Estimating General Equilibrium Effects of Major Education Reforms*

Michael Gilraine, New York University
Hugh Macartney, Duke University and NBER
Robert McMillan, University of Toronto and NBER

December 23, 2018

Abstract

This paper sets out an approach for credibly estimating general equilibrium sorting effects of major education reforms. We develop our strategy in the context of California's vast class size reduction program, and use multiple differencing to identify the reform's direct and indirect effects on a common scale (using test scores), along with their persistence. The estimates reveal a large direct class size effect and an even larger indirect effect via changes in school demographics. Further, both effects persist, generating a substantial longer-run policy impact. Our approach is applicable more broadly to large-scale reforms that cover a subset of school grades.

Keywords: Education Reform, General Equilibrium, Education Production, Sorting, Class Size Reduction, Persistence, Multiple Differencing, Difference-in-Differences

JEL Classifications: H40, I21, I22

*This paper is an updated version of NBER Working Paper 24191. We would like to thank Pat Bayer and Gregorio Caetano for helpful discussions, and Damon Clark, Ted Rosenbaum, Wilbert van der Klaauw, and workshop participants at the University of Bristol (CPMO), the University of Toronto, the 2018 SOLE Meetings, the Federal Trade Commission, the CIREQ Applied Economics Conference, and the New York Federal Reserve for additional comments. Financial support is gratefully acknowledged from the IES, SSHRC, Duke University, and the University of Toronto Mississauga. All remaining errors are our own. Contact emails: Gilraine, mike.gilraine@nyu.edu; Macartney, hugh.macartney@duke.edu; McMillan, mcmillan@chass.utoronto.ca.

1 Introduction

Empirical policy analysis often focuses on the direct, intended effects of policies "holding all else equal." Measuring such direct effects accurately is an important ingredient in policy making, although it is well appreciated that large-scale reforms may also have *indirect* general equilibrium impacts that work in offsetting or reinforcing ways. Because of their potential to alter the policy-making calculus significantly, estimating the size of these indirect effects is of considerable interest. Yet doing so presents challenges for empirical research, given that they cannot be identified separately from the direct effects without additional sources of independent variation. As a consequence, the literature seeking to gauge their extent is relatively undeveloped.[1]

This paper sets out an estimation approach to help fill the gap – one that allows us to measure indirect effects of large-scale reforms in a credible way and place them on a common footing with the direct effects of policy for the first time. We focus on an education context and a type of general equilibrium response likely to matter whenever (i) a reform improves public school quality significantly – the basic goal of most education reforms, and (ii) private school options are popular pre-reform.[2] In this common configuration of circumstances, some households will tend to re-sort by switching out of private schools, potentially changing the public school student mix. In turn, to the extent that induced compositional changes influence education production (either directly or via peer interactions), they should affect measured outcomes – the indirect effects we seek to identify.

We develop our approach in the context of California's class size reduction (CSR) program of the late-1990s – up to that point, the largest state-led education reform ever implemented in the United States.[3] Inspired by Project STAR in Tennessee, a well-known experimental evaluation in education and the subject of a number of influential studies,[4] the California legislature sought to replicate Project STAR's publicized experimental benefits at an altogether larger scale. To that end, the CSR program targeted kindergarten through third grade (as did Project STAR), and cut class sizes in these early grades by a substantial amount throughout the state.

Large-scale reforms often have institutional features that can be used to identify policy impacts. This is certainly the case with CSR, which involved very specific timing in its implementation. Smaller classes were phased in exclusively for first grade during the 1996-97 school year, second

[1] Notable existing contributions include papers by Jepsen and Rivkin (2009) and Dinerstein and Smith (2016), discussed below, and Bianchi (2017), who analyzes indirect effects arising from an Italian university reform.

[2] Other forms of general equilibrium response include effects on teacher labor markets and household residential sorting. Our emphasis is on providing clean identification of a general equilibrium response that turns out to be quantitatively important.

[3] This assessment comes from the 1998/99 report of the CSR Research Consortium.

[4] Prominent among these are Krueger (1999), Krueger and Whitmore (2001), and Chetty et al. (2011).

grade became eligible in 1997-98, and schools could then seek to reduce class sizes in either kindergarten or third grade the following year (and the remaining grade the year after). In order to be eligible for CSR funding, schools had to hire sufficient teachers to lower class sizes below 20, and were required to do so in the grade-specific order of the roll-out.

Given the combination of the strong financial incentives (around $650 per student[5]) to implement the reform according to this timetable, the substantial reductions in class size it produced – a 20 percent reduction on average in elementary schools across the state – and the sheer scale of California's education system, one would expect CSR to have broader effects. Compelling work by Jepsen and Rivkin (2009) has already highlighted impacts on the teacher labor market, showing that there was a sudden, significant increased need for new teacher hires, which dampened CSR's benefits in the short term. Further, the effects of the policy were non-uniform, with some schools experiencing reductions in class size without any decline in teacher quality – changes that might be expected to induce sociodemographic sorting from private to public schools in response to CSR.

We first document that significant general equilibrium sorting did indeed occur. A difference-in-differences research design that compares treated versus untreated grades before and after the reform reveals that CSR caused a sizeable reduction in local private school shares of 1.4 percentage points for relevant elementary grades[6] and also pronounced changes in public school sociodemographics. Thus the reform led to an inflow of students into public schools and especially students of higher socioeconomic status, findings complementing recent research by Dinerstein and Smith (2016).[7]

This evidence of general equilibrium sorting responses motivates a framework for estimating both the policy's direct and indirect effects that is central to our approach. The framework parameterizes the public school education production function to capture the reform's effects in terms of test scores, contemporaneously and also carrying over into the future, as in influential recent research (see Chetty et al. 2011 and Chetty, Friedman and Rockoff 2014). Specifically, we assume the technology is linear and additive (in line with much of the education literature), and express each cohort's grade-year average test score as a function of parameters governing the current and persisting effects of school resources and sociodemographics.

We show the direct and indirect sorting effects of the reform can be identified using a multiple differencing approach. This draws on two main assumptions, both supported in the data – that

[5]The average per-pupil expenditures in 1995-96 were $6,068.

[6]This is equivalent to 12 percent of the pre-CSR K-3 average private school share of 11.7 percent and a sixth of its standard deviation.

[7]Those authors provide persuasive evidence that increased funding for public schools in New York City drew private school students into the public system, both by choice and through the forced closure of (typically small) private schools.

differences across grades arising independently of the reform are time-invariant, and that there are no input differences across treated cohorts, controlling for teacher quality. Using observations for untreated cohorts to serve as controls for their treated counterparts, our framework generates a system of linear equations depending on the unknown parameters. We show that the *direct* effect of the reform can be recovered, given the additive linear structure and the identical treatment assumption, by taking differences across adjacent cohorts in the same academic year, leveraging their differential exposure to the reform over time. For the *indirect* effect, we draw on exogenous variation associated with elementary school grade spans. Specifically, we use the fact that students in areas served by K-5 schools have earlier opportunities to move back to the private sector for middle school than students attending K-6 public schools, generating exogenous differences in school compositions. The indirect sorting effect is then isolated by contrasting the performance of sixth grade students under the two types of grade configuration.[8]

Our estimates show that both direct and indirect effects are significant and positive.[9] Of note, the precisely estimated indirect sorting effect is even greater in magnitude (0.16σ in terms of mathematics scores) than the direct effect (0.11σ).[10] Further, we find that both the direct and indirect effects persist into subsequent years, at approximately the same rate in each case. The findings are important from a measurement perspective, as estimates of the direct effect would be biased (as we demonstrate) when applying a regular difference-in-differences approach in the presence of general equilibrium sorting or persistence. Together, the parameter estimates and our framework indicate that the overall benefits of CSR in the longer term are substantial when accounting for general equilibrium sorting and persistence.[11]

The magnitudes of our direct and indirect estimates in the context of CSR support the view that researchers should treat household sorting as a primary factor when assessing the impact of large-scale reforms that alter public school quality. This is especially likely to be the case when – as we have highlighted – private school enrollment is high pre-reform, and a large number of households are on the margin of switching, as in a California context (judging by the switching behavior in response to CSR). These conditions hold in many US states, for example. Our analysis

[8] In utilizing distinct sources of variation to identify the separate effects of interest, our estimation approach resembles the influential analysis of Boston's Metco desegregation program by Angrist and Lang (2004), differences in the types of policy and the goals of the two studies to one side.

[9] Across a variety of settings, studies in the class size literature have found positive effects on achievement (for instance, Krueger 1999, Angrist and Lavy 1999, Krueger and Whitmore 2001, and Gilraine 2017) as well as no effects (including Hoxby 2000 and Angrist, Battistin and Vuri 2017). Research studying the Californian context has delivered mixed results similar to the broader literature (see Bohrnstedt and Stecher 2002 for non-effects, and Unlu 2005 and Jepsen and Rivkin 2009 for positive test score effects).

[10] A bounding exercise in Section 6.3 indicates that peer spillovers are likely to account for a significant portion – well over half – of the indirect sorting effect, with an implied social multiplier in line with Graham (2008).

[11] See Section 6.2. There, we also show that the indirect effect inflates recurrent costs by 20 percent. Using estimates provided in Chetty et al. (2011), we compute a suggestive net benefit-cost ratio that is well above one.

thus indicates that studies ignoring sorting effects are likely to misstate the overall impact of major education reforms to a high degree in often-encountered circumstances.

From a methodological perspective, the estimation approach we develop is applicable when two sources of variation are available: First, a major reform applies to some groups (in our case, school grades) and not others – given constrained public funds, this is a common occurrence. Second, there is variation in school grade spans (used to identify indirect sorting effects); in practice, this is a widespread feature of many public education systems. Our estimation approach is relevant in schooling applications given that persistence is a key aspect of education production, where the stock of knowledge naturally accumulates, and because both indirect effects and persistence are found to matter; in such circumstances, we show simpler designs produce biased estimates. Further, the data requirements of the approach are quite minimal and are likely to be satisfied in many settings; all the data in this study are observational, involve school-grade-year averages and are available publicly, for example. Taking these elements together, our paper provides researchers with a transparent means of gauging the extent of indirect general equilibrium sorting effects alongside the direct effects of policy in a variety of settings.

The rest of the paper is organized as follows: Section 2 describes the conceptual framework that shapes our empirical approach, and mechanisms likely to operate following CSR's introduction. Section 3 sets out the institutional background to CSR in California, and the data set we have assembled. In Section 4, we show reduced-form evidence indicating that the reform caused significant general equilibrium sorting. This motivates the multiple differencing approach at the heart of the paper, set out in Section 5. Estimates using the approach are presented and interpreted in Section 6, which also develops policy implications, and Section 7 concludes.

2 Conceptual Framework

We describe a standard policy setting in which general equilibrium sorting effects may arise following a major reform. For concreteness, we focus on a state's education system (such as California's), where the students are served by public and private schools. We first present a formal framework based on an education production function, which we use to define the empirical quantities of interest. Then we discuss the mechanisms one might expect to operate following a large-scale reform like CSR, anticipating the empirical analyses that follow.

2.1 Education Production Technology

Consider an environment in which an outcome y depends on inputs consisting of a policy variable and a set of other relevant factors. A reform is implemented through a change in this policy variable, which may give rise to direct and indirect effects. We will think of outcomes primarily as test scores, given the education setting we focus on. In response to a major policy change, the policy's effects can then be understood in terms of an education production technology in which education inputs together affect measured test score performance – in our application, the *public school* technology.[12]

Our interest is in uncovering the parameters of the technology. To that end, we will make explicit the assumptions that allow us to apply a transparent multiple differencing approach developed in Section 5 below. We focus on a specification in which education output is affected by three main inputs $\{R, X, Q\}$, where R measures school resources (under the direct control of the policy maker), X represents productive student characteristics (their ability and possible peer interactions) in the school, and Q is teacher quality. There is also a further noise component, ϵ, reflecting unobservable random influences on contemporaneous test scores.

Assumption 1: The production technology is linear.

We approximate the true education production function, following the bulk of education literature, with a technology that is additive in its inputs and the error.

Assumption 2: The production technology is cumulative.

We allow current inputs in one period to have persistent impacts in subsequent periods as students acquire (and retain) knowledge and skills, in light of compelling recent evidence that education inputs have cumulative effects (see Rivkin, Hanushek and Kain 2005, and Chetty, Friedman and Rockoff 2014, for example).

Based on these two assumptions, the following technology accounts for the current and persistent effects of class size, student sociodemographics and teachers on scores in period t:

$$y_t = \gamma_R \sum_{\tau=0}^{L} (\delta_R)^\tau R_{t-\tau} + \gamma_X \sum_{\tau=0}^{L} (\delta_X)^\tau X_{t-\tau} + h(Q_t) + \epsilon_t \,. \tag{2.1}$$

This specification will provide the basis for our estimation approach, chosen in light of what we will be able to reasonably identify using data aggregated to the school-cohort level and the sources of exogenous variation we have access to.

[12] In the background, there is a local education market with public and private schooling options. Public schools are free and have to admit all students who wish to enroll, typically within a given attendance zone around the school. Private schools, in contrast, charge tuition and can be selective.

The large-scale reform can be represented by the change in school resources (ΔR) associated with the policy intervention – for example, comparing before and after.[13] The introduction of the reform is expected to have two contemporaneous effects on learning: (1) the *direct* effect, given by the product of the γ_R parameter and the change in resources ΔR; and (2) the *indirect* general equilibrium effect arising from student sorting between private and public school systems, given by the product of the γ_X parameter and the induced change in student composition ΔX (associated with ΔR) due to a demand shift.[14]

A demographic change in the composition of students in the public system may affect test score outcomes through two channels. First, if the incoming students are of higher ability than students already enrolled in public school and score more highly themselves, then outcomes will improve through what we term the 'own' effect. Second, the change in the demographic composition of public schools may result in *spillover* benefits to incumbent public school students, perhaps via positive peer influences in the classroom. The aggregate nature of our data – described in Section 3.1 – limits our capacity to separate these two channels, so we will combine them for the most part. Below (see Section 6.3), we will be able to conduct an informative bounding exercise, however.

An additional effect may arise through induced changes in teacher quality Q_t, which influences test scores in equation (2.1) through $h(Q_t)$. We omit teacher quality from the current discussion for clarity, returning to it in Section 5; as explained there, we will account for the impact of teacher quality non-parametrically, and use a control technique to address changes in it.

In terms of *persistence* in our specification, the effect of past resources (smaller classes) on current test scores is parametrized by δ_R – a parameter of interest in prior research (see, for example, Krueger and Whitmore 2001 and Ding and Lehrer 2010). Specifically, δ_R measures the persistent effect on test scores of a one-unit increase in resources one period ago into the present. Further, resources from at most L periods ago are allowed to influence current test scores, following a geometric decay. Similarly, the parameter δ_X captures the persistent effects of induced prior school demographic compositions on current test scores. Adding persistence allows direct and indirect effects from earlier periods to accumulate over time: the structure implies expressions for each effect.[15]

[13] We will think of the extra resources as appearing exogenously. This corresponds reasonably well to the case of the California CSR reform.

[14] As public school quality rises, so more households will prefer the public school option, leading to an increase in enrollment and the change in X. Clearly, to the extent that public and private school student populations differ initially, so the inflow to the public system will change school demographics more.

