is the total number of observations kept across strata, $N_0 = N_1 + \cdots + N_J$.

Corollary 1. *With these definitions, a consistent Quantile Regression estimator under VP sampling is given by the following sample moment condition:*

$$\sum_{i=1}^{N} \sum_{j=1}^{J} p_j^{-1} h_{ij} s_{ij} \left(\tau - 1 \left[y_{ij} - \mathbf{x}_{ij}' \tilde{\boldsymbol{\beta}}_\tau \leq 0 \right] \right) \mathbf{x}_{ij} = \mathbf{0} \tag{3.12}$$

All proofs are provided in Appendix B. Note that the outer summation in equation (3.12) is over N, which includes discarded observations. In practice one can use

$$\sum_{i=1}^{N_0} p_j^{-1} \left(\tau - 1[y_{ij} - \mathbf{x}_{ij}' \tilde{\boldsymbol{\beta}}_\tau \leq 0] \right) \mathbf{x}_{ij} = \mathbf{0} \tag{3.13}$$

The asymptotic distribution of $\tilde{\boldsymbol{\beta}}_\tau$ follows from Newey et al. (1994), Theorem 7.1:

Corollary 2. *If the conditions of Newey et al. (1994) Theorem 7.1 are satisfied,* $\sqrt{N}(\tilde{\boldsymbol{\beta}}_\tau - \boldsymbol{\beta}_\tau) \overset{a}{\sim} \mathcal{N}(0, A_2^{-1} B_2 A_2^{-1})$*, where*

$$A_2 = E[f_{y|\mathbf{x}}(\mathbf{x}\boldsymbol{\beta}_\tau)\mathbf{x}\mathbf{x}'] \tag{3.14}$$

and

$$B_2 = \sum_{j=1}^{J} p_j^{-1} E\left[s_{ij}\mathbf{q}\mathbf{q}'\right] \tag{3.15}$$

and q is as defined above.

We estimate (3.14) using

$$\hat{A}_2 = N^{-1} \sum_{i=1}^{N_0} p_j^{-1} \hat{f}_{i,y|\mathbf{x}} \mathbf{x}_{ij} \mathbf{x}'_{ij} \tag{3.16}$$

where $\hat{f}_{y|x}$ is defined above. (3.15) can be estimated using

$$N^{-1} \sum_{i=1}^{N_0} p_j^{-2} \hat{\mathbf{q}}_{ij} \hat{\mathbf{q}}'_{ij} \tag{3.17}$$

Although both (3.16) and (3.17) depend on N, which is typically not observed, these cancel out in the expression of Avar($\tilde{\boldsymbol{\beta}}_\tau$).

3.4 Simulation Results

We compare the performance of the above analytic standard errors to those generated by Stata's qreg command both with and without the "pweight" option. We also compare them to bootstrapped standard errors, where the bootstrapping procedure ignores the sampling scheme, but in each of 1,000 bootstrap replications coefficients are estimated with regard to the sampling scheme, i.e., using the "pweight" option in Stata. Our data generating process follows the multiplicative heteroskedasticity model of Cameron et al. (2009):

$$
\begin{aligned}
y &= 1 + x_1 + x_2 + u \\
u &= (0.1 + 0.5x_1) \times \varepsilon \\
x_1 &\sim \chi_1^2 \\
x_2 &\sim \mathcal{N}(0, 25) \\
\varepsilon &\sim \mathcal{N}(0, 25)
\end{aligned}
$$

An advantage of this DGP is that each quantile is linear in $\mathbf{x}$:

$$Q_\tau(y|\mathbf{x}) = \alpha_\tau + \beta_{1,\tau}x_1 + \beta_{2,\tau}x_2 = [1 + 0.1F_\varepsilon^{-1}(\tau)] + x_1 + \left[1 + 0.5F_\varepsilon^{-1}(\tau)\right] x_2$$

We create a population of 51 strata, with sizes proportional to the population of the 50 US states and the District of Columbia. We present results with both exogenous and endogenous stratification, and under endogenous stratification we present results for two sample sizes. For the case of exogenous stratification, the u are sorted randomly across strata, and the sample size is the smaller of the two. In the case of endogenous stratification, observations are sorted across strata such that the most populous strata have the largest values of u. Since u is correlated with x_1, stratification is not exogenous, and estimators that ignore stratification are inconsistent. The SS sampling case sets $N_j = 20\ \forall j$ for the smaller sample size, and $N_j = 50\ \forall j$ for the larger sample size. For the VP sampling case, we set p_j proportional to the inverse of the population of stratum j, and therefore in expectation each stratum is equally represented in the sample. The scaling factor is set so that $\mathrm{E}(N) = 1,020$ for the smaller sample size, and $\mathrm{E}(N) = 2550$ for the larger sample size. For both SS and VP sampling we draw 10,000 samples from the population.