[15] For example, a shock to class sizes l periods ago will give rise to an indirect sorting effect in the current period of $\gamma_X \Delta X (\delta_X)^l$, where ΔX measures the induced within-period sorting response l periods ago. In turn, taking a forward-looking perspective, a class size shock at t will have a total indirect sorting effect on scores that propagates into the future in an amount equal to $\gamma_X \Delta X_t \sum_{\tau=0}^{L} (\delta_X)^\tau$. We use the estimates and this structure to compute the

2.2 Looking Ahead to the Empirics

The main empirical goals of the paper are to estimate, on a common scale, the direct and indirect effects of major education reforms, along with their persistence. To those ends, we will take the framework to the data in the context of California's CSR, described in detail in the next section.

California features a mixture of public and private schools, as we will document, with the two populations differing markedly, in ways one would expect. The change in the public system associated with CSR was substantial, with early grades in public schools experiencing significant reductions in class size. The implied relative changes in quality should boost demand for public schools in the affected grades at the expense of private schools: we will show causal evidence of that in terms of enrollment changes in Section 4. Further, given that costs of switching are not zero, it is plausible to think that students induced into the public system as a result of CSR might remain there even after the class size benefit ended (in fourth grade and above): we are able to shed light on the importance of this type of pattern using exogenous variation in public school grade spans, which will play an important role in our identification strategy.

3 Institutional Background and Data

In this section, we discuss the policy context and relevant institutional background to the CSR reform. Doing so serves to emphasize both that the reform was very large in scope and so likely to have broader effects, and that it was rolled out in a particular way, useful for applied research. We also discuss the data we have assembled – the sources and descriptive evidence.

California's CSR program was introduced in the spring of 1996, and was vast in scale, being the largest state-led education reform implemented in the United States up to that time.[16] Impetus for the reform arose in the wake of disappointing national test score rankings four years earlier, when National Assessment of Educational Progress (NAEP) scores became available on a state-by-state basis for the first time. These revealed California to be among the worst-performing states in both mathematics and reading. Further, it became clear that the low performance issue was persistent.[17]

California lawmakers, motivated in part by Project STAR,[18] enacted the class size reduction reform in a bid to address these problems in July 1996. While the policy was widely supported by

overall test score benefits of CSR in Section 6.2.

[16] After full implementation, California's CSR program cost about $1.5 billion each year. Following California, several large-scale CSR programs were implemented, with the federal government spending $1.2 billion a year from 1999 to 2001 on class size reduction, and Florida instituting a CSR program in 2002 that cost over $2 billion a year.

[17] For instance, the 1994 NAEP results showed California to be the very bottom state (along with Louisiana) in fourth grade reading, and in 1996, it tied with Tennessee at the bottom of the eighth grade mathematics rankings.

[18] See, for example, a report from the associated legislative discussions, available at `http://files.eric.ed.gov/fulltext/ED407699.pdf`.

both parents and teachers, fierce disagreement between the Republican Governor and the California Teachers' Association over education policy meant that its implementation did not arise in a consensual way, with the Governor adamant that extra funding available from the state's budget surplus in the mid-1990s (which by a 1988 constitutional initiative had to be spent on education) would not be used as discretionary funding that could flow into higher teacher salaries. To ensure this, the Governor avoided funding the union-dominated education boards by arranging to give the money directly to schools that had class sizes below a certain threshold.

The Reform: The reform provided targeted incentives to reduce class sizes in early grades from a statewide average of 28.5 down to 20.[19] For the first year of operation, school year 1996-97, the program applied only to first graders (as noted in the Introduction). Second grade classes then became subject to the program incentives in the following year (1997-98), and schools were able to choose to implement CSR in either kindergarten or third grade beginning in 1998-99. We exploit this differential timing of implementation by grade when studying changes in private school share, sociodemographic compositions, and test scores.

Even though participation was voluntary, substantial financial payments of $650 per pupil enrolled in a class of 20 or fewer students (relative to average 1995-96 per-pupil expenditures of $6,068) led to nearly universal adoption by districts and schools such that 88 percent of first graders were in a CSR-compliant class in the first year of the reform.[20] Forty-two (out of 895) districts did not implement CSR in the first year, either because (1) they had class sizes just above twenty and did not think it was worth seeking the extra funding to hire a new teacher, or (2) they already had many class sizes below twenty and did not realize they were eligible.[21] At the school level, around 10 percent of schools delayed their implementation of CSR, primarily because of a lack of space.[22] Due to endogeneity concerns, we assume all schools and districts that were eligible did adopt, with school adoption decisions only being used to show robustness of our main findings (see the triple-differences designs in Appendix B and Appendix C).[23]

The coverage of the policy is shown in Figure 1, making clear the nature of the roll-out. In line with that, Figure 2 highlights the broad impact of CSR that we exploit, plotting student-to-teacher

[19] This subsection draws on the lively account of the background to CSR in Schrag (2006). As detailed there, an unidentified staffer for the Governor stated that the class size goal of 20 was set based primarily on what could be afforded.

[20] For districts to participate in CSR, they only needed to opt into the program, whereas schools only received CSR funding if they reduced class sizes in the relevant grade.

[21] See `http://www.lao.ca.gov/1997/021297_class_size/class_size_297.html`.

[22] In a survey by the CSR Research Consortium, eighty percent of principals who had not implemented CSR stated that space issues were the main impediment. See `http://www.classize.org/summary/97-98/summaryrpt.pdf`.

[23] The assumption of full adoption will lead us to understate the true effects (direct and indirect) of CSR, the implementation rates suggesting true effects that are underestimated by around ten percent.

ratios in elementary and middle schools for school years 1990-91 through 2006-07. It shows a clear drop in student-to-teacher ratios for elementary schools when CSR was implemented in 1996-97, with no comparable change in the middle schools.

Several factors make studying CSR challenging. First, despite the scale of the reform, no systematic program evaluation method was put in place.[24] This was a consequence of the initial announcement and roll-out of the actual policy being sudden and unanticipated, generating headlines such as "Sacramento Surprise – Extra Funds / Governor wants to use money to cut class size" in the *San Francisco Chronicle* (Lucas 1996). (We note it also means that no districts, schools or parents could anticipate the reform's introduction.) Second, in terms of measuring student performance, student testing did not begin until the 1997-98 school year, when the Standardized Testing and Reporting Program – another initiative of the Republican Governor – began. Thus, researchers do not have access to a comparable pre-reform test.[25] We address this issue by using exogenous differences in treatment, described below. Third, additional data limitations include a lack of individual student or classroom-level data and an inability to track teachers over time.[26] Our measurement approach makes use of data aggregated to the school-grade-year level: we will discuss in Section 5 how such aggregated data can still be used to identify the effects of interest based on a multiple differencing strategy we propose.

3.1 Data

The main data set we have assembled draws on several useful public data sources provided by the California Department of Education (CDE). (Appendix Table A.1 offers a more detailed description.) The first consists of the student enrollment for all public schools and districts at the grade level from the 1990-91 through the 2008-09 school years.[27] We augment these enrollment data with additional CDE data describing demographic characteristics, including race, 'English as a Second Language' (ESL) status, and Free or Reduced-Price Meal status.[28] Second, the CDE also provides grade-level enrollment data for private schools from 1990-91 to 2008-09 inclusive; no demographics are available beyond overall private school enrollments from this source. Together, these two data

[24] The legislature did create the CSR Research Consortium to conduct a four-year comprehensive study to evaluate the implementation and impact of CSR, though it had to confront the same data limitations that we highlight.

[25] Earlier tests in the state – the CLAS test, for instance – were discontinued in the face of budget cuts and union resistance. Appendix A offers a quick primer on California statewide testing.

[26] California's teacher identifiers were scrambled each year to prevent following the same teacher over time. They continue to be scrambled in the statewide files to the present.

[27] We stop in 2008-09 due to a CSR funding formula change in the following academic year so that schools would not lose all their CSR funding if class sizes exceeded twenty students. This caused a substantial rise in K-3 class sizes.

[28] This serves as a measure of the poverty rate of the entire student body since only students whose household income is below a threshold based on a percentage of the poverty line are eligible. It is not available at the grade level, unlike our other public school demographic variables.

sets allow us to study the effects of CSR on local private school shares, starting well before CSR's introduction – this will be an advantage relative to the test score information that is available.

The third data source provides test score data from California's Standardized Testing and Reporting Program (STAR) for second grade and higher. All students in second through eleventh grade took the Stanford Achievement Test in both mathematics and English near the end of the academic year (with some minor exceptions[29]). The Stanford Achievement Test was a national norm-referenced multiple-choice test introduced in the 1997-98 school year. Because the policy was in place for first grade since the 1996-97 school year and included second grade beginning in 1997-98, we do not observe a purely pre-reform period in terms of test scores. Thus, identifying the effect of CSR on test scores necessarily involves exploiting differences in treatment over time once the reform came into effect; our estimation strategy is designed to use that variation.[30]

To keep test scores similar over time, we use the percentile ranking as our test score measure. This ranking reflects the percentage of students in a nationally-representative sample of students, in the same grade, tested at a comparable time of the school year, who fall below the test score for the mean student in a given school-grade-year. For example, if the average student in a school-grade-year scored at the 60th percentile on the standardized test, this would mean they did as well as (or better than) 60 percent of the students in the national sample. The shading in Figure 1 indicates the availability of these data alongside the CSR policy rollout (with summary statistics for the mathematics test score data in Table A.2).

Table 1 provides summary statistics for the main variables used in our analysis. We provide overall means and also break these down in the next three columns – into the period *preceding* the introduction of the CSR reform in California (1990-91 through 1995-96), the period *during* its phase-in across grades (1996-97 through 1999-00), and the period *following* its full implementation (2000-01 through 2008-09).

The evolution of the student-teacher ratio in elementary schools over time (shown in the first row) indicates that the CSR reform had a dramatic effect: the ratio fell from 25 to 21.6, reflecting a 15 percent decline in class size, although the actual class size decline in K-3 was likely much larger.[31]

[29]Students were exempted if they were special education students or if a parent or guardian submitted a written request for an exemption. Test taking rates were high nonetheless. For example, in 1998-99, over ninety-three percent of students in grades 2-11 took the relevant test.

[30]We restrict some of our analyses to the academic years 1997-98 through 2001-02, even though test scores are reported until 2012-13. This is because the monotonicity of scores by grade is no longer preserved for the 2002-03 academic year and onward due to a change in testing regimes for the 2002-03 school year (see Appendix A).

[31]The student-teacher ratio of the school is used as a proxy for class size as we do not observe teacher assignment data prior to the introduction of CSR. Given that elementary schools often include grades 4-6, we underestimate the decline in CSR grades (K-3) since non-CSR grades (4-6) are included in the calculation. Schrag (2006) indicates that the pre-CSR K-3 average class size was about 28.5. Given that post-CSR average K-3 class sizes are about 19.5, the actual class size decline caused by CSR in grades K-3 is likely closer to 30 percent.

The private school share of enrollment at the state level also declined during the period of interest, falling from 9.9 percent prior to CSR implementation to 8.8 percent afterward. Because there was a similar trend of declining private school shares nationally during the time period (Buddin, 2012), we adopt a grade-by-grade research design in the next section to assess whether CSR had a causal impact over and above the national trend. In addition, the table shows a marked change in the composition of students in public schools, with a reduction of about 10 percentage points in the share of white students and a corresponding increase in the fraction of Hispanic students. We will provide a formal analysis of such demographic changes alongside overall enrollment changes next.

4 Evidence of General Equilibrium Responses

In this section, we present causal evidence that sheds light on general equilibrium sorting responses to the reform. This evidence will motivate our approach for identifying the reform's direct and indirect effects in the following section.

We first investigate the impact of the reform on private school shares, drawing out the scale of the general equilibrium sorting response. Second, we provide complementary evidence relating to effects on public school demographics and private school entry and exit. Third, we examine whether the sorting induced by CSR was transitory or not. This is relevant when separating the reform's direct and indirect effects – a key aspect of our identification, which will exploit differences in the extent to which students return to the private system over time based on exogenous variation in public school grade spans.

4.1 Private Schools

We explore the effect of CSR on private school shares by taking advantage of the reform's grade-by-grade roll-out in grades K-3. For each period t, we define the treatment group as any grade that implements CSR and the control group as any grade that does not. Thus, we assume that all eligible grades adopted CSR according to the state's roll-out, abstracting from the voluntary participation decision by districts and schools – doing so is likely to understate the true effects of CSR (as noted above). We then apply a difference-in-differences approach, which compares treatment and control grades before and after the reform came into effect. The analysis uses the following regression (weighted by district-grade-year enrollment[32]):

$$share_{dgt} = \beta_0 + \beta_1 post_{gt} + \beta_2 treat_g + \beta_3 (post_{gt} * treat_g) + \eta_d + \theta_t + \phi X_{dgt} + \epsilon_{dgt}\,, \tag{4.1}$$

[32]Weighting is used to account for smaller districts that do not contain any private schools. Alternatively, the regression can be restricted to only those school districts with a private school option. We present results for the 'weighting' method, as the sample restriction produces similar estimates and so is omitted.

where $share_{dgt}$ is the private school share for district d in grade g at time t,[33] $post_{gt}$ indicates whether (or not) CSR had been implemented for grade g, $treat_g$ indicates whether grade g was ever subject to the CSR reform, X_{dgt} is a set of district-grade-year covariates (percent ESL, race and enrollment), and η_d and θ_t are district and time fixed effects, respectively.

The difference-in-differences coefficient of interest is β_3. It is identified under the assumption that CSR and non-CSR grades would have experienced the same change in private school share in the absence of the reform. While this 'parallel trends' assumption is not directly testable, evidence (below) indicating a lack of differential pre-trends is suggestive of its validity.

Results: To visualize our difference-in-differences approach, which exploits variation in the time when different grades became subject to CSR, Figure 3 plots the change in private school enrollment share overall (the dashed line) and in CSR grades (the lines in different shades) over time. The visual evidence is clear: when CSR is first implemented in the public system for a particular grade, the corresponding private share for that grade declines noticeably relative to other grades, suggesting that the reform attracted private school students into the public system. For example, the share of students in private schools in the entire state in first grade is flat in the two academic years preceding 1995-96. Then by the start of 1996-97 (the first year that CSR affects public school class sizes in first grade), there is a pronounced dip down in first grade while the shares for other grades remain steady, consistent with there being a switch into public schools for that grade.[34]

Given the patterns in Figure 3, we begin by reporting the most basic contrast in Table 2 – private school shares 'before' and 'after' CSR's introduction among treated and untreated grades. We find that treated grades experienced a precisely-estimated 1.1 percentage point decline in private school share relative to untreated grades following the implementation of the policy. This corresponds to the estimate from the difference-in-differences estimator in equation (4.1) without including any controls.

Next, we estimate the equation including various controls and report the results in Table 3 – the estimates in column (1), without controls, correspond to those in Table 2. Our preferred specification with all controls included is similar to its unconditional counterpart: treated grades experience a 1.4 percentage point decline in private school share relative to untreated grades as a result of CSR. This decline is equivalent to 12 percent of the pre-CSR K-3 average private school share of 11.7 percent – a significant amount – and 17 percent of its standard deviation.[35]

[33] Formally, $share_{dgt}$ is defined as the enrollment in private schools for the district-grade-year combination, *d-g-t*, divided by the total enrollment for *d-g-t*.

[34] Similarly, when second grade becomes eligible for CSR in public schools at the start of 1997-98, we also see a decline in the private school share relative to the previous academic year (and relative to other grades). The same is true for kindergarten and third grade in the first year when those grades became eligible (1998-99).

[35] Appendix B shows the robustness of the private school share results. Specifically, we conduct a triple-differences

We provide support for the 'parallel trends' assumption that underlies these results by plotting coefficient estimates by year, and find no evidence of differential pre-trends prior to the reform (see Figure A.2).

The steep decline in private school share caused by CSR makes it likely that the extensive margin – the number of schools – would also be affected, as in Dinerstein and Smith (2016). Figure 4 plots the number of private schools per 1000 school-aged children in California and the rest of the country.[36] As expected, there is a noticeable reduction in private schools per capita following the 1996-97 reform in California relative to the rest of the country.[37] We estimate a 0.06 decline in the number of private schools per 1000 school aged children, amounting to closing 360 private schools in California – a ten percent decline from the 3,467 private schools in the state prior to CSR's implementation.[38]

4.2 Public Student Composition

The previous set of results indicates that CSR caused students in relevant grades to switch from private schools into the public system. Our public school data allow us to explore the impact of this influx of new students from private schools on public school sociodemographic compositions at the school-grade-year level. To do so, we exploit variation in private school presence locally. Specifically, our econometric approach involves a triple-differences design, using the same grade and time differencing in equation (4.1) in addition to a third dimension of differencing related to whether a private school is nearby: our preferred specification defines 'nearby' as within 3 km. (A more detailed description of our approach is given in Appendix C.)

Given that the proportion of white students in private school is initially about fifteen percent higher and the proportion of Hispanic students about twenty three percent lower compared to their public counterparts (see Table A.5), inflows to the public system are likely to consist mainly of these two groups.[39] This is indeed what we find: the evidence in the table shows that CSR led to

specification that uses district-level CSR participation intensity as an additional dimension of differencing and find qualitatively similar results.

[36] To make this comparison, we use data from the Private School Universe Survey, conducted by the National Center for Education Statistics. It is available at `https://nces.ed.gov/surveys/pss/pssdata.asp`.

[37] These can be further broken into private school entry and exit responses. After the 1996-97 CSR reform, Figures A.4(a) and A.4(b) show a sharp increase in private school exit rates and a decline in entry rates in California relative to the rest of the country.

[38] See Table A.4 for estimates of the extensive margin effects of CSR in difference-in-differences and triple-differences frameworks (Appendix B provides a fuller description). As the three layers of differencing, the latter uses time (pre- versus post-CSR), state (California versus the rest of the country), and whether (or not) the private school served CSR grades.

[39] While we do not have detailed private school demographic data, the NCES provides school-level demographics for the 1997-98 school year and every two years thereafter. Based on this data source, the public-private demographic disparities we report are thus one year after CSR began in 1996-97.

a 2.9 percentage point increase in the fraction of white students and a decline of 1.5 percentage point in the fraction of Hispanic students in public schools with nearby private alternatives (relative to public schools without nearby private competitors), indicative of pronounced sociodemographic sorting.

4.3 Sorting: Transitory or Permanent?

The causal evidence relating to the initial impact of the reform prompts the question whether the sorting we have documented is transitory or not. This will be relevant when separating the reform's direct and indirect effects, as we show in the next section. There are three possibilities: students previously in the private school system might return to private schools directly upon completion of third grade when the CSR treatment ends; they might return after completing all grades offered by the public school they switched into (say, after fifth grade in a K-5 school); or they might remain in the public system for the duration of their primary and secondary education.

Our data do not provide measures of individual switching behaviour directly, but we are able to shed light on this issue using private school share data aggregated to the district level. Specifically, we exploit the differential exposure of cohorts to the reform, drawing on the idea that *if* the second possibility above holds, then pronounced changes in private school share should line up with elementary school grade spans. To that end, we implement the following regression discontinuity design:

$$grade\text{`}i\text{'}share_{dc} = \beta_0 + \beta_1 D_{dc} + \beta_2 f(cohort_{dc}) + \beta_3 D_{dc} * f(cohort_{dc}) + \eta_d + \epsilon_{dc}\,, \tag{4.2}$$

for $-b \leq cohort_{dc} \leq b$, where $grade\text{`}i\text{'}share_{dc}$ is the private school share in grade i belonging to cohort c in district d, indicator D_{dc} denotes whether cohort c was exposed to CSR, $f(\cdot)$ is a flexible polynomial function, $cohort_{dc}$ is the cohort number (defined by the year that the student enters kindergarten and normalized by that year's relation to the year the reform was introduced),[40] η_d is a district fixed effect, and b is some bandwidth.

This regression discontinuity design identifies CSR's impact on the private school share for each grade. Our coefficient of interest is β_1, which represents the effect of CSR on the private school share of cohorts in grade i. Given that by far the most common grade configurations in California are K-5 and K-6,[41] the second possibility rehearsed above would imply an increase in β_1 from elementary school non-CSR grades (4-6) to the middle school grades (7-8), while the first

[40] The cohort entering kindergarten in 1995-96 is designated 'cohort zero' as it is the first cohort to be exposed to CSR in first grade. Since the cohort variable is discrete, we add 0.5 to each value so that zero is the midpoint between the first treated and untreated cohorts.

[41] Table A.11 reports the numbers and percentages of elementary schools by grade configuration in California. It shows that elementary schools are divided approximately equally between K-5 and K-6 configurations.

possibility would imply no such increase.

In terms of the persistence of sorting, Figure A.5 plots the estimated effect of CSR on private school share for each grade, and in Table 4, we report average effects grouped by elementary, middle, and high school to increase power. The effect for kindergarten should be considered a placebo, as the first CSR cohort was exposed in first grade only; this is borne out by an estimate statistically indistinguishable from zero. Estimates for subsequent grade spans indicate that the CSR reform induced private school students to enter the public school system and that they remained there until completion of the elementary grades. Approximately two-thirds of the CSR 'treatment effect' on private school share disappears when making the transition to middle school, consistent with a significant share of students transitioning back into the private system at that time. Thus the estimates indicate that a sizeable fraction (around two thirds[42]) of the sixth-grade students attending K-5 public schools switched back to the private system for middle school one year earlier than those attending K-6 public schools.

5 Estimation Approach

The reduced-form evidence of significant general equilibrium sorting responses to CSR motivates the multiple differencing approach at the heart of our analysis. This exploits independent variation across cohorts and school configurations to determine the direct and indirect effects of the reform, while allowing each effect to persist differentially. Before describing our estimation approach in detail and highlighting its generality, we derive the main estimating equation we will take to the data, based on the conceptual framework in Section 2.

5.1 Estimation Framework

Our starting point is the linear cumulative production technology with geometric decay given in equation (2.1). We adapt this to reflect the available aggregated data, writing the school-grade-year (s-g-t) test score y_{sgt} as a function of current and past inputs:

$$y_{sgt} = \gamma_R \sum_{\tau=0}^{L} (\delta_R)^\tau R_{s,g-\tau,t-\tau} + \gamma_X \sum_{\tau=0}^{L} (\delta_X)^\tau X_{s,g-\tau,t-\tau} + \epsilon_{sgt}\,. \tag{5.1}$$

To keep track of grades and time, we use the τ index to increment both successive grades ($g \in \{0, 1, \ldots, 6\}$) and academic years ($t \in \{1996\text{-}97, 1997\text{-}98, \ldots\}$). All unobserved determinants of the test score are represented by ϵ_{sgt}. (Teacher quality Q is suppressed in (5.1) for expositional clarity:

[42] This fraction is implied by the effect size falling from -0.30 in elementary non-CSR grades to -0.10 in middle school grades.

we return to this later in the section.)

Given our interest in modeling the response of test scores to the introduction of a major education reform like CSR, we draw notional contrasts between observed scores at the school-grade-year level and 'counterfactual' scores that would have prevailed had the reform not been enacted. The rationale for differencing in this way is to embed the counterfactual control into our estimating equation; thus, the resulting estimates are relative to a baseline in which the reform never came into effect.

The comparisons we make involve school averages, where the total number of schools is given by N_s. Averaging over all schools serving grade g in year t, define $\Delta y_{gt} \equiv y_{gt} - y^u_{gt} \equiv \frac{1}{N_s}\sum_s (y_{sgt} - y^u_{sgt})$ as the difference between the actual average test score for that grade-year combination and the unobserved (superscripted by 'u') average score that would arise in a counterfactual setting in which the reform had never been implemented. Analogously, we define ΔR_{gt} and ΔX_{gt} based on average differences between actual and counterfactual school resources and sociodemographics respectively.[43]

In practice, we cannot construct Δy_{gt} directly, given that y^u_{sgt} is unobserved. Instead, we obtain a prediction of the counterfactual score (denoted by $\hat{y}^u_{gt}$) based on data for untreated cohorts under assumptions stated below. Forming the predicted difference $\widehat{\Delta y}_{gt} \equiv y_{gt} - \hat{y}^u_{gt}$, the estimating equation is given by:

$$\widehat{\Delta y}_{gt} = \gamma_R \sum_{\tau=0}^{L} (\delta_R)^\tau \Delta R_{g-\tau,t-\tau} + \gamma_X \sum_{\tau=0}^{L} (\delta_X)^\tau \Delta X_{g-\tau,t-\tau} + \Delta\epsilon_{gt}\,, \tag{5.2}$$

where $\Delta\epsilon_{gt} \equiv y^u_{gt} - \hat{y}^u_{gt}$,[44] and $\Delta R_{g-\tau,t-\tau}$ and $\Delta X_{g-\tau,t-\tau}$ represent the change in school resources and the mix of students arising from CSR (relative to the counterfactual baseline) for students in grade $g-\tau$ and academic year $t-\tau$.

For *treated* grade-years, represented in Figure 1 with a '$\times$' symbol, resources increase ($\Delta R_{gt} \neq 0$) and sociodemographics adjust ($\Delta X_{gt} \neq 0$), as the descriptive evidence in Section 4 shows. Thus we would expect to see non-zero test score effects relative to the 'no-CSR' counterfactual baseline (yielding $\Delta y_{gt} \neq 0$) for treated cohorts, according to the sequence of relevant educational inputs (which we term the 'input trajectory') received by students as they progress through the education system. For all control combinations (such as third grade and above in 1997-98), represented in Figure 1 with a '$\cdot$' symbol, we have $\Delta R_{gt} = \Delta X_{gt} = 0$, which implies that $\Delta y_{gt} = 0$ (from equation (5.2)).[45]

[43] So, for example, $\Delta R_{gt} \equiv R_{gt} - R^u_{gt} \equiv \frac{1}{N_s}\sum_s (R_{sgt} - R^u_{sgt})$.

[44] To see this, note that $\widehat{\Delta y}_{gt} = \Delta y_{gt} + \Delta\epsilon_{gt}$. Thus, $\Delta\epsilon_{gt} = \widehat{\Delta y}_{gt} - \Delta y_{gt} = (y_{gt} - \hat{y}^u_{gt}) - (y_{gt} - y^u_{gt}) = y^u_{gt} - \hat{y}^u_{gt}$. Intuitively, $\Delta\epsilon_{gt}$ shrinks as the prediction improves.

[45] For control grade-year observations, it must be the case that $\hat{y}^u_{gt} \equiv y^u_{gt}$, so that $\widehat{\Delta y}_{gt} \equiv \Delta y_{gt}$.

5.2 Identification

Applying multiple differencing, we show how the direct effect, the indirect effect and the persistence parameters are identified, before extending our strategy to account for teacher quality.

Identifying the Direct Effect: To estimate the direct effect, we carry out a within-year comparison of two adjacent cohorts that received differential exposure to the CSR reform, given not all grades are treated by CSR.

Consider the input trajectories of students in third and fourth grades in 2001-02. These are displayed in Figure 5, highlighted by two diagonal outlines that each enclose four points. It is clear that third graders in that school year had received four successive years of direct exposure to smaller classes, while the fourth graders received only three.

We will argue that differencing the two trajectories using our estimation framework will isolate the direct effect of interest. The differencing argument draws on two assumptions. First, under the assumption that the indirect effects of the policy are the same for the two cohorts in that year, applying the same cross-cohort differencing will 'difference out' all the indirect effects. Second, a 'common trends' assumption, which is plausible in this context, allows untreated grade-year observations to provide valid counterfactuals, and implies that the counterfactual prediction error will be zero, which yields the direct effect alone (as we show).

At the outset, to specify the *indirect* input trajectory, we appeal to the following fact:

Fact 1: *Nearly all private school students drawn into the public system by CSR remain there until they transition to middle school, at which point approximately two-thirds return to the private system.*

Evidence supporting this fact comes from the regression discontinuity results in Table 4 showing the estimated effect of CSR on private school shares by grade span.[46] The import of this fact is that the indirect general equilibrium effects of the reform should influence elementary school grades whether or not they are subject to CSR. Students do not return to the private system immediately after third grade (presumably because of the switching costs involved), so that fourth grade classrooms (for instance) will also be affected *indirectly* through induced changes in student composition, even though fourth grade students were never subject to CSR directly.

The estimation framework allows us to make the identification argument formally. The argument has the following structure: First, we use estimating equation (5.2) to obtain an expression for the average predicted test score difference for the third grade cohort in 2001-02, $\widehat{\Delta y}_{3,01\text{-}02}$, and

[46] The coefficient is negative and significant for CSR grades (first through third) and is unchanged for non-CSR elementary grades (fourth through sixth). The magnitude of the point estimate then drops by two-thirds for middle school.

similarly $\Delta y_{4,01\text{-}02}$ for the fourth grade cohort in the same year. Second, we form the difference, $\Delta y_{3,01\text{-}02} - \widehat{\Delta y}_{4,01\text{-}02}$, deducting the latter from the former. (The relevant expressions are set out in full in Appendix D.) Third, under a plausible assumption regarding input trajectories – Assumption 3 below – this difference simplifies to the following expression:

$$\widehat{\Delta y}_{3,01\text{-}02} - \widehat{\Delta y}_{4,01\text{-}02} = \gamma_R \Delta R_{3,01-02} + (\Delta \epsilon_{3,01\text{-}02} - \Delta \epsilon_{4,01\text{-}02}). \tag{5.3}$$

Then fourth, under a parallel trends assumption (Assumption 4 below), the differenced error term in (5.3) will equal zero. This implies that the RHS of the expression will consist solely of the direct effect of interest, $\gamma_R \Delta R_{3,01-02}$. Further, the same parallel trends assumption will allow us to express the LHS in terms of known quantities – differences in grade-year average test scores – completing the identification argument.

We now set out the two assumptions formally, and justify each one. First is an assumption (in two parts) about input levels experienced by cohorts that were treated (at least in some years) under CSR:

Assumption 3(a): $\Delta R_{gt} = \Delta R_{g't} \quad \forall g, g'$.

In words, all grades treated by CSR in a given year t experience the same class size 'treatment.' Supporting evidence in Table A.6 shows that CSR grades had similar class sizes once the reform was implemented. In addition, Bohrnstedt and Stecher (2002) report that, once fully implemented in 2000-01, 95-98 percent of students in each CSR grade were in CSR-compliant classrooms, indicating little heterogeneity in grade-level implementation rates.

Assumption 3(b): $\Delta X_{gt} = \Delta X_{g't} \quad \forall g, g'$.

Part (b) says the indirect effects of CSR in a given year t are the same across grades. This is supported by the regression discontinuity evidence presented in Table 4.