We present results for infeasible estimates of A_1 and A_2 that rely on knowledge of the true $f_{y|x}$, the standard errors for which are denoted f_i, as well as feasible estimates that use equation (3.9). For the bandwidth, we rely on Stata's three methods, the Hall-Sheather, Bonger, and Chamberlain methods, the standard errors based on which are denoted $\hat{f}_{i,1}$, $\hat{f}_{i,2}$, and $\hat{f}_{i,3}$, respectively. Each bandwidth is a function of $\sum_{i=1}^{N} \text{weight}_{ij}$, and τ, where weight_{ij} is the weight for observation i in stratum j. Thus, for the VP case, the bandwidths are random, since N is random, but in the SS case the bandwidths are not random.

Table 3.1 presents the results for the SS sampling case under exogenous stratification. For reference, the true values of the parameters are listed in the first row of each panel. Throughout the simulation results, estimates that do not have a w subscript use Stata's qreg command without weights, while those with the w subscript use weights. Confirming theory, the unweighted estimates well approximate the true values, and are more precise than the weighted estimates. Since the precision of Quantile Regression estimators is determined in part by the amount of data in the neighborhood of y_i around $y_i - Q_\tau(y|\mathbf{x}) = 0$, the estimators are noisier at the tails, e.g., at the 10^{th} and 90^{th} percentiles.

Stata's unweighted standard errors, which are correct under exogenous stratification, well approximate the standard deviation from the empirical distribution of the unweighted estimators. In contrast, the standard errors that use weights routinely underestimate the standard deviation of the empirical distribution of weighted estimators. Among the estimates of standard errors of weighted estimators, those obtained by bootstrapping perform best. Both the infeasible and the feasible analytic standard errors tend to underestimate variation in the estimators. The bandwidth for $\hat{f}_{i,2}$ at the 25th percentile is approximately 0.35,[6] and therefore it is not possible to estimate $\boldsymbol{\beta}_{0.25-h}$.

Table 3.2 presents results under endogenous stratification. Not surprisingly, estimates of β_1 and α that fail to account for stratification do a poor job, while those that account for stratification well approximate the true values. The coefficients on x_2, which is random across strata, are unaffected by weighting. The variability of $\hat{\beta}_{1,w}$ does not exhibit the symmetric pattern observed under exogenous stratification, and is instead monotonically increasing in τ. This is due to the sampling scheme: the endogenous stratification over samples observations in the neighborhood of y_i around $y_i - Q_{0.10}(y|\mathbf{x}) = 0$, and under samples observations in the neighborhood of y_i around $y_i - Q_{0.90}(y|\mathbf{x}) = 0$.

Stata's weighted standard errors appear completely insensitive to this fact, and instead exhibit a U-shaped pattern.[7] This leads to dramatic overstatement of variability at lower quantiles, and still nonetheless understatement of variability at higher quantiles. The bootstrapped standard errors capture the monotonic pattern of increasing in τ, but for estimates of α and β_1 the bootstrapped standard errors are too low for $\tau = 0.10$, and perform fairly well at $\tau = 0.90$. In stark contrast, both the feasible and infeasible standard errors for α_w and $\beta_{1,w}$ are too high for $\tau = 0.90$, but perform well for lower values of τ.

We present the results for the larger sample size, $N_j = 50$, in Table 3.3. $\hat{\beta}_{1,w}$ is more precisely estimated across quantiles, with the proportional reduction in the empirical

[6] Recall that the bandwidth is not random under SS sampling.

[7] We believe Stata's weighted standard errors have two flaws in implementing Equation (3.11): they use the outer-product of the score, and the weighting factor, Q_j/H_j is not squared.

standard deviations being about constant across τ. The bootstrapped standard errors again tend to overstate variability in estimates of α and $\beta_{1,w}$, particularly at lower values of τ. With the larger sample size, both the feasible and infeasible analytic standard errors tend to outperform bootstrapped standard errors at each value of τ, and for each coefficient.

Table 3.4 presents results for the VP case under exogenous sampling. Again, confirming theory, when stratification is exogenous, unweighted estimates well approximate the true values, and are efficient relative to weighted estimates. The standard deviation from the empirical distribution of the estimates across all 10,000 simulations follows a U-shaped pattern across τ, with less variation at $\tau = 0.50$, and the most variation in the tails. Stata's unweighted standard errors accurately estimate the true variation in estimates, but the weighted standard errors are more than an order of magnitude too small.[8] The bootstrapped and infeasible standard errors perform well, though both underestimate variation of $\tilde{\beta}_1$ at $\tau = 0.90$. The feasible standard errors consistently underestimate variation in estimates.

Results under endogenous stratification with $\mathrm{E}(N) = 1020$ are presented in Table 3.5. Again, as in the SS case, unweighted estimates perform poorly under endogenous stratification, while the weighted estimates well approximate the true values. Variation in the estimates is increasing in τ, which, as in the SS case, is a product of our sampling scheme: observations with y_i in the neighborhood of $y_i - Q_\tau(y|x) = 0$ are systematically under sampled at $\tau = 0.90$, and oversampled at $\tau = 0.10$. Stata's weighted standard errors are again wildly inaccurate. Across all coefficients and values of τ, the bootstrapped standard errors perform quite well, while both the infeasible and feasible analytic standard errors

[8] We obtain numerically identical results to Stata's weighted standard errors when we use p^{-1}, instead of p^{-2}, in Equation (3.17)

tend to understate variation at higher values of τ.