Assumption 3 (combined) implies – as suggested above – that differences in class size inputs across treated grades in the same year are all zero and so drop out, and that differences in sociodemographic inputs across grades in the same year (comparing the relevant cohorts) are also zero and can be ignored. The only remaining school input affecting score differences will be the change in class size in third grade in 2001-02.

The expression in (5.3) makes clear that we still have to attend to the error difference term, $(\Delta \epsilon_{3,01\text{-}02} - \Delta \epsilon_{4,01\text{-}02})$, on the RHS. This captures any error introduced by the prediction of the counterfactual: identification of the direct effect thus hinges on the quality of that prediction. The following parallel trends assumption is useful in that regard.

Assumption 4: In the absence of the reform, test score differences between grades g

and g' are time-invariant: $y^u_{gt} - y^u_{g't} = y^u_{gt'} - y^u_{g't'}$.

The assumption implies that no other contemporaneous reforms affect grades differentially. Support for this assumption in our setting comes from the fact that grade differences for untreated cohorts are statistically indistinguishable from each other over time (see Table A.8).

Assumption 4 allows observations for untreated cohorts to serve as controls for their never-treated counterparts. A natural candidate is the within-year difference between third and fourth grades in 1997-98 (highlighted in the figure by two encircled points), as neither of these cohorts was ever affected by the reform. In particular, given that $\Delta y_{gt} \equiv y_{gt} - y^u_{gt} = 0$ for never-treated 'control' grade-year combinations, Assumption 4 implies that $y_{3,97\text{-}98} - y_{4,97\text{-}98} = y^u_{3,97\text{-}98} - y^u_{4,97\text{-}98} = y^u_{3,01\text{-}02} - y^u_{4,01\text{-}02}$.

Two relevant consequences follow from this assumption. First, the difference $\widehat{\Delta y}_{3,01\text{-}02} - \widehat{\Delta y}_{4,01\text{-}02}$ can be expressed in terms of known quantities only – specifically, $(y_{3,01\text{-}02} - y_{4,01\text{-}02}) - (y_{3,97\text{-}98} - y_{4,97\text{-}98})$.[47] The justification is that, under Assumption 4, $\widehat{\Delta y}_{3,01\text{-}02} - \widehat{\Delta y}_{4,01\text{-}02} = \Delta y_{3,01\text{-}02} - \Delta y_{4,01\text{-}02}$ – that is, the counterfactual score difference will be equal to the *true* score difference. This follows from the fact that the difference in known test scores $(y_{3,01\text{-}02} - y_{4,01\text{-}02}) - (y_{3,97\text{-}98} - y_{4,97\text{-}98})$ is equal to $(y_{3,01\text{-}02} - y_{4,01\text{-}02}) - (y^u_{3,01\text{-}02} - y^u_{4,01\text{-}02})$ under the assumption, and that expression (appealing to the definitions) is equal to the true counterfactual score difference, $\Delta y_{3,01\text{-}02} - \Delta y_{4,01\text{-}02}$.

The second consequence is that the test score difference is itself equal to the direct effect of the reform – the quantity of interest. This will be the case if the error term difference $(\Delta\epsilon_{3,01\text{-}02} - \Delta\epsilon_{4,01\text{-}02})$ in (5.3) is equal to zero, as it is under Assumption 4.[48]

To summarize, identification of the direct effect ($\gamma_R \Delta R$) requires average test scores of two different cohorts to be compared within the same treated year, controlling for any non-CSR differences by deducting off corresponding average scores for older cohorts not subject to the reform. Based on the timing of CSR's implementation, this sequence of treatments occurs in our context for third and fourth grade in the 2001-02 school year, with the same grades in 1997-98 accounting for the non-CSR counterfactual.[49]

Identifying the Indirect Effect: Using a similar differencing strategy, the framework can in turn be used to recover the indirect effect ($\gamma_X \Delta X$), this time differencing based on school grade span configurations. In particular, we compare sixth grade test scores in areas with K-6 versus K-5

[47] The reasoning is that, definitionally, $\widehat{\Delta y}_{3,01\text{-}02} - \widehat{\Delta y}_{4,01\text{-}02} = (y_{3,01\text{-}02} - y_{4,01\text{-}02}) - (\hat{y}^u_{3,01\text{-}02} - \hat{y}^u_{4,01\text{-}02})$, where the term on the right of the = sign can be written $(y_{3,01\text{-}02} - y_{4,01\text{-}02}) - (y_{3,97\text{-}98} - y_{4,97\text{-}98})$ – in terms of known quantities (under the assumption).

[48] To see why, use actual scores in 1997-98 to predict counterfactual scores in 2001-02 and apply the assumption. Then we have: $\Delta\epsilon_{3,01\text{-}02} - \Delta\epsilon_{4,01\text{-}02} = (y^u_{3,01\text{-}02} - \hat{y}^u_{3,01\text{-}02}) - (y^u_{4,01\text{-}02} - \hat{y}^u_{4,01\text{-}02}) = (y^u_{3,01\text{-}02} - y^u_{4,01\text{-}02}) - (y^u_{3,97\text{-}98} - y^u_{4,97\text{-}98}) = 0$.

[49] Second grade cannot be used, as no pre-reform observations are available to construct the relevant counterfactual.

configurations. In doing so, we will control for two additional layers of differencing: first, contrasting the sixth grade difference between configurations with the fifth grade difference in a school year when CSR is affecting sixth grade cohorts (2001-02, for instance), and second, comparing the resulting difference-in-difference in 2001-02 to its untreated analog in a pre-treatment year (e.g., 1997-98).

Our focus on K-6 and K-5 schools is motivated by Fact 1 (already discussed) and also a second fact.

Fact 2: *Schools with K-5 and K-6 grade spans account for the majority – 47 and 42 percent, respectively – of schools serving elementary grades in California.*

We show this in Appendix Table A.11. Given Fact 1 – that around two-thirds of K-5 students initially drawn into the public system by CSR returned to the private sector for sixth grade while their K-6 counterparts did not – Fact 2 gives rise to exogenous grade span variation that generates differential spillovers from CSR.

We can use the estimation framework to write an expression for the average predicted achievement difference for sixth grade K-6 students in 2001-02 (relative to the no-CSR counterfactual):

$$\begin{aligned}\widehat{\Delta y}_{6,01-02,K6} = {} & (\delta_R)^5\gamma_R\Delta R_{1,96-97} + (\delta_R)^4\gamma_R\Delta R_{2,97-98} + (\delta_R)^3\gamma_R\Delta R_{3,98-99} \\ & + (\delta_X)^5\gamma_X\Delta X_{1,96-97} + (\delta_X)^4\gamma_X\Delta X_{2,97-98} + (\delta_X)^3\gamma_X\Delta X_{3,98-99} \\ & + (\delta_X)^2\gamma_X\Delta X_{4,99-00} + \delta_X\gamma_X\Delta X_{5,00-01} + \gamma_X\Delta X_{6,01-02,K6} + \Delta\epsilon_{6,01-02,K6} \quad (5.4)\end{aligned}$$

where the extra subscript 'K6' represents students in schools with a K-6 configuration; similarly, we will use the 'K5' subscript to denote students in areas where the elementary school grade span is K-5.[50] The RHS of equation (5.4) reflects the input trajectory, involving both resources and sociodemographics, that sixth graders in 2001-02 have been exposed to while in their K-6 schools.

This trajectory is illustrated in the schematic Figure 6(a), making clear the CSR 'resource shock' only applied for three of those six years, while sociodemographic spillovers from CSR applied in *each* of the six years. The average predicted achievement difference for sixth grade K-5 configuration students in 2001-02 can be written analogously, with Figure 6(b) representing the associated trajectory.

We make the following assumption about inputs for K-6 and K-5 schools:

Assumption 5: (a) $\Delta R_{g,t,K6} = \Delta R_{g,t,K5} \quad \forall\ g \leq 5$**, and (b)** $\Delta X_{g,t,K6} = \Delta X_{g,t,K5} \quad \forall\ g \leq 5$**.**

This is a refinement to Assumption 3 above: the increased resources and student composition changes due to CSR are assumed to be identical across grade configurations for all common grades

[50] Here, given our focus on variation in grade configurations, averages are taken over all schools in a particular grade configuration.

up to and including fifth grade. Assumption 5(a) holds since the reform was applied uniformly across configurations (see Table A.11). Assumption 5(b) is supported by Fact 1 as well as the lack of significant differences in demographic sorting across K-5 and K-6 schools in a triple-differences design, described in Appendix C.

Taking the difference between sixth grade students in K-6 versus K-5 configurations yields:

$$\begin{aligned}\widehat{\Delta y}_{6,01-02,K6} - \widehat{\Delta y}_{6,01-02,K5} &= \gamma_X \Delta X_{6,01-02,K6} - \gamma_X \Delta X_{6,01-02,K5} + (\Delta\epsilon_{6,01-02,K6} - \Delta\epsilon_{6,01-02,K5}) \\ &= \psi\gamma_X \Delta X_{6,01-02,K6} + (\Delta\epsilon_{6,01-02,K6} - \Delta\epsilon_{6,01-02,K5}), \end{aligned} \tag{5.5}$$

where the parameter $\psi \leq 1$ gives the proportion of students induced into the public system by CSR who then *exit* the public system during the transition to middle school. Specifically, we let $\Delta X_{6,t,K5} = (1-\psi)\Delta X_{6,t,K6}$, and estimate ψ to be equal to two-thirds (Fact 1).

Under the parallel trends assumption (Assumption 4), 1997-98 scores could serve as valid counterfactuals for the scores of K-5 and K-6 schools in 2001-02 in the absence of CSR (implying $\Delta\epsilon_{6,01-02,K6} - \Delta\epsilon_{6,01-02,K5} = 0$). The indirect effect of interest ($\gamma_X \Delta X_{6,01-02,K6}$) could in turn be recovered from the known left-hand side of equation (5.5), given by

$$\widehat{\Delta y}_{6,01-02,K6} - \widehat{\Delta y}_{6,01-02,K5} = (y_{6,\text{01-02},K6} - y_{6,\text{01-02},K5}) - (y_{6,\text{97-98},K6} - y_{6,\text{97-98},K5}).$$

We relax that assumption here, requiring that parallel trends hold using difference-in-differences (rather than first differences), using *two* layers of differencing to provide the counterfactual. As the first layer, we account for systematic differences between K-5 and K-6 schools by differencing out fifth grade test scores in K-5 ($y_{5,01-02,K5}$) and K-6 schools ($y_{5,01-02,K6}$). As the second, which requires a weaker assumption than the one layer of differencing under Assumption 4, we use the pre-reform test scores for both fifth and sixth grades in K-5 and K-6 schools as counterfactuals for the observed test scores in fifth and sixth grades in the 2001-02 school year.[51] Therefore, we have:

$$\begin{aligned}\psi\gamma_X \Delta X_{6,01-02} =& [y_{6,01-02,K6} - y_{5,01-02,K6} - (y_{6,97-98,K6} - y_{5,97-98,K6})] \\ &- [y_{6,01-02,K5} - y_{5,01-02,K5} - (y_{6,97-98,K5} - y_{5,97-98,K5})], \end{aligned} \tag{5.6}$$

which allows us to identify the indirect effect ($\gamma_X \Delta X$) based on observed scores on the RHS and our estimate of ψ.[52] Effectively, the identification of the indirect effect in our framework involves comparing sixth grade versus fifth grade in K6 schools versus K5 schools for cohorts affected by

[51] Here, we are over-identified. We could use 1997-98, 1998-99 and 1999-00 as counterfactuals since cohorts in fifth and sixth grades were not subject to CSR in those years. In practice, we use all three and take an average of the estimates, although estimates are quantitatively similar regardless which counterfactual year we use.

[52] While the identification of the indirect effect ($\gamma_X \Delta X$) requires comparing cohorts in K-6 and K-5 schools for 2001-02 or later, a change to test scores for the 2002-03 school year prevents us from using subsequent cohorts in practice.

CSR versus those that were not: this type of comparison is analogous to running a triple-differences regression, which will provide the basis for a useful robustness check described in the next section.

Identifying the Persistence Parameters: We are able to isolate the parameters governing the persistence of the reform (δ_R) and the persistence of changes in student demographics (δ_X) using a similar approach.[53] To do so, we take the estimated contemporaneous effects $\gamma_R \Delta R$ and $\gamma_X \Delta X$ as given and construct the following two differences: (i) between fourth- and third-grade test scores in the 2000-01 school year, and (ii) between fourth- and fifth-grade test scores in the 2000-01 school year. This forms a system of two non-linear equations (equations (D.6) and (D.7) in Appendix D.1) with two unknowns (δ_R and δ_X), which we then solve for, computing bootstrapped standard errors for each.

Intuitively, we identify the persistence parameters by exploiting variation across third, fourth, and fifth grade cohorts in 2000-01. All three cohorts were affected through the direct class size reduction channel for three years, but were affected by the indirect channel for different lengths of time (three, four and five years for third, fourth and fifth grade, respectively). This allows us to separate the persistence of the indirect effect (δ_X), which affected some grades more than others, from the persistence of the direct effect (δ_R), which influenced all grades equally, although it was applied at different points in time.

Accounting for Teacher Quality: Our identification strategy can be extended to include indirect teacher quality effects in the equations used to identify the direct effect, indirect student sorting effect and persistence parameters. We allow teacher quality to differ according to year, whether the grade was subject to CSR or not, and how removed the treatment is from the grade-year combination of interest (i.e., the lag).

In essence, we follow Jepsen and Rivkin (2009) by appealing to variation in observable teacher experience as a proxy for teacher quality.[54] Those authors document a pronounced increase in the overall proportion of inexperienced teachers following the introduction of CSR, and a subsequent decline to pre-CSR levels after a few years. Given that our framework relies on variation across CSR and non-CSR grades over time, we draw on evidence relating to the way in which teacher inexperience evolved by grade, presented in Appendix Table A.12.[55] Such changes are observable each year, allowing our treatment of these indirect teacher quality effects to be non-parametric,

[53] We are restricted to identifying the persistence parameters only, rather than the non-parametric effect of the reform in each period, due to the change in the test format for the 2002-03 school year.

[54] This decision is necessary given that the number of parameters exceeds what can be identified through variation in test scores alone: there is a parameter for every '$\times$' or '$\cdot$' contained within a cohort's trajectory, across all cohorts analyzed.

[55] Jepsen and Rivkin (2009) control implicitly for teacher observables that evolve by grade, using school-grade-year controls and grade-year fixed effects, and so do not document patterns at the grade-year level.

rather than following the 'geometric decay' treatment used for class size and student sorting. (We set out the full reasoning in Appendix D.3.)

5.3 Difference-in-Differences Design

Following on from the discussion of identification, it is instructive to show why estimating the impact of a large-scale reform using a difference-in-differences ('D-in-D') specification will not, in general, be appropriate for estimating the direct effect. Doing so would entail comparing the pre-/post- reform difference in average scores of students in a grade who became subject to the policy with the corresponding scores of students in a control grade (as discussed). We show that even if the linear technology and parallel trends assumptions (Assumptions 1 and 4 above) hold, such a strategy will produce biased estimates as long as at least one of two statements is true: (i) the effect of past resources persists ($\delta_R \neq 0$), or (ii) there are general equilibrium spillover effects ($\gamma_X \Delta X \neq 0$).

To see this, consider students in third and fourth grade in the 1998-99 school year. As already rehearsed, the third grade cohort in 1998-99 is affected by the CSR reform in first, second and third grade, both directly and indirectly (due to the general equilibrium spillover), while the fourth grade cohort in 1998-99 is never affected. Based on the definition of Δy_{gt} and Assumption 4, the D-in-D specification comparing third (CSR) and fourth (non-CSR) grades from 1997-98 to 1998-99 is:

$$\begin{aligned} \Delta y_{3,98\text{-}99} - \Delta y_{4,98\text{-}99} &= y_{3,98\text{-}99} - y_{4,98\text{-}99} - (y^u_{3,98\text{-}99} - y^u_{4,98\text{-}99}) \\ &= y_{3,98\text{-}99} - y_{4,98\text{-}99} - (y_{3,97\text{-}98} - y_{4,97\text{-}98}). \end{aligned} \tag{5.7}$$

Taking the terms on the LHS, $\Delta y_{3,98\text{-}99}$ is given by:

$$\begin{aligned} \Delta y_{3,98\text{-}99} = (\delta_R)^2 \gamma_R \Delta R_{1,96\text{-}97} + \delta_R \gamma_R \Delta R_{2,97\text{-}98} + \gamma_R \Delta R_{3,98\text{-}99} \\ + (\delta_X)^2 \gamma_X \Delta X_{1,96\text{-}97} + \delta_X \gamma_X \Delta X_{2,97\text{-}98} + \gamma_X \Delta X_{3,98\text{-}99} + \Delta \epsilon_{3,98\text{-}99}\,, \end{aligned} \tag{5.8}$$

and from equation (5.2), we have $\Delta y_{4,98\text{-}99} = 0$. The direct effect of the reform in this instance is $\gamma_R \Delta R_{3,98\text{-}99}$. Yet it is clear that the RHS of (5.7), $y_{3,98\text{-}99} - y_{4,98\text{-}99} - (y_{3,97\text{-}98} - y_{4,97\text{-}98}) \neq \gamma_R \Delta R_{3,98\text{-}99}$, the RHS of (5.8), unless $\delta_R = 0$ *and* there are no general equilibrium spillover effects ($\gamma_X \Delta X_{gt} = 0$) in any of the three years. That is, the direct effect is not identified if there is persistence in the student learning technology or there are spillovers.[56]

This issue applies more broadly, to other data structures. If the reform applied to all eligible grades directly upon its introduction, for example, researchers would still need an approach for

[56] This conclusion holds when making other post-reform grade comparisons (e.g., $\Delta y_{3,99\text{-}00} - \Delta y_{4,99\text{-}00}$).

estimating the indirect effect and controlling for that when estimating the direct effect. Otherwise, the latter would continue to be biased using a D-in-D approach.

5.4 Generality of the Estimation Approach

We have used the multiple differencing approach in this application to identify the direct and indirect effects of a large-scale reform in the presence of data restrictions (not least, a lack of pre-reform data) that make the empirical analysis of California's CSR program challenging. The approach we propose is more widely applicable – to settings in which a policy reform treats a subset of grades shared by two or more grade configurations. Combined variation of these two forms is widespread: schools are often subject to a major reform that extends across co-existing grade configurations; and our approach applies as long as only a subset of grades is treated by the reform.[57]

Recall that our approach for uncovering indirect sorting effects exploits the differential sorting behavior of students attending schools with different grade spans. This provides variation in student exposure to indirect inputs (school demographics) while leaving their exposure to direct inputs (school resources) unchanged; the latter can then be differenced away. The indirect sorting effect can thus be identified when such grade-span variation is available, allowing the relevant triple difference to be constructed.

In terms of estimating the direct effect, if the implementation of the reform is staggered by grade (as in the case of CSR), then our differencing approach for identifying the direct effect is required. If, instead, the rollout of the reform is not staggered, but rather affects treated grades simultaneously, an alternative would be to use a standard difference-in-differences approach, although we emphasize that this would need to be combined with the methodology we propose for estimating (and so controling for) the indirect effect – as is clear from the discussion of biases using a regular D-in-D approach, the estimated effect combines the impacts of resources and sociodemographics. The fact that a regular difference-in-differences research design alone is unable to identify the direct effect, irrespective of whether the introduction of the reform is staggered or not, underscores the need for the multiple differencing estimation approach we propose.

From an implementation standpoint, the multiple differencing approach is appealing in that it can be used in observational settings without the use of individual data, making it viable in contexts where researchers face common data limitations. Of note, all the data used in our analysis are available publicly, and are averaged to the school-grade-year level. In other settings where researchers

[57] For example, kindergarten through fifth grades are shared by K-5 and K-6 configurations, and our approach simply requires that at least one of those grades is untreated (in the K-5 configuration).

do not face such data restrictions, applying our approach will generate over-identified estimating equations, in turn allowing researchers to conduct diagnostics of the assumptions underlying our methodology.

6 Estimates

This section presents results from implementing the multiple differencing approach set out in Section 5. We also discuss the magnitudes from a policy perspective, decompose the indirect effect, and consider the broader relevance of the approach in practice.

6.1 Main Results

Table 5 provides the main estimates – for the parameters governing the contemporaneous partial equilibrium effect of CSR (γ_R), the general equilibrium effect of CSR on school composition (γ_X), and the persistence of resources associated with CSR (δ_R) and school sociodemographics (δ_X), respectively.

The table's layout, organized in four columns in pairs of two, reflects a bounding exercise that relates to the sorting parameter, γ_X. The first and second pair of columns are calculated using two different assumptions about the proportion of students, ψ, who return to private school after completing all grades offered by the public school they switched to initially. In columns (1) and (2), we follow the regression discontinuity evidence in Table 4 and treat $\psi = \frac{2}{3}$, using the fact that two-thirds of the students are estimated to return to the private system when the middle school transition occurs; these contain our preferred estimates. A lower bound estimate for γ_X is then provided in columns (3) and (4) by assuming that *all* students who were drawn into the public system by CSR return to the private system during the middle school transition ($\psi = 1$); if fewer students return when the transition occurs, then our estimate gets scaled up.[58]

Relative to columns (1) and (3), columns (2) and (4) add county fixed effects. In keeping with the evidence in Jepsen and Rivkin (2009), we find that controlling for teacher quality is important, and so only report estimates that include teacher quality controls. (The teacher quality estimates themselves are given in Table A.13.) All the estimates in columns (2) and (4) are significant, and somewhat more precise than those without county fixed effects, in columns (1) and (3) – the standard errors for the γ parameters are recovered using the delta method, while we bootstrap the standard errors for the δ (persistence) parameters.

[58] This is apparent from the LHS of equation (5.6) above, $\psi\gamma_X\Delta X_{6,01-02}$, where a higher value of ψ implies a lower value of γ_X for a given change in sociodemographics.

Focusing on our preferred estimates in column (2), based on the estimated share $\psi = \frac{2}{3}$ and including county fixed effects, the direct impact of CSR accounts for a 2.2 unit increase in the mean percentile rank of students, which corresponds to a 0.11σ increase in the school-grade test score distribution. The magnitude of this estimate is in line with experimental estimates: for instance, Krueger and Whitmore (2001) find that Project STAR raised test scores by around 0.1 standard deviations.

By way of contrast, consider results from a difference-in-differences specification,[59] shown in Table A.9 (with Figure A.6 indicating that pre-trends hold for test scores). These imply treatment effects in the region of 0.07σ for mathematics scores, the difference-in-differences understating our preferred estimates of the class size effect that account for persistence by over a third.[60]

Alongside the direct effect, the general equilibrium sorting effect accounts for a 3.3 unit increase in the mean percentile rank of students, which is equivalent to a 0.16σ increase in the school-grade test score distribution. It is precisely estimated, and it is larger in magnitude than the direct effect.[61] This is the first estimate in the empirical education literature of general equilibrium sorting effects placed on the same footing (in terms of test scores) as the direct effects of major reforms.

We noted in the previous section that identification of the indirect effect in our framework was analogous to running a triple-differences regression. In Appendix C, we implement the triple-differences analog to our differencing method, and find it yields similar estimates of the indirect effect. Framing our indirect effect in this manner provides a useful check on our identification strategy: differences between K6 schools and K5 schools among cohorts affected and unaffected by CSR should only appear in the sixth grade versus fifth grade comparison as a result of re-sorting to the private system and not the other grade comparisons we can also make. Figure A.7 shows that this is indeed the case: the triple-differences estimates between all other grade g and $g-1$ comparisons are statistically indistinguishable from zero.

Turning to the persistence parameters, δ_R and δ_X, we find that in our preferred specification, both the direct and indirect effects fade out in the range of 45-57 percent each year. These estimates accord with much of the literature on fade-out, which finds that the class size test score gain is

[59] Specifically, we run the following event study regression: $y_{sgt} = \beta_0 + \beta_1 post_{sgt} + \beta_2 treat_{sg} + \beta_3 post_{sgt} * treat_{sg} + \eta_s + \theta_t + \delta_g + \phi X_{sgt} + \epsilon_{sgt}$, where y_{sgt} is the average test score in school s in grade g at time t, $post_{sgt} \equiv \mathbb{1}\{CSR_3 \geq 0\}$ is an indicator variable that school s has implemented CSR in third grade, $treat_{sg}$ is a third grade dummy, and η_s, θ_t and δ_g represent school, year and grade fixed effects, respectively. Our coefficient of interest is β_3.

[60] As an aside, we note informally that the D-in-D results align with our main estimates if the structural δs and γs are inserted into equation (5.2). This is suggestive that the additivity assumption provides a reasonable approximation.

[61] We discuss the interpretation of this effect below – specifically, whether it is plausible to think that spillovers from incoming to existing public school students might be important.

"reduced approximately to half to one quarter of its previous magnitude" (Krueger and Whitmore 2001, page 11), although such test score gains then reappear later in the labor market (Chetty et al. 2011). These estimates are also consistent with fade-out estimates in the teacher effects literature (see Jacob, Lefgren and Sims 2010, and Kinsler 2012).

To summarize, the estimates indicate that the reform has a significant direct effect of 0.11σ (in terms of mathematics scores) and an indirect effect due to student sorting, measured on the same basis, that is even larger. Thus, the effects of general equilibrium sorting in response to a major quality-improving education reform are first order – a point we develop next. Further, we find that both direct and indirect effects persist strongly, at similar rates.

6.2 Magnitudes – Policy

Next, we turn to the implications of these estimates for the policy calculus. The scale of the general equilibrium sorting effect suggests that focusing only on the direct channel may substantially underestimate the overall impact of CSR. We assess the extent to which this is true by considering the test score benefits, followed by the fiscal consequences.

Benefits: We can use our framework to construct measures of the overall benefits of CSR, both in the short and longer term. After one year of exposure, students in the public school system are predicted to score 0.15σ higher – in other words, almost one-and-a-half times the direct effect. This total is the sum of 0.11σ due to the direct effect, 0.10σ from the indirect effect *excluding* its composition component,[62] and a 0.06σ *decline* in test scores due to the general equilibrium reduction in teacher quality.[63]

For the longer term, we estimate the effects of CSR cohorts experiencing the full four years of the program, based on the linear technology from equation (5.1) combined with our estimated persistence parameters. Here, we assume that teacher quality effects persist at the same rate as the direct effect (similar to the persistence estimates in Chetty, Friedman and Rockoff 2014). At the end of third grade, students who experienced CSR in grades K-3 are estimated to score fully

[62] We multiply the indirect effect by $\frac{5}{8}$ to eliminate its compositional component, in line with the average private-public switcher scoring at the seventy fifth percentile of the private school test score distribution – see Table 6. (We expand on the rationale in the next subsection.)

[63] This latter estimate comes from subtracting CSR teacher quality from non-CSR teacher quality in the final year we have data (2001-02) for in Table A.13. It is significantly larger in absolute terms than the -0.01σ found in Jepsen and Rivkin (2009).

0.29σ higher[64] than they would have in the absence of CSR.[65]

To monetize the longer term benefit of CSR, we carry out a suggestive calculation using estimates in Chetty et al. (2011), who find that a one standard deviation improvement in kindergarten class quality raises student test scores by 0.32σ, in turn raising lifetime earnings by approximately $39,100 for the average individual in 2009 dollars. Given that CSR increases student test scores by 0.15σ after one year, we conjecture that CSR increased class quality by about half a standard deviation in the units of Chetty et al. (2011) (since one σ improvement raises scores by 0.32σ in that setting), suggesting that CSR raised the present value of individual earnings by $18,300 (in 2009 dollars).

These estimates shed new light on the benefits of class size reduction, both given the size of the indirect sorting effect and the evidence of positive persistence above. In this regard, our results accord with the convincing studies that document longer-term benefits of class size reduction, focusing on Project STAR – see Krueger and Whitmore (2001) and Chetty et al. (2011). To the extent that CSR is representative of other major reforms intended to improve school quality, consideration of the test score benefits suggests that estimates that abstract from induced sorting are likely to suffer from considerable omitted variables bias. This view is only reinforced when assessing the fiscal costs, which follows next.

Costs: Once fully implemented, the State of California expected CSR to cost approximately $1.2 billion per year (in 1996 dollars), multiplying the 1.827 million eligible K-3 public school students by the $650 per student CSR funding.[66] In this calculation, the state neglected the fact that improving public school quality would then bring private school students into the public system, as we have documented. Specifically, CSR caused around 37,500 K-3 students to switch from private to public schools. The state average per student expenditure (including CSR funding) was $5,800, implying that the general equilibrium sorting effect raised the per-year cost of the program by about $220 million dollars (=37,500*5,800), or by 20 percent. By not taking account of the

[64]The test score effect for (end-of-grade) third grade students is calculated by substituting in the parameter estimates from Tables 5 and A.13 (in standard deviation units) into equation (5.1), adding an additional input for teacher quality. Specifically, we use our parameter estimates ($\hat{\gamma}_R = 0.11$, $\hat{\gamma}_X = 0.16$, $\hat{\gamma}_Q = -0.06$, $\hat{\delta}_R = 0.45$, $\hat{\delta}_X = 0.57$, δ_Q(assumed)$= \hat{\delta}_R$) into: $\hat{y}_3 = \sum_{\tau=0}^{3} \hat{\delta}_R^{\tau} \hat{\gamma}_R + \sum_{\tau=0}^{3} \hat{\delta}_X^{\tau} \hat{\gamma}_X + \sum_{\tau=0}^{3} \delta_Q^{\tau} \hat{\gamma}_Q$. In addition, we multiply $\hat{\gamma}_X$ by $\frac{5}{8}$ to eliminate the composition component of the indirect effect.

[65]For reference, our estimate is in the range reported by Unlu (2005), who compared California NAEP scores to other states before and after CSR and found that four years of exposure to CSR raised fourth grade mathematics test scores by 0.2-0.3σ.

[66]The state's cost expectations are available for the first few years only, and those values are on the low side since only a few grades were affected; for instance, the budgeted cost in the first year was $971 million. According to Brewer et al. (1999), actual costs in 1997-98 were $1.5 billion, but those include one-time funding of $300 million for facilities.

indirect costs of the reform, the state thus under-estimated the reform's total cost by *one-fifth*, highlighting the magnitude of the general equilibrium sorting response to CSR.[67] We add that these are the recurrent costs predicted in steady state, to be incurred each year. Given that, we expect CSR to cost about $2 billion a year in 2009 dollars, or around $1100 per student. In light of the $18,300 increase in the individual present value earnings created by CSR, our rough net benefit calculation implies a substantially positive benefit-to-cost ratio, on the order of fifteen.[68]

6.3 Decomposing the Indirect Effect

Given the size of the indirect effect we have estimated, one might wonder about the likely extent of spillovers experienced by existing public school students, arising as a result of general equilibrium sorting.