The results from endogenous stratification with $\mathrm{E}(N) = 1020$ are precisely mirrored in the results under endogenous stratification with $\mathrm{E}(N) = 2550$, presented in Table 3.6. Again Stata's weighted standard errors are woefully inaccurate, while bootstrapped standard errors perform quite well across coefficients and values of τ. Both the feasible and infeasible analytic standard errors tend to understate variability for $\tau = 0.90$.

3.5 Conclusion

This paper extends the results from Bartalotti (2012), which addressed the issue of inference for Quantile Regressions when the data are obtained through Standard Stratified sampling, to the case where data are obtained through Variable Probability sampling. We develop the asymptotic distribution of Quantile Regression under VP sampling. Valid estimators for the asymptotic variance matrix of those estimators are provided. Given the insights provided by Quantile Regression, and the fact that many data sets are obtained through complex random sampling techniques, this paper fills an important gap in the literature.

We provide simulation results that confirm theory in showing that unweighted estimates perform well under exogenous stratification, and are in that case efficient relative to weighted estimators. We demonstrate the importance of weighting for consistent estimates under SS and VP sampling when the sampling scheme is endogenous. Under SS sampling, neither bootstrapped nor analytic standard errors are always best, though with larger sample sizes the analytic standard errors tended to do better.

Under VP sampling, bootstrapped standard errors performed best across coefficients, quantiles, and sample sizes, while analytic standard errors underestimated variability around quantiles that were undersampled. A consistent finding throughout the simulation results is that Stata's standard errors are erroneous, and should not be used.

APPENDICES

APPENDIX A - TABLES

Table 3.1: Exogenous SS, Simulation Results

		10th	25th	50th	75th	90th
	β_1	-2.200	-0.690	1.000	2.690	4.200
	$\hat{\beta}_1$	-2.219	-0.699	0.997	2.671	4.192
		(0.300)	(0.239)	(0.226)	(0.238)	(0.301)
	$\hat{\beta}_{1,w}$	-2.180	-0.685	0.992	2.660	4.162
		(0.439)	(0.355)	(0.326)	(0.353)	(0.449)
Standard Errors	Stata's Unweighted	0.299	0.240	0.223	0.238	0.302
	Stata's Weighted	0.291	0.237	0.218	0.235	0.296
	Bootstrapped	0.423	0.357	0.331	0.355	0.430
	$\hat{f}_{i,1}$	0.405	0.342	0.315	0.335	0.399
	$\hat{f}_{i,2}$	0.444		0.322	0.345	0.412
	$\hat{f}_{i,3}$	0.384	0.328	0.306	0.324	0.382
	f_i	0.417	0.343	0.316	0.337	0.398
	β_2	1.000	1.000	1.000	1.000	1.000
	$\hat{\beta}_2$	1.001	1.000	1.000	1.000	1.000
		(0.013)	(0.011)	(0.010)	(0.011)	(0.013)
	$\hat{\beta}_{2,w}$	1.001	1.001	1.001	1.001	1.001
		(0.021)	(0.016)	(0.015)	(0.016)	(0.021)
Standard Errors	Stata's Unweighted	0.013	0.010	0.010	0.010	0.012
	Stata's Weighted	0.012	0.010	0.009	0.010	0.012
	Bootstrapped	0.022	0.017	0.016	0.017	0.022
	$\hat{f}_{i,1}$	0.018	0.014	0.013	0.014	0.017
	$\hat{f}_{i,2}$	0.020		0.014	0.015	0.018
	$\hat{f}_{i,3}$	0.017	0.013	0.012	0.013	0.016
	f_i	0.019	0.016	0.014	0.015	0.019
	α	0.350	0.660	1.000	1.340	1.640
	$\hat{\alpha}$	0.366	0.670	1.006	1.347	1.648
		(0.082)	(0.066)	(0.062)	(0.067)	(0.085)
	$\hat{\alpha}_w$	0.345	0.660	1.002	1.344	1.658
		(0.125)	(0.100)	(0.092)	(0.101)	(0.128)
Standard Errors	Stata's Unweighted	0.083	0.066	0.061	0.067	0.082
	Stata's Weighted	0.082	0.065	0.060	0.065	0.081
	Bootstrapped	0.130	0.104	0.095	0.104	0.131
	$\hat{f}_{i,1}$	0.116	0.094	0.088	0.093	0.110
	$\hat{f}_{i,2}$	0.130		0.090	0.097	0.116
	$\hat{f}_{i,3}$	0.106	0.088	0.082	0.087	0.102
	f_i	0.118	0.096	0.088	0.094	0.112

Note: Estimates come from 10,000 simulations. $\hat{\beta}_1$ is estimated without weights. $\hat{\beta}_{w,1}$ is estimated with Stata's "pweight" option. Numbers in parentheses are standard deviation of estimates across the 10,000 simulations. Bootstrapped standard errors come from 1,000 repetitions, where each draws from the sample with equal probability and uses weighted estimate. $\hat{f}_{i,1}$, $\hat{f}_{i,2}$, and $\hat{f}_{i,1}$ are from Hall-Sheather, Bonger, and Chamberlain methods of estimating bandwidth, while f_i uses known distribution.