Conceptually, the indirect effect can be divided into two components – the compositional ('own') effect and the spillover effect. The compositional effect occurs mechanically because students who would have enrolled in private schools in the absence of the reform would typically be expected to score higher on standardized tests (on average) than their public school counterparts.[69] In turn, the spillover effect occurs when public school students receive benefits from their new, higher-scoring classmates, most likely through positive peer influences.

In order to decompose the 0.16σ indirect effect estimated in Section 6 into these two components, as is our goal here, one needs to know the test score of the marginal private school student induced into the public school system due to CSR – the 'private-public switcher.' While the average test score of the private-public switchers is unobserved in our data, we are able to carry out the decomposition under different scenarios relating to the percentile rank that the average switcher is drawn from, and construct informative bounds.

Table 6 sets out the relevant results. Each column reports a different percentile of the private school test score distribution the private-public switcher could be drawn from. We take public and private school test score distributions from the 1996 California NAEP fourth grade results.[70] The first row reports the average test score of the private-public switchers in terms of the public school test score distribution. Since class-level standard deviations are much smaller than individual-level standard deviations, we multiply increases in individual-level standard deviations by three to place

[67] Here we only focus on the permanent costs of the reform to the government, and not the temporary funds for helping schools make the transition to smaller classes.

[68] Of course, policy makers should place CSR alongside other feasible alternative policies, many of which are significantly less costly.

[69] Angrist and Lang (2004) are able to leverage individual data to estimate compositional effects of the Metco program in a convincing way.

[70] See Table 2.7A in `https://files.eric.ed.gov/fulltext/ED425943.pdf`. All effects sizes are normalized at the school-grade level to be mean zero and standard deviation one in the public school system.

them in the distribution of school-grade test scores.[71]

It is reasonable to expect that private-public switchers are relatively high up the private school test score distribution – high-ability, low-income students are likely to be the most responsive to an increase in public school quality (see Epple and Romano (1998), for example). Under the scenario where the private-public switcher is at the 75^{th} percentile of the private school test score distribution (column (2) in the table), around sixty percent (or 0.10/0.16) of the indirect effect comes from positive spillovers onto public school students arising from peer effects.[72] The implied social multiplier of 1.73 if the private-public switcher is at the 75^{th} percentile is very similar to that in the thorough study by Graham (2008), who uses a linear-in-means peer effects model and Project STAR data to estimate a social multiplier of 1.86.

Overall, for plausible scenarios in which the private-public switcher is drawn from the mean or higher in the private school score distribution, the spillover effect accounts for between 50 and 80 percent of the estimated indirect effect (the entries in the third row of the table, divided by 0.16 – see columns (1)-(3)). While this evidence is suggestive, it does point to additional benefits from the reform that many pre-existing public students are likely to enjoy.

6.4 Indirect Sorting Effects: General Relevance

The estimated size of the indirect sorting effect we obtained from California's CSR reform is likely to carry over to other settings when certain pre-conditions hold (as they do in California). The reform-related shock to public school quality should be large; pre-reform, the private school share needs to be high; and further, the characteristics of students in private versus public schools have to differ so that changes in peer quality occur post-reform.

The same pre-conditions are found in other US states, for example. At the time of CSR, California ranked 20th (out of 50 states) in its private school enrollment rate (Yun and Reardon 2005). And California's large private-public test score gap is typical of most other states – Altonji, Elder and Taber (2005) report a national eighth grade private-public test score gap of 0.4σ, for instance. Thus it is reasonable to expect similar sorting responses to large reform-related shocks

[71] See Finn and Achilles (1990), where effect sizes are three-fold in the distribution of class means relative to individual means, for instance.

[72] The relevant calculation starts from the 1.4 percent decline in private school share estimated in Column (4) of Table 3. Based on this, we expect that an average school-grade with enrollment of fifty-five students will (in expectation) receive 0.77 of a private school student entering their school. Considering the average switcher at the 75^{th} percentile of the private school distribution, using the column (2) scenario for illustration, will lead to a 0.02 ($= \frac{1.42*0.77}{56}$) increase in the *student*-level distribution. We then multiply by three to convert this increase in the student-level distribution to the school-grade-level distribution. This gives the 'Composition Effect' entry of 0.06 in the second row of the third column. The 'Spillover Effect' entry on the third row, of $0.10 = 0.16 - 0.06$, is the total indirect effect minus the composition effect. The implied social multiplier in the fourth row is then given by the ratio of the spillover effect to the composition effect (in the second column, $1.73 = 0.10/0.06$).

to public school quality elsewhere in the United States.

The size of the sorting effect will be larger to the extent that private schools are relatively passive to the reform, and students in private schools are more responsive to relative changes in quality between public and private schools, placing a larger share on the margin of switching. In terms of the former, we find some suggestive evidence of adjustments on the part of private schools, serving to mitigate the size of the sorting effect we have estimated. Specifically, following CSR, fewer private schools entered in the state (relative to trend), and more private schools exited, consistent with evidence from New York City presented in Dinerstein and Smith (2016). On the quality margin, we also see suggestive evidence that private schools responded to the boost in public school quality associated with CSR by lowering their own class sizes.[73] To keep track of the proportion of marginal students, a model of public and private school behavior and individual 'consumer' choice comes naturally to mind, although estimating that would require more disaggregated information about the decisions of the relevant economic agents than our California data provide.[74]

7 Conclusion

In this paper, we have presented a transparent approach for estimating general equilibrium sorting effects of major reforms. While such effects may be sizeable, they are typically difficult to identify, and so have not been a prime focus of policy-oriented empirical research.

Central to our approach is a framework that relates education inputs to measurable outcomes, allowing those inputs to have persistent effects. We have shown how the framework's key parameters can be identified by applying a multiple differencing procedure – one that leverages two sources of exogenous variation: in the way local public goods are provided (differences in school grade span, for instance), and in the reform's coverage, where some groups are affected while others are not (as may occur due to budgetary considerations). Both sources of variation are available in many settings. Further, the data requirements for the approach we propose are minimal.

We developed the empirical anlysis in the context of a major education reform of the late-1990s – California's CSR program. Using the grade-specific timing of the reform, we first showed that CSR caused a significant decrease in private school shares and marked compositional changes in the public school system. Then, applying our multiple differencing strategy, we estimated an indirect sorting effect of the policy at least as important (in our education setting) as the *direct* policy effect, which has been the focus of careful measurement in the prior literature.

[73] These latter results are available on request.

[74] For example, Bayer, Ferreira and McMillan (2004) uses an equilibrium sorting model estimated using census microdata to gauge the reinforcing effect of improvements to public school quality that work through household location choices.

Applying the approach, we were also able to recover the persistence of the direct and indirect effects of the reform. Once these are accounted for, we showed how the combined benefits of CSR in the short term are almost one and a half times the direct effect; in the longer term, the combined benefits are even greater. On the cost side, we showed that indirect sorting leads to recurrent expenditures that are a fifth higher than in the state's own projections. And a suggestive net benefit calculation points to a benefit-cost ratio substantially in excess of one. (Of course, policy markers should still weigh CSR's net benefit alongside those of other feasible reforms as a means of improving education outcomes.)

Beyond class size reduction policies, our approach and estimates are relevant when assessing the effects of major reforms in other contexts. Alternative education reforms with different cost implications – for instance, incentive-based policies – that boost public school quality are also likely to change the mix of students across public and private systems, with consequences for education production of the kind our framework can accommodate. We leave using the approach to estimate indirect general equilibrium effects in other settings for future work.

References

Altonji, Joseph G., Todd E. Elder, and Christopher R. Taber. 2005. "Selection on observed and unobserved variables: Assessing the effectiveness of Catholic schools." *Journal of Political Economy*, 113(1): 151–184.

Angrist, Joshua D, and Kevin Lang. 2004. "Does school integration generate peer effects? Evidence from Boston's Metco Program." *American Economic Review*, 94(5): 1613–1634.

Angrist, Joshua D, and Victor Lavy. 1999. "Using Maimonides' rule to estimate the effect of class size on scholastic achievement." *Quarterly Journal of Economics*, 114(2): 533–575.

Angrist, Joshua D., Erich Battistin, and Daniela Vuri. 2017. "In a small moment: Class size and moral hazard in the Italian Mezzogiorno." *American Economic Journal: Applied Economics*, 9(4): 216–49.

Bayer, Patrick, Fernando Ferreira, and Robert McMillan. 2004. "Tiebout sorting, social multipliers and the demand for school quality." National Bureau of Economic Research Working Paper 10871.

Bianchi, Nicola. 2017. "The indirect effects of educational expansions: Evidence from a large enrollment increase in STEM majors." Unpublished manuscript.

Bohrnstedt, George W., and Brian M. Stecher. 2002. "What we have learned about class size reduction in California. Capstone report." Unpublished manuscript.

Brewer, Dominic J., Cathy Krop, Brian P. Gill, and Robert Reichardt. 1999. "Estimating the cost of national class size reductions under different policy alternatives." *Educational Evaluation and Policy Analysis*, 21(2): pp. 179–192.

Buddin, Richard. 2012. "The impact of charter schools on public and private school enrollments." *Cato Institute Policy Analysis*, 707.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014. "Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood." *American Economic Review*, 104(9): 2633–2679.

Chetty, Raj, John N. Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore-Schanzenbach, and Danny Yagan. 2011. "How does your kindergarten classroom affect your earnings? Evidence from Project STAR." *Quarterly Journal of Economics*, 126(4): 1593–1660.

Dinerstein, Michael, and Troy Smith. 2016. "Quantifying the supply response of private schools to public policies." Unpublished manuscript.

Ding, Weili, and Steven F. Lehrer. 2010. "Estimating treatment effects from contaminated multiperiod education experiments: The dynamic impacts of class size reductions." *Review of Economics and Statistics*, 92(1): 31–42.

Epple, Dennis, and Richard E. Romano. 1998. "Competition between private and public schools, vouchers, and peer-group effects." *American Economic Review*, 33–62.

Finn, Jeremy D, and Charles M Achilles. 1990. "Answers and questions about class size: A statewide experiment." *American Educational Research Journal*, 27(3): 557–577.

Gilraine, Michael. 2017. "Identifying multiple treatments from a single discontinuity with an application to class size caps." Unpublished manuscript.

Graham, Bryan S. 2008. "Identifying social interactions through conditional variance restrictions." *Econometrica*, 76(3): 643–660.

Hoxby, Caroline M. 2000. "The effects of class size on student achievement: New evidence from population variation." *Quarterly Journal of Economics*, 115(4): 1239–1285.

Jacob, Brian A., Lars Lefgren, and David P. Sims. 2010. "The persistence of teacher-induced learning." *Journal of Human Resources*, 45(4): 915–943.

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): 223–250.

Kinsler, Josh. 2012. "Beyond levels and growth estimating teacher value-added and its persistence." *Journal of Human Resources*, 47(3): 722–753.

Krueger, Alan B. 1999. "Experimental estimates of education production functions." *Quarterly Journal of Economics*, 114(2): 497–532.

Krueger, Alan B., and Diane M. Whitmore. 2001. "The effect of attending a small class in the early grades on college-test taking and middle school test results: Evidence from Project STAR." *Economic Journal*, 111(468): 1–28.

Lucas, Greg. 1996. "Sacramento Surprise – Extra Funds / Governor wants to use money to cut class size." *San Francisco Chronicle*.

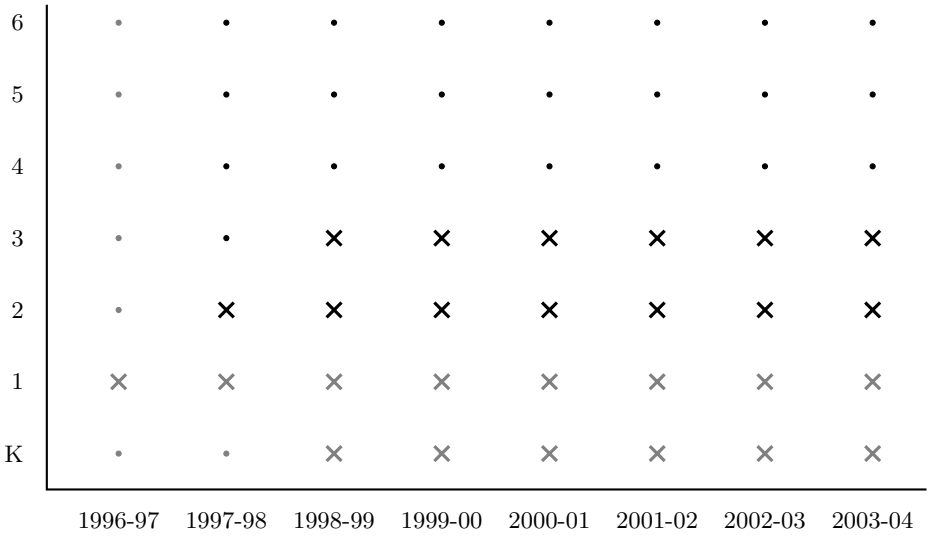
Rivkin, Steven G., Eric A. Hanushek, and John F. Kain. 2005. "Teachers, schools, and academic achievement." *Econometrica*, 73(2): 417–458.

Schrag, Peter. 2006. "Policy from the Hip: Class size reduction in California." *Brookings Papers on Education Policy*, 2006(1): 229–243.

Unlu, Fatih. 2005. "California class size reduction reform: New findings from the NAEP." Unpublished manuscript.

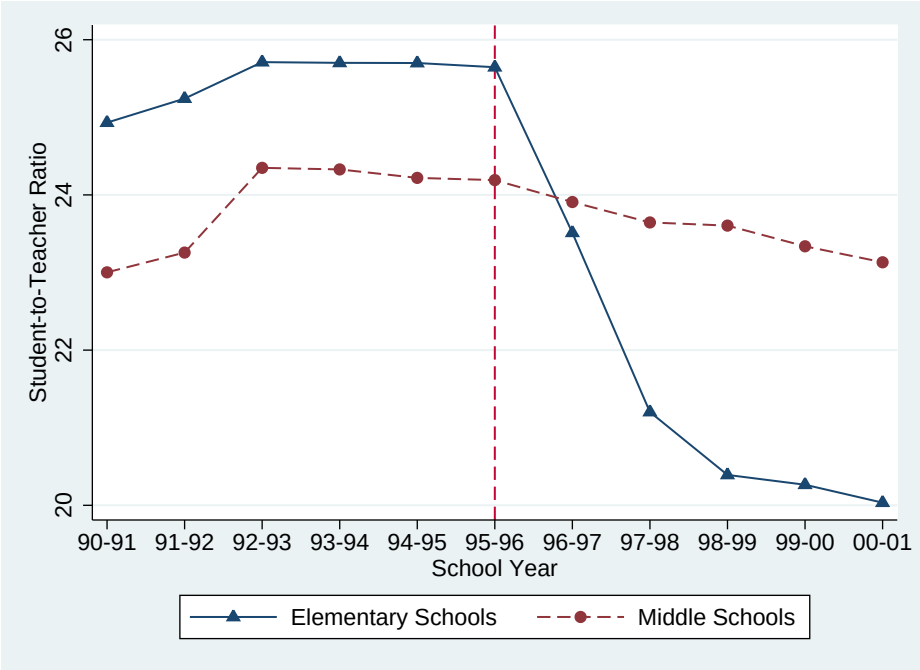
Yun, John T., and Sean F. Reardon. 2005. "Private school racial enrollments and segregation." *School choice and diversity: What the evidence says*, 42–58.