Table 3.2: Endogenous SS, Simulation Results

		10^{th}	25^{th}	50^{th}	75^{th}	90^{th}
	β_1	-2.200	-0.690	1.000	2.690	4.200
Standard Errors	$\hat{\beta}_1$	-3.959	-2.835	-1.682	-0.635	0.163
		(0.173)	(0.124)	(0.102)	(0.112)	(0.141)
	$\hat{\beta}_{1,w}$	-2.213	-0.689	0.992	2.686	4.144
		(0.145)	(0.175)	(0.264)	(0.442)	(0.681)
	Stata's Unweighted	0.210	0.153	0.129	0.130	0.163
	Stata's Weighted	0.303	0.241	0.221	0.231	0.277
	Bootstrapped	0.201	0.251	0.384	0.544	0.702
	$\hat{f}_{i,1}$	0.140	0.168	0.250	0.411	0.567
	$\hat{f}_{i,2}$	0.151		0.255	0.426	0.588
	$\hat{f}_{i,3}$	0.137	0.165	0.243	0.390	0.542
	f_i	0.138	0.166	0.248	0.424	0.599
	β_2	1.000	1.000	1.000	1.000	1.000
Standard Errors	$\hat{\beta}_2$	0.999	1.001	1.001	1.001	1.003
		(0.018)	(0.013)	(0.012)	(0.014)	(0.022)
	$\hat{\beta}_{2,w}$	1.001	1.001	1.001	1.001	1.001
		(0.011)	(0.012)	(0.015)	(0.021)	(0.031)
	Stata's Unweighted	0.016	0.013	0.012	0.014	0.021
	Stata's Weighted	0.013	0.010	0.009	0.010	0.012
	Bootstrapped	0.012	0.012	0.016	0.023	0.036
	$\hat{f}_{i,1}$	0.011	0.011	0.013	0.018	0.025
	$\hat{f}_{i,2}$	0.012		0.014	0.019	0.026
	$\hat{f}_{i,3}$	0.010	0.010	0.012	0.017	0.024
	f_i	0.011	0.011	0.014	0.020	0.027
	α	0.350	0.660	1.000	1.340	1.640
Standard Errors	$\hat{\alpha}$	0.206	0.557	0.960	1.387	1.886
		(0.088)	(0.059)	(0.051)	(0.062)	(0.107)
	$\hat{\alpha}_w$	0.360	0.662	1.001	1.346	1.688
		(0.057)	(0.057)	(0.079)	(0.122)	(0.193)
	Stata's Unweighted	0.101	0.078	0.074	0.082	0.135
	Stata's Weighted	0.084	0.066	0.060	0.064	0.081
	Bootstrapped	0.070	0.074	0.101	0.142	0.219
	$\hat{f}_{i,1}$	0.057	0.055	0.073	0.108	0.149
	$\hat{f}_{i,2}$	0.061		0.075	0.115	0.158
	$\hat{f}_{i,3}$	0.054	0.054	0.069	0.098	0.138
	f_i	0.055	0.054	0.073	0.113	0.152

Note: Estimates come from 10,000 simulations. $\hat{\beta}_1$ is estimated without weights. $\hat{\beta}_{w,1}$ is estimated with Stata's "pweight" option. Numbers in parentheses are standard deviation of estimates across the 10,000 simulations. Bootstrapped standard errors come from 1,000 repetitions, where each draws from the sample with equal probability and uses weighted estimate. $\hat{f}_{i,1}$, $\hat{f}_{i,2}$, and $\hat{f}_{i,1}$ are from Hall-Sheather, Bonger, and Chamberlain methods of estimating bandwidth, while f_i uses known distribution.

Table 3.3: Large Sample Endogenous SS, Simulation Results

		10th	25th	50th	75th	90th
	β_1	-2.200	-0.690	1.000	2.690	4.200
	$\hat{\beta}_1$	-3.960	-2.833	-1.681	-0.632	0.168
		(0.107)	(0.078)	(0.066)	(0.072)	(0.089)
	$\hat{\beta}_{1,w}$	-2.212	-0.687	0.992	2.679	4.181
		(0.093)	(0.110)	(0.165)	(0.283)	(0.437)
Standard Errors	Stata's Unweighted	0.131	0.096	0.080	0.081	0.102
	Stata's Weighted	0.190	0.152	0.139	0.148	0.183
	Bootstrapped	0.129	0.159	0.242	0.346	0.462
	$\hat{f}_{i,1}$	0.090	0.109	0.162	0.272	0.401
	$\hat{f}_{i,2}$	0.094		0.164	0.280	0.414
	$\hat{f}_{i,3}$	0.089	0.107	0.159	0.261	0.383
	f_i	0.089	0.108	0.161	0.276	0.410
	β_2	1.000	1.000	1.000	1.000	1.000
	$\hat{\beta}_2$	1.000	1.001	1.001	1.001	1.003
		(0.011)	(0.008)	(0.007)	(0.009)	(0.014)
	$\hat{\beta}_{2,w}$	1.001	1.001	1.001	1.000	1.000
		(0.007)	(0.007)	(0.009)	(0.013)	(0.019)
Standard Errors	Stata's Unweighted	0.010	0.008	0.008	0.009	0.014
	Stata's Weighted	0.008	0.007	0.006	0.006	0.007
	Bootstrapped	0.007	0.008	0.010	0.014	0.020
	$\hat{f}_{i,1}$	0.007	0.007	0.009	0.012	0.015
	$\hat{f}_{i,2}$	0.008		0.009	0.012	0.016
	$\hat{f}_{i,3}$	0.006	0.007	0.008	0.011	0.015
	f_i	0.007	0.007	0.009	0.013	0.018
	α	0.350	0.660	1.000	1.340	1.640
	$\hat{\alpha}$	0.211	0.557	0.958	1.385	1.877
		(0.054)	(0.037)	(0.033)	(0.039)	(0.067)
	$\hat{\alpha}_w$	0.360	0.662	1.001	1.342	1.658
		(0.036)	(0.036)	(0.049)	(0.077)	(0.116)
Standard Errors	Stata's Unweighted	0.063	0.049	0.047	0.052	0.083
	Stata's Weighted	0.053	0.042	0.038	0.041	0.050
	Bootstrapped	0.043	0.046	0.063	0.088	0.126
	$\hat{f}_{i,1}$	0.036	0.035	0.048	0.072	0.099
	$\hat{f}_{i,2}$	0.038		0.049	0.075	0.106
	$\hat{f}_{i,3}$	0.034	0.035	0.046	0.065	0.092
	f_i	0.035	0.035	0.048	0.074	0.105

Note: Estimates come from 10,000 simulations. $\hat{\beta}_1$ is estimated without weights. $\hat{\beta}_{w,1}$ is estimated with Stata's "pweight" option. Numbers in parentheses are standard deviation of estimates across the 10,000 simulations. Bootstrapped standard errors come from 1,000 repetitions, where each draws from the sample with equal probability and uses weighted estimate. $\hat{f}_{i,1}$, $\hat{f}_{i,2}$, and $\hat{f}_{i,1}$ are from Hall-Sheather, Bonger, and Chamberlain methods of estimating bandwidth, while f_i uses known distribution.

Table 3.4: Exogenous VP Sampling, Simulation Results

		10th	25th	50th	75th	90th
	β_1	-2.200	-0.690	1.000	2.690	4.200
	$\tilde{\beta}_1$	-2.223	-0.697	1.001	2.674	4.194
		(0.295)	(0.236)	(0.224)	(0.237)	(0.305)
	$\tilde{\beta}_{1,w}$	-2.186	-0.682	0.992	2.659	4.166
		(0.437)	(0.358)	(0.329)	(0.356)	(0.450)
Standard Errors	Stata's Unweighted	0.300	0.241	0.223	0.238	0.301
	Stata's Weighted	0.015	0.012	0.012	0.012	0.015
	Bootstrapped	0.426	0.357	0.330	0.355	0.431
	$\tilde{f}_{i,1}$	0.367	0.322	0.303	0.318	0.368
	$\tilde{f}_{i,2}$	0.398	0.362	0.316	0.333	0.388
	$\tilde{f}_{i,3}$	0.346	0.302	0.283	0.293	0.359
	f_i	0.427	0.350	0.323	0.345	0.407
	β_2	1.000	1.000	1.000	1.000	1.000
	$\tilde{\beta}_2$	1.001	1.000	1.000	1.000	1.000
		(0.013)	(0.011)	(0.010)	(0.011)	(0.013)
	$\tilde{\beta}_{2,w}$	1.001	1.000	1.001	1.000	1.001
		(0.020)	(0.016)	(0.015)	(0.016)	(0.020)
Standard Errors	Stata's Unweighted	0.013	0.010	0.010	0.010	0.012
	Stata's Weighted	0.001	0.001	0.000	0.001	0.001
	Bootstrapped	0.022	0.017	0.016	0.017	0.022
	$\tilde{f}_{i,1}$	0.016	0.013	0.012	0.013	0.016
	$\tilde{f}_{i,2}$	0.017	0.016	0.013	0.014	0.016
	$\tilde{f}_{i,3}$	0.017	0.014	0.013	0.014	0.017
	f_i	0.020	0.016	0.015	0.016	0.019
	α	0.350	0.660	1.000	1.340	1.640
	$\tilde{\alpha}$	0.367	0.670	1.005	1.347	1.648
		(0.082)	(0.066)	(0.062)	(0.068)	(0.085)
	$\tilde{\alpha}_w$	0.346	0.658	1.001	1.345	1.657
		(0.125)	(0.100)	(0.092)	(0.101)	(0.128)
Standard Errors	Stata's Unweighted	0.083	0.066	0.061	0.067	0.082
	Stata's Weighted	0.004	0.003	0.003	0.003	0.004
	Bootstrapped	0.130	0.104	0.096	0.104	0.131
	$\tilde{f}_{i,1}$	0.098	0.082	0.079	0.082	0.096
	$\tilde{f}_{i,2}$	0.108	0.102	0.087	0.091	0.103
	$\tilde{f}_{i,3}$	0.099	0.083	0.076	0.081	0.099
	f_i	0.120	0.098	0.090	0.096	0.115

Note: Estimates come from 10,000 simulations. $\hat{\beta}_1$ is estimated without weights. $\hat{\beta}_{w,1}$ is estimated with Stata's "pweight" option. Numbers in parentheses are standard deviation of estimates across the 10,000 simulations. Bootstrapped standard errors come from 1,000 repetitions, where each draws from the sample with equal probability and uses weighted estimate. $\hat{f}_{i,1}$, $\hat{f}_{i,2}$, and $\hat{f}_{i,1}$ are from Hall-Sheather, Bonger, and Chamberlain methods of estimating bandwidth, while f_i uses known distribution.