Figure 1: Policy Coverage and Data Availability



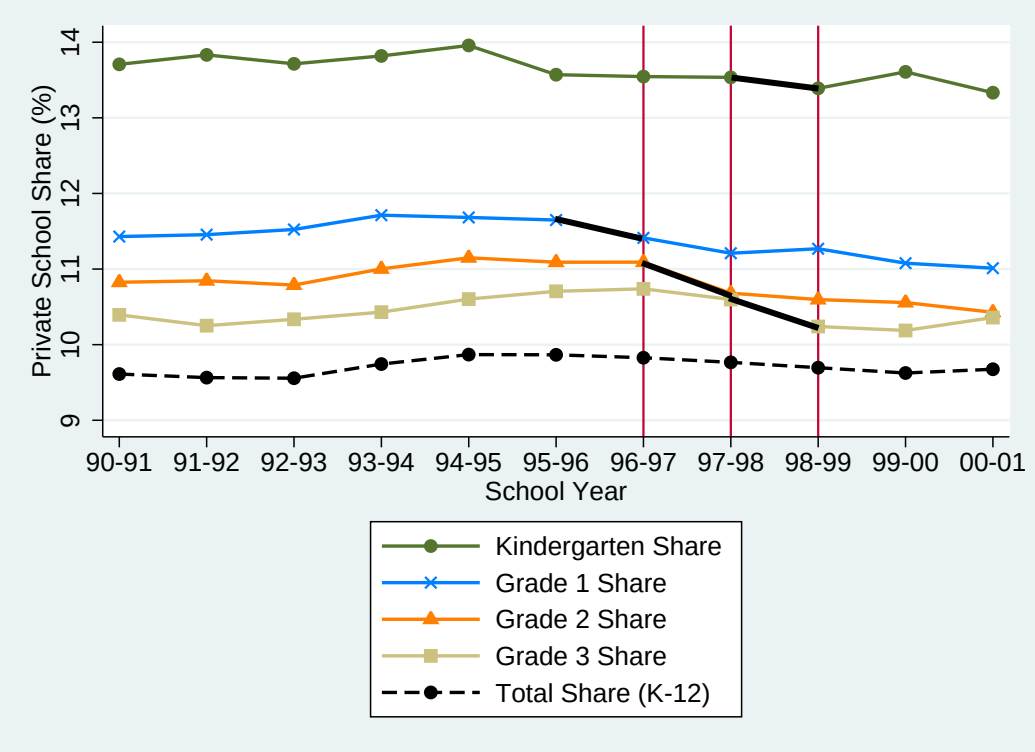
Notes: The reform is in effect for a particular grade (vertical axis) and year (horizontal axis) combination if the corresponding cell contains a '×' symbol and it is not if it contains a '·' symbol. While the earliest grade of implementation is kindergarten (K), test score data are only available for grades two and above and from 1997-98 onward. The bottom two rows and leftmost column use a lighter shading to reflect this.

Figure 2: Class Sizes in California over Time



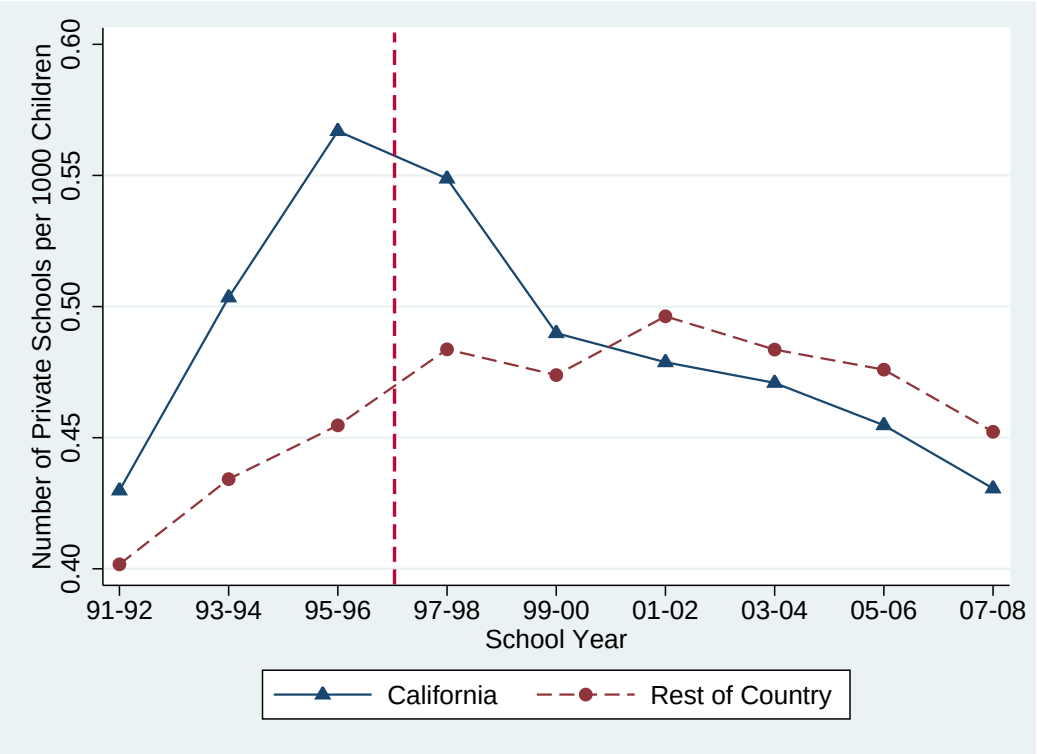
Notes: This figure shows student-to-teacher ratios by year for school years 1990-91 through 2000-01. The student-to-teacher ratio is defined as the number of students in a school divided by the number of teachers at that school. Given that CSR only affects grades K-3, we expect that this will substantially underestimate the change in K-3 class sizes induced by CSR. Elementary schools are defined as any school that includes grades K-3 and whose highest grade is 6 or below. Middle schools are schools that do not have a K-3 grade and whose highest grade is 9 or below. The vertical line represents the start of the 1995-96 school year, the last year before CSR was implemented in the 1996-97 school year.

Figure 3: Private School Share Trends by Grade



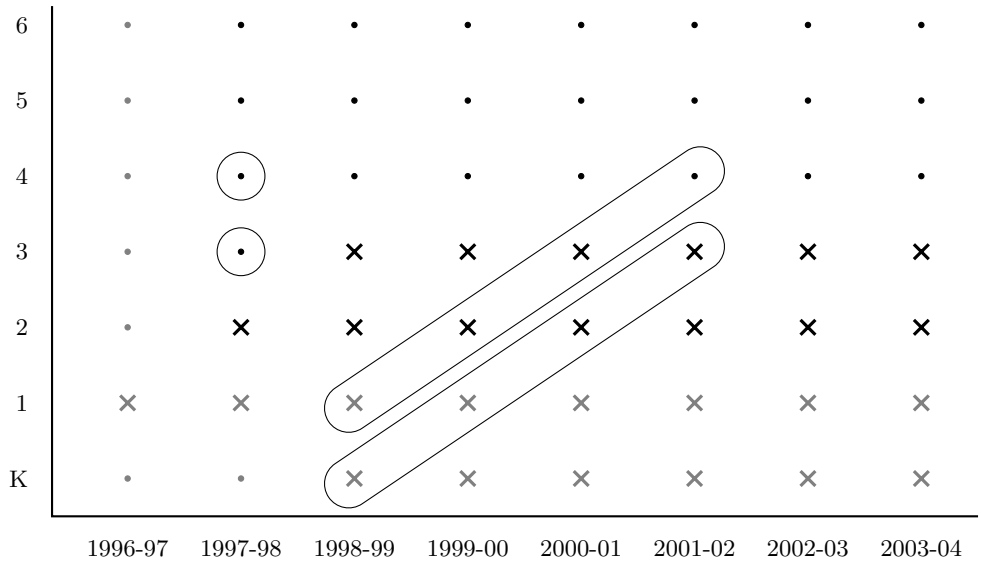
Notes: This figure shows aggregate private school share trends by grade over the years surrounding CSR. 'Private School Share' is defined as the aggregate number of students in private school in each grade in the state divided by the total number of public and private school students in that grade. The vertical lines represent the start of school years 1996-97, 1997-98 and 1998-99 respectively, when different grades became eligible for CSR. Specifically, first grade became eligible for the 1996-97 school year, second grade for the 1997-98 school year, and third grade and kindergarten for the 1998-99 school year. The darkened thick line segments indicate the effect of CSR on the grade-level private school share when CSR was first implemented for that particular grade.

Figure 4: Number of Private Schools per 1000 School-Aged Children by Year



Notes: The dashed vertical line indicates the 1996-97 introduction of the CSR reform. Data are available only every two years. The figure only includes private schools that primarily serve CSR grades. A private school is determined to serve CSR grades if, on average, at least twenty percent of its student body is in K-3 grades in the 1989-90 through 2013-14 school years. The population of children are defined as all individuals aged 5-17 living in a state.

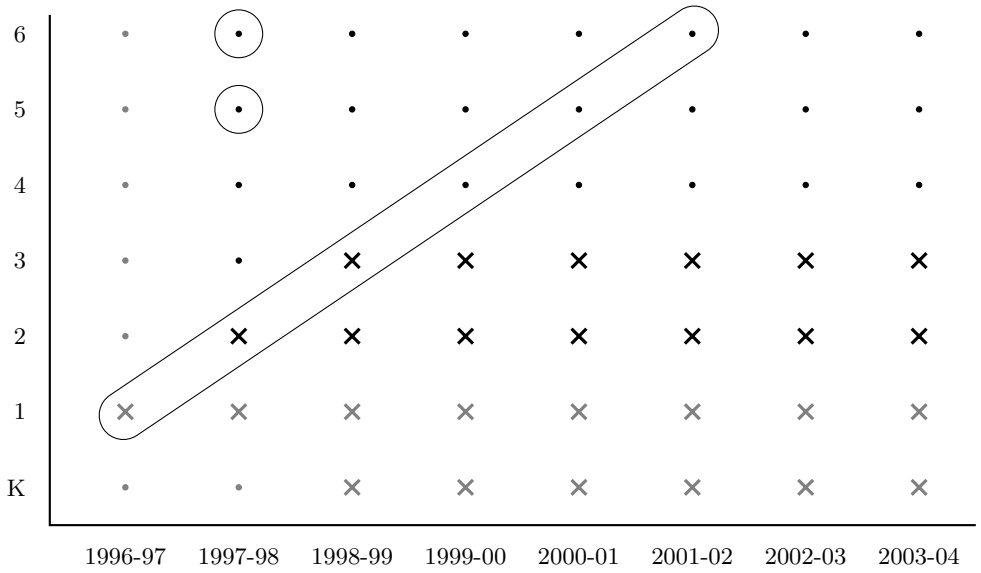
Figure 5: Key Differencing Variation for the Direct Effect



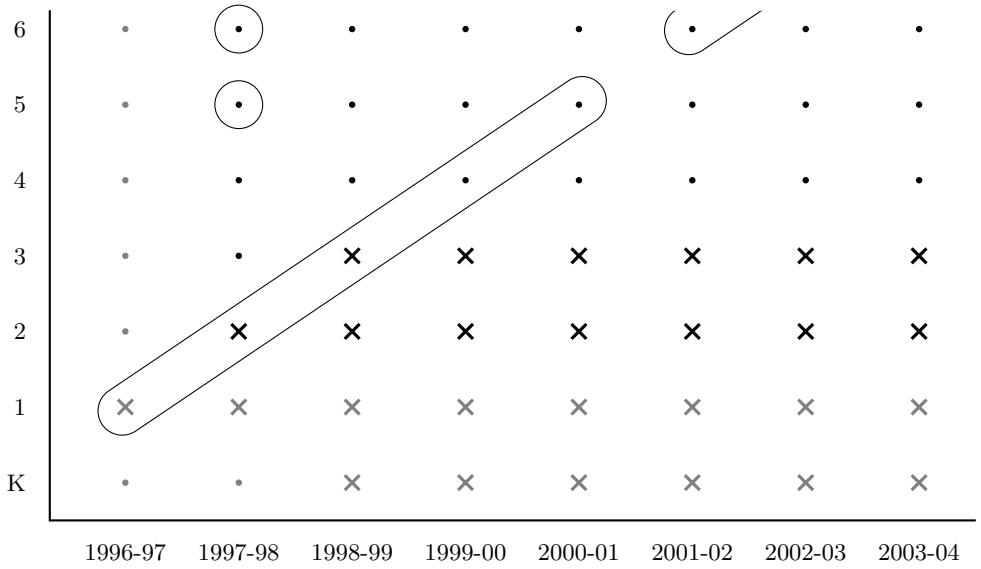
Notes: Policy coverage and data availability are as described in the notes to Figure 1. The key variation used to recover the direct effect through differencing is highlighted using four different outlines.

Figure 6: Key Differencing Variation for the Indirect Effect

(a) K-6 Grade Configuration



(b) K-5 Grade Configuration



Notes: Policy coverage and data availability are as described in the notes to Figure 1. The key variation used to recover the indirect effect through differencing is highlighted using the outlines in each panel (contrasting K-6 and K-5 grade configurations).

Table 1: Descriptive Statistics

	Overall Mean (1990-91 to 2008-09)	Pre-CSR (90-91 to 95-96)	CSR (96-97 to 99-00)	Post-CSR (00-01 to 08-09)
School Data				
Elementary Student-to-Teacher Ratio[1]	22.7	24.9	22.6	21.6
Private School Share (%)	9.3	9.9	9.9	8.8
	(8.5)	(8.5)	(8.5)	(8.4)
CSR Intensity[2]	89.9	90.2	90.0	89.7
	(17.1)	(16.4)	(17.0)	(17.4)
% English Learner[3]	35.8	32.1	34.3	38.1
	(19.8)	(19.8)	(20.0)	(19.5)
% White	36.1	42.6	38.6	32.2
	(25.2)	(26.2)	(25.7)	(23.8)
% Hispanic	42.3	36.6	40.3	45.8
	(24.6)	(23.4)	(24.1)	(24.8)
% Black	8.3	8.7	8.7	8.0
	(8.5)	(9.4)	(9.0)	(7.9)
% Asian	8.3	8.3	8.4	8.3
	(9.9)	(9.0)	(9.6)	(10.5)
Enrollment	578	533	572	606
	(2260)	(2135)	(2249)	(2331)
% Free and Reduced Price Meals[4]	47.7	41.9	48.0	50.6
	(23.1)	(21.6)	(23.1)	(23.3)
Observations (District-Grade-Year)	208,285	63,983	32,761	111,541

Notes: This table shows descriptive statistics of outcome variables along with student demographics before, during and after CSR implementation. All variables are weighted by district-grade-year enrollment with the exception of enrollment. Demographic data only include public school students.

[1] Elementary Student-to-Teacher Ratio is calculated as the number of students in a school divided by the number of teachers in that elementary school. Elementary schools are defined as any school that includes grades K-3 and whose highest grade is 6 or below.