Table 3.5: Endogenous VP Sampling, Simulation Results

		10th	25th	50th	75th	90th
	β_1	-2.200	-0.690	1.000	2.690	4.200
	$\tilde{\beta}_1$	-3.956	-2.830	-1.680	-0.634	0.165
		(0.189)	(0.142)	(0.122)	(0.127)	(0.157)
	$\tilde{\beta}_{1,w}$	-2.212	-0.691	0.989	2.636	4.107
		(0.202)	(0.252)	(0.374)	(0.530)	(0.728)
Standard Errors	Stata's Unweighted	0.210	0.153	0.129	0.130	0.165
	Stata's Weighted	0.016	0.013	0.012	0.012	0.014
	Bootstrapped	0.199	0.249	0.377	0.541	0.710
	$\tilde{f}_{i,1}$	0.191	0.239	0.350	0.458	0.571
	$\tilde{f}_{i,2}$	0.197	0.257	0.362	0.480	0.576
	$\tilde{f}_{i,3}$	0.180	0.223	0.325	0.457	0.556
	f_i	0.198	0.246	0.367	0.512	0.643
	β_2	1.000	1.000	1.000	1.000	1.000
	$\tilde{\beta}_2$	1.000	1.001	1.001	1.001	1.002
		(0.018)	(0.013)	(0.012)	(0.014)	(0.022)
	$\tilde{\beta}_{2,w}$	1.001	1.001	1.000	1.000	1.000
		(0.011)	(0.012)	(0.015)	(0.021)	(0.031)
Standard Errors	Stata's Unweighted	0.016	0.013	0.012	0.014	0.021
	Stata's Weighted	0.001	0.001	0.000	0.001	0.001
	Bootstrapped	0.012	0.012	0.016	0.023	0.036
	$\tilde{f}_{i,1}$	0.010	0.010	0.013	0.017	0.026
	$\tilde{f}_{i,2}$	0.011	0.012	0.013	0.018	0.024
	$\tilde{f}_{i,3}$	0.010	0.010	0.014	0.021	0.029
	f_i	0.012	0.012	0.015	0.020	0.027
	α	0.350	0.660	1.000	1.340	1.640
	$\tilde{\alpha}$	0.204	0.554	0.959	1.389	1.887
		(0.097)	(0.068)	(0.064)	(0.078)	(0.135)
	$\tilde{\alpha}_w$	0.359	0.662	1.002	1.357	1.689
		(0.068)	(0.072)	(0.098)	(0.137)	(0.198)
Standard Errors	Stata's Unweighted	0.101	0.078	0.074	0.082	0.135
	Stata's Weighted	0.004	0.004	0.003	0.003	0.004
	Bootstrapped	0.070	0.074	0.100	0.143	0.223
	$\tilde{f}_{i,1}$	0.060	0.066	0.083	0.110	0.155
	$\tilde{f}_{i,2}$	0.066	0.073	0.090	0.114	0.148
	$\tilde{f}_{i,3}$	0.055	0.060	0.083	0.127	0.169
	f_i	0.067	0.070	0.094	0.127	0.161

Note: Estimates come from 10,000 simulations. $\hat{\beta}_1$ is estimated without weights. $\hat{\beta}_{w,1}$ is estimated with Stata's "pweight" option. Numbers in parentheses are standard deviation of estimates across the 10,000 simulations. Bootstrapped standard errors come from 1,000 repetitions, where each draws from the sample with equal probability and uses weighted estimate. $\hat{f}_{i,1}$, $\hat{f}_{i,2}$, and $\hat{f}_{i,1}$ are from Hall-Sheather, Bonger, and Chamberlain methods of estimating bandwidth, while f_i uses known distribution.

Table 3.6: Large Sample Endogenous VP Sampling, Simulation Results

		10th	25th	50th	75th	90th
	β_1	-2.200	-0.690	1.000	2.690	4.200
	$\tilde{\beta}_1$	-3.959	-2.831	-1.679	-0.632	0.171
		(0.119)	(0.091)	(0.077)	(0.081)	(0.099)
	$\tilde{\beta}_{1,w}$	-2.210	-0.687	0.991	2.657	4.154
		(0.129)	(0.157)	(0.237)	(0.338)	(0.465)
Standard Errors	Stata's Unweighted	0.132	0.096	0.081	0.081	0.102
	Stata's Weighted	0.017	0.013	0.012	0.012	0.015
	Bootstrapped	0.128	0.159	0.240	0.344	0.464
	$\tilde{f}_{i,1}$	0.125	0.156	0.230	0.308	0.388
	$\tilde{f}_{i,2}$	0.127	0.164	0.235	0.323	0.416
	$\tilde{f}_{i,3}$	0.120	0.150	0.216	0.300	0.377
	f_i	0.127	0.157	0.236	0.333	0.435
	β_2	1.000	1.000	1.000	1.000	1.000
	$\tilde{\beta}_2$	1.000	1.001	1.001	1.001	1.003
		(0.011)	(0.008)	(0.007)	(0.009)	(0.014)
	$\tilde{\beta}_{2,w}$	1.001	1.001	1.001	1.001	1.001
		(0.007)	(0.007)	(0.009)	(0.013)	(0.019)
Standard Errors	Stata's Unweighted	0.010	0.008	0.008	0.009	0.102
	Stata's Weighted	0.001	0.001	0.000	0.000	0.001
	Bootstrapped	0.007	0.008	0.010	0.014	0.020
	$\tilde{f}_{i,1}$	0.006	0.007	0.008	0.011	0.015
	$\tilde{f}_{i,2}$	0.007	0.008	0.009	0.011	0.015
	$\tilde{f}_{i,3}$	0.006	0.006	0.008	0.011	0.017
	f_i	0.007	0.007	0.009	0.013	0.018
	α	0.350	0.660	1.000	1.340	1.640
	$\tilde{\alpha}$	0.210	0.556	0.958	1.386	1.877
		(0.059)	(0.043)	(0.040)	(0.049)	(0.085)
	$\tilde{\alpha}_w$	0.360	0.662	1.001	1.346	1.662
		(0.042)	(0.045)	(0.061)	(0.087)	(0.121)
Standard Errors	Stata's Unweighted	0.063	0.049	0.047	0.052	0.083
	Stata's Weighted	0.004	0.004	0.003	0.003	0.004
	Bootstrapped	0.043	0.046	0.062	0.089	0.127
	$\tilde{f}_{i,1}$	0.039	0.044	0.057	0.070	0.091
	$\tilde{f}_{i,2}$	0.042	0.047	0.060	0.078	0.098
	$\tilde{f}_{i,3}$	0.036	0.040	0.050	0.071	0.097
	f_i	0.042	0.045	0.060	0.083	0.109

Note: Estimates come from 10,000 simulations. $\hat{\beta}_1$ is estimated without weights. $\hat{\beta}_{w,1}$ is estimated with Stata's "pweight" option. Numbers in parentheses are standard deviation of estimates across the 10,000 simulations. Bootstrapped standard errors come from 1,000 repetitions, where each draws from the sample with equal probability and uses weighted estimate. $\hat{f}_{i,1}$, $\hat{f}_{i,2}$, and $\hat{f}_{i,1}$ are from Hall-Sheather, Bonger, and Chamberlain methods of estimating bandwidth, while f_i uses known distribution.

APPENDIX B - PROOFS

Proof of Corollary 1. It suffices to show that (3.12) converges in probability to (3.5). Using the facts that h_{ij} is independent of $(s_{ij}, y_{ij}, \mathbf{x}_{ij})$, $\mathrm{E}(h_{ij}) = p_j$, and $\sum_{j=1}^{J} s_{ij} = 1$, we have the following:

$$
\begin{aligned}
N^{-1}\sum_{i=1}^{N}\sum_{j=1}^{J} & p_j^{-1} h_{ij} s_{ij}\left[\tau - 1\left[y_{ij} - \mathbf{x}_{ij}'\boldsymbol{\beta}_\tau \le 0\right]\right]\mathbf{x}_{ij} \\
& \xrightarrow{p} \sum_{j=1}^{J} p_j^{-1}\mathrm{E}\left[h_{ij}s_{ij}(\tau - 1[y_{ij} - \mathbf{x}_{ij}'\boldsymbol{\beta}_\tau \le 0])\mathbf{x}_{ij}\right] \\
& = \sum_{j=1}^{J} p_j^{-1}\mathrm{E}(h_{ij})\mathrm{E}\left[s_{ij}(\tau - 1[y_{ij} - \mathbf{x}_{ij}'\boldsymbol{\beta}_\tau \le 0])\mathbf{x}_{ij}\right] \\
& = \mathrm{E}\left[\sum_{j=1}^{J} s_{ij}(\tau - 1[y_{ij} - \mathbf{x}_{ij}'\boldsymbol{\beta}_\tau \le 0])\mathbf{x}_{ij}\right] \\
& = \mathrm{E}\left[(\tau - 1[y_{ij} - \mathbf{x}_{ij}'\boldsymbol{\beta}_\tau \le 0])\mathbf{x}_{ij}\right]
\end{aligned}
$$