[2] 'CSR Intensity' measures the proportion of K-3 students in CSR school-grades in the 1998-99 school year. The measure varies slightly year-to-year due to district closures and missing data from some districts in some years (87% of observations are from districts with at least 20 years of data).

[3] Some observations are missing values for this variable. There are a total of 185,249 observations with non-missing values.

[4] This variable is only available at the district-year level and has 19,311 observations.

Table 2: Effect of CSR on Private School Share

Dependent Variable: Private School Share (%)			
	Untreated Grades (Grades 4-12)	Treated Grades (Grades K-3)	Difference (Untreated-Treated)
Before CSR	8.88	11.76	-2.87
	(8.80)	(7.62)	(0.25)
After CSR	8.43	10.19	-1.76
	(8.79)	(7.38)	(0.29)
Change	0.45	1.56	-1.11
(Before-After)	(0.14)	(0.24)	(0.17)
Observations	136,408	71,877	208,285

Notes: This table shows changes private school share means for 'treated' CSR grades (K-3) and 'untreated' non-CSR grades (4-12) before and after CSR is implemented for that grade. Differences-in-means are then taken between these 'treated' and 'untreated' grades. Taking the difference in these difference-in-means yields the difference-in-differences estimate from equation (4.1) without controls. This point estimate also corresponds to column (1) of Table 3. Observations are at the district-grade-year level, and cover 1990-91 through 2008-09 school years. Means are weighted by district-grade-year enrollment. Standard errors for the difference-in-means cells are clustered at the district level.

Table 3: Difference-in-Differences Estimates of CSR on Private School Share

Outcome Variable: Private School Share (%)				
	(1)	(2)	(3)	(4)
Treatment*Post	-1.11***	-1.04***	-0.99***	-1.40***
	(0.17)	(0.18)	(0.27)	(0.27)
Post	-0.45***	0.24	0.10	0.38**
	(0.14)	(0.15)	(0.16)	(0.18)
Treatment	2.87***	-	-	-
	(0.25)			
Year/Grade FE	No	Yes	Yes	Yes
Demographic Controls	No	No	Yes	Yes
District FE	No	No	No	Yes
Observations	208,285	208,285	173,129	173,129

Notes: This table shows results from the difference-in-differences regression described by equation (4.1) with varying levels of controls. Observations are at the district-grade-year level and cover the 1990-91 through 2008-09 school years. Demographic controls include student race, gender, English second language, enrollment and enrollment squared. The 'treatment' variable is omitted for columns (2)-(4) since it is collinear with the grade fixed effects. All regressions are weighted by district-grade-year enrollment. Standard errors are clustered at the district level. ***,** and * denote significance at the 1%, 5% and 10% levels, respectively.

Table 4: Regression-Discontinuity Estimates by Grade Span

Outcome Variable: Private School Share for Grade Span					
	Kindergarten (Placebo)	Elementary School CSR Grades (1-3)	Elementary School non-CSR Grades (4-6)	Middle School Grades (7-8)	High School Grades (9-12)
	(1)	(2)	(3)	(4)	(5)
Average Effect	-0.07	-0.30**	-0.30**	-0.10	0.03
	(0.22)	(0.15)	(0.15)	(0.28)	(0.13)
Observations	2,874	8,825	9,251	6,390	11,680

Notes: This table reports results from the regression discontinuity design defined in equation (4.2) that exploits differential exposure of cohorts to the reform. The kindergarten effect here represents a placebo test as kindergarten was not a CSR grade for the cohorts around the discontinuity. Observations are at the district-cohort-grade level. To calculate average effects across grade spans, we estimate a separate local linear regression allowing for a different functional form on either side of the cutoff (see equation (4.2)) for each grade. We then average these grade-level estimates to find the average effect over the grade span. The bandwidth used is three. Standards errors are calculated using the delta method and are clustered at the district level. Demographic controls and district fixed effects are used in all regressions. ***,** and * denote significance at the 1%, 5% and 10% levels, respectively.

Table 5: Structural Estimates

Outcome Variable: Mathematics Test Scores				
	With $\psi = \frac{2}{3}$		With $\psi = 1$	
	(1)	(2)	(3)	(4)
γ_R	2.10***	2.22***	2.10***	2.22***
	(0.20)	(0.20)	(0.20)	(0.20)
γ_X	2.27	3.26**	1.63	2.42**
	(1.69)	(1.54)	(1.24)	(1.13)
δ_R	0.49*	0.45**	0.50*	0.46**
	(0.26)	(0.21)	(0.26)	(0.21)
δ_X	0.64**	0.57**	0.70**	0.62**
	(0.30)	(0.27)	(0.31)	(0.30)
County FE	No	Yes	No	Yes
Observations	147,636	147,636	147,636	147,636

Notes: This table shows estimates of the structural parameters described in Section 5. Observations are at the school-grade-year level, and cover the 1997-98 through 2001-02 school years. Mathematics test scores are shown in percentile ranks relative to a national norming sample, where one percentile rank roughly equates to 0.05σ in the distribution of school-grade level test scores. All parameter estimates include controls for teacher quality. Standard errors for γ_R and γ_X are computed using the delta method and are clustered at the school level. Standard errors for δ_R and δ_X are bootstrapped. ***,** and * denote significance at the 1%, 5% and 10% levels, respectively.

Table 6: Compositional and Spillover Effects by Private-Public School Switcher Percentile

	Average Test Score Percentile of Private-Public Switchers				
	90^{th}	75^{th}	Mean	50^{th}	25^{th}
	(1)	(2)	(3)	(4)	(5)
Average 'Switcher' Test Score	1.88	1.42	0.82	0.79	0.21
Compositional Effect	0.08	0.06	0.03	0.03	0.01
Spillover Effect	0.08	0.10	0.13	0.13	0.15
Implied Social Multiplier	1.06	1.73	3.73	3.91	17.47

Notes: This table decomposes the 0.16σ spillover effect estimated in Section 6 into compositional and spillover components. Given that the test score of the marginal private-public switcher is not observed, each column reports a different percentile of the private school test score distribution the average private-public switcher could be drawn from. Public and private school test score distributions are taken from the 1996 California NAEP fourth grade results (see Table 2.7A in `https://files.eric.ed.gov/fulltext/ED425943.pdf`). All effect sizes are normalized at the school-grade level to be mean zero and standard deviation one in the public school system. (The table layout is explained in detail in Section 6.3.)

Appendix A California State Testing – a Quick Primer

Statewide testing in California started in 1961 for mathematics, reading and writing in grades 5, 8 and 10. In 1972, the California Assessment Program was created, which tested reading in grades 2 and 3 and mathematics, reading and writing in grades 6 and 12. It lasted (with a few test additions) until 1991, when it was replace by the California Learning Assessment System (CLAS), which covered reading, writing and mathematics in grades 4, 5, 8 and 10.

In 1994, under public pressure from civil rights groups that the CLAS was inaccurate and intruded upon students' privacy (due to numerous race-based questions on the test), the Governor vetoed a Senate bill to extend CLAS.[75] As a consequence, there were no statewide tests for the 1994-95 and 1995-96 school years, although districts often did conduct standardized tests during this time; the state even provided funding for this through the Pupil Testing Incentive Program.

In the 1996-97 school year, the Standardized Testing and Reporting program (STAR) – an initiative of the Governor – was implemented, which tested reading, writing and math in grades 2-8 and reading, writing, mathematics, history, and science in grades 9-11. The test used by STAR was the Stanford 9, a nationally normed multiple-choice achievement test. Additional test items in language arts and in mathematics were included in the 1999-00 through 2002-03 tests to cover material in the California content standards that were not addressed by the Stanford 9. Besides this small addition, the STAR program was relatively unchanged until 2002-03 (see below). These are the tests we use in this study.

In time for the 2002-03 school year, California's STAR program was reauthorized and the State Board of Education issued a request for potential contractors to submit proposals for administering STAR. The contract was won by CTB/McGraw-Hill, and led to the test being changed from the Stanford Achievement Test (run by Harcourt Educational Measurement) to the California Achievement Tests. Test scores reported by the two tests differed dramatically, with no systematic linking of scores between the two tests being conducted. Given this test change, we focus on the 1996-97 through 2001-02 (inclusive) test scores in this paper. California's STAR testing program was officially terminated after the 2012-13 school year and was replaced by the California Assessment of Student Performance and Progress.

[75] The Governor stated that his veto was due to the fact that it did not give teachers and parents individual student achievement scores (scores were available at the school level only).

Appendix B Private School Evidence: Robustness

In this appendix, we explore the robustness of our estimates of the impact of CSR on private school share described in Section 4.1. We do so in two ways: (i) we extend our difference-in-differences design to a triple-differences design using district-level CSR participation intensity as an additional dimension of differencing, and (ii) we provide support for the 'parallel trends' assumption by plotting coefficient estimates by year and looking for any significant pre-trends in outcomes. We also present the estimating equation we use to identify the effect of CSR on the number of private schools.

To extend our difference-in-differences design in Section 4.1 to a triple-differences design, we calculate a measure of the intensity of CSR implementation by school district. This takes advantage of the fact that, while most districts opted into CSR,[76] the school-grade level implementation was uneven across them. As school-level CSR participation data are only available for the 1998-99 through 2003-04 school years, we define our 'local intensity' measure (CSR_d) as the percentage of K-3 students in a CSR participating school-grade within a district for the 1998-99 school year.[77] Formally,

$$CSR_d = \frac{\sum_{s \in d} \sum_{g=0}^{3} \mathbb{1}\{CSR_{sg}\} * (enroll_{sg})}{\sum_{s \in d} \sum_{g=0}^{3} enroll_{sg}}, \tag{B.1}$$

where $enroll_{sg}$ is the enrollment of grade g students in school s and district d (kindergarten is defined as $g = 0$), and $\mathbb{1}\{CSR_{sg}\}$ is an indicator for whether the school implemented CSR for the particular grade in the 1998-99 school year.

Using this local intensity measure, the triple-differences analysis is implemented by estimating the following weighted regression:

$$\begin{aligned} share_{dgt} = {} & \beta_0 + \beta_1 post_{gt} + \beta_2 (post_{gt} * treat_g) + \beta_3 (post_{gt} * CSR_d) + \beta_4 (treat_g * CSR_d) \\ & + \beta_5 (post_{gt} * treat_g * CSR_d) + \eta_d + \theta_t + \delta_g + \phi X_{dgt} + \epsilon_{dgt}\,, \end{aligned} \tag{B.2}$$

where all variables other than the intensity measure CSR_d are identical to those in equation (4.1). The triple-differences coefficient of interest is β_5. Identification of the parameter depends on a less restrictive variant of the parallel trends assumption in Section 4.1: the difference in the evolution of private school share between CSR and non-CSR grades would have been the same for low- and high-share CSR districts in the absence of the reform.

[76] In the first year of CSR, only 56 of 895 districts in California did not opt-in. In the following year, twenty districts remained non-participating districts. For every year thereafter in our sample period, the number of non-participating districts was about ten.

[77] Results are similar if this variable is averaged over the 1998-99 through 2003-04 school years.

Given that our triple-differences identification strategy exploits variation in the local intensity of adoption, Figure A.1 shows the spatial variation in our district-level CSR adoption intensity measure, CSR_d. There substantial geographic variation in our CSR intensity measure, with high levels of CSR adoption in regions such as San Diego and the Bay Area and low levels in regions such as the southern end of the Central Valley.

Using district-level CSR participation intensity as an additional dimension of differencing, our preferred triple-differences analysis from equation (B.2) yields similar findings to the difference-in-differences estimates from Section 4.1.[78] With all controls, column (4) of Table A.3 shows that CSR is associated with a 1.3 percentage point decline in private school share. Thus, private school share experienced a substantial reduction as a result of the reform, concentrated precisely in the grades that were treated and in school districts that implemented the reform in a faithful way.

Finally, we provide support for the 'parallel trends' assumption that underlies these results by plotting coefficient estimates by year. Figure A.2(a) does so for the main difference-in-differences specification defined by equation (4.1), while Figure A.2(a) does the same for the triple-differences specification given by equation (B.2). Both figures show that there is no effect on private school share prior to the implementation of the reform,[79] followed by a clear decline afterwards.

Number of Private Schools: We also conduct an analysis of the effect of CSR on the number of private schools. To do so, we rely on data from the Private School Universe Survey and the U.S. Census (see Table A.1 for more details) and implement a difference-in-differences approach, comparing the number of private schools in California to the rest of the United States before and after the CSR reform came into effect. We then add an additional layer of differencing by comparing private schools that predominantly serve students in CSR grades to those predominantly serving students in non-CSR grades.[80] Specifically, we run the following triple-differences regression:

$$\begin{aligned} private_{cst} = \beta_0 + \beta_1 CSR_c + \beta_2 (CSR_c * CA_s) + \beta_3 (CSR_c * post_t) + \beta_4 (CA_s * post_t) \\ + \beta_5 (CSR_c * CA_s * post_t) + \gamma_s + \theta_t + \epsilon_{cst}\,, \end{aligned} \tag{B.3}$$

where $private_{cst}$ is the number of private schools per one thousand 5-17 year old children with grade configuration c in state s in year t, CSR_c is an indicator equal to one if more than twenty percent of the private school's student body is in CSR grades, CA_s is an indicator for the state of

[78] It is important to note that the difference-in-differences and triple-differences estimates are not directly comparable, since almost all districts have some level of CSR implementation. Thus, the triple-differences coefficient cannot be interpreted as the effect of CSR relative to a non-CSR baseline, as such a comparison extends beyond the support of the data.

[79] More formally, a chi-squared test finds that the impacts before the reform are not jointly significant.

[80] We define a school as 'predominantly serving students in CSR grades' if more than twenty percent of their student body is in a K-3 grade.

California, $post_t$ indicates whether or not CSR has been implemented and γ_s and θ_t are state and year fixed effects, respectively. The triple-differences coefficient of interest is β_5, which identifies the impact of CSR on the number of private schools under the assumption that the difference in the evolution of the number of private schools serving CSR and non-CSR grades would have been the same for California and the rest of the United States in the absence of the reform. Results from this regression are reported in Table A.4.

Appendix C Public School Composition

In this appendix, we describe our econometric approach (alluded to in Section 4.2) for assessing the extent to which re-sorting between private and public schools altered the composition of students in public school. To do so, we implement a triple-differences design that starts with the first two layers of differencing from (4.1) whereby CSR grades are compared to non-CSR grades before and after the reform was implemented. We then add a third dimension of differencing that takes into account whether a private school is nearby (which we discretize). The weighted estimating equation is:

$$\begin{aligned} demo_{sgt} = \beta_0 + \beta_1(post_t * treat_g) + \beta_2(post_t * \mathbb{1}\{Buffer < x\ km\}_s) + \beta_3(treat_g * \mathbb{1}\{Buffer < x\ km\}_s) \\ + \beta_4(post_t * treat_g * \mathbb{1}\{Buffer < x\ km\}_s) + \eta_s + \theta_t + \delta_g + \phi X_{sgt} + \epsilon_{sgt}\,, \end{aligned} \tag{C.1}$$

where $demo_{sgt}$ is the demographic share of interest for grade g student in school s at time t, $post_t$ indicates whether CSR had been implemented, $treat_g$ indicates whether grade g was subject to the CSR reform, $\mathbb{1}\{Buffer < x\ km\}_s)$ is an indicator for whether a private school serving CSR grades[81] is within a x km radius