□

Proof of Corollary 2. To apply the results from Newey et al. (1994), note that

$$
\begin{aligned}
\nabla_{\boldsymbol{\beta}_\tau} & \sum_{j=1}^{J}\mathrm{E}\left[p_j^{-1}h_{ij}s_{ij}\left(\tau - 1\left[y_{ij} - \mathbf{x}_{ij}'\boldsymbol{\beta}_\tau \le 0\right]\right)\mathbf{x}_{ij}\right] \\
& = \nabla_{\boldsymbol{\beta}_\tau}\sum_{j=1}^{J}\mathrm{E}\left[s_{ij}\left(\tau - 1\left[y_{ij} - \mathbf{x}_{ij}'\boldsymbol{\beta}_\tau \le 0\right]\right)\mathbf{x}_{ij}\right]
\end{aligned}
$$

$$
\begin{aligned}
&= \bigtriangledown_{\boldsymbol{\beta}} \, \mathrm{E}\left[\left(\tau - 1\left[y_{ij} - \mathbf{x}_{ij}'\boldsymbol{\beta}_\tau \leq 0\right]\right)\mathbf{x}_{ij}\right] \\
&= \bigtriangledown_{\boldsymbol{\beta}_\tau} \, \mathrm{E}\left[\mathrm{E}\left[\left(\tau - 1\left[y_{ij} - \mathbf{x}_{ij}'\boldsymbol{\beta}_\tau \leq 0\right]\right)\mathbf{x}_{ij} \mid \mathbf{x}\right]\right] \\
&= \bigtriangledown_{\boldsymbol{\beta}_\tau} \, \mathrm{E}\left[\tau - F_{y|x}(\mathbf{x}_{ij}'\boldsymbol{\beta}_\tau)\mathbf{x}_{ij}\right] \\
&=\mathrm{E}\left[f_{y|x}(\mathbf{x}_{ij}'\boldsymbol{\beta})\mathbf{x}_{ij}\mathbf{x}_{ij}'\right] = A_{VP}
\end{aligned}
$$

and

$$
B_{VP} = \quad \mathrm{E}\left[\left(\textstyle\sum_{j=1}^{J} p_j^{-1} h_{ij} s_{ij} \mathbf{q}_{ij}\right)' \left(\textstyle\sum_{j=1}^{J} p_j^{-1} h_{ij} s_{ij} \mathbf{q}_{ij}\right)\right]
$$

Cross products cancel out since $h_{ij}s_{ij}h_{km}s_{km} = 0$ For any $k \neq i$ or $m \neq j$. Note that $(h_{ij}s_{ij})^2 = h_{ij}s_{ij}$.

$$
B_{VP} = \mathrm{E}\left[\sum_{j=1}^{J} p_j^{-2} h_{ij} s_{ij} \mathbf{q}_{ij}' \mathbf{q}_{ij}\right] = \sum_{j=1}^{J} p_j^{-2} \mathrm{E}\left[h_{ij} s_{ij} \mathbf{q}_{ij}' \mathbf{q}_{ij}\right] = \sum_{j=1}^{J} p_j^{-1} \mathrm{E}\left[s_{ij} \mathbf{q}_{ij}' \mathbf{q}_{ij}\right]
$$

□

BIBLIOGRAPHY

BIBLIOGRAPHY

Albrecht, James, Anders Björklund, and Susan Vroman (2003). Is there a glass ceiling in Sweden? *Journal of Labor Economics* 21.1, pp. 145–177.

Bartalotti, Otávio (2012). "Essays in econometrics". UMI Number 3509178.

Buchinsky, Moshe (1998). Recent advances in Quantile Regression models: A practical guideline for empirical research. *The Journal of Human Resources* 33.1, pp. 88–126.

Buchinsky, Moshe (2001). Quantile regression with sample selection: Estimating women's return to education in the U.S. *Empirical Economics* 26.1, pp. 87–113.

Cameron, Adrian Colin and Pravin K Trivedi (2009). *Microeconometrics Using Stata.* Vol. 5. Stata Press College Station, TX.

Koenker, Roger (2005). *Quantile Regression.* 38. Cambridge university press.

Machado, José AF and José Mata (2005). Counterfactual decomposition of changes in wage distributions using quantile regression. *Journal of Applied Econometrics* 20.4, pp. 445–465.

Martins, Pedro S and Pedro T Pereira (2004). Does education reduce wage inequality? Quantile regression evidence from 16 countries. *Labour Economics* 11.3, pp. 355–371.

Melly, Blaise (2005). Decomposition of differences in distribution using quantile regression. *Labour Economics* 12.4, pp. 577–590.

Newey, Whitney K and Daniel McFadden (1994). Large sample estimation and hypothesis testing. *Handbook of Econometrics* 4, pp. 2111–2245.

Powell, James L (1991). Estimation of monotonic regression models under quantile restrictions. *Nonparametric and Semiparametric Methods in Econometrics,(Cambridge University Press, New York, NY)*, pp. 357–384.

Solon, Gary, Steven J Haider, and Jeffrey Wooldridge (2013). *What are we weighting for?* Tech. rep. National Bureau of Economic Research.

Wooldridge, Jeffrey M (1999). Asymptotic properties of weighted M-estimators for variable probability samples. *Econometrica* 67.6, pp. 1385–1406.

Wooldridge, Jeffrey M (2001). Asymptotic properties of weighted M-estimators for standard stratified samples. *Econometric Theory* 17.02, pp. 451–470.

Wooldridge, Jeffrey M (2010). *Econometric analysis of cross section and panel data.* Second Edition. MIT Press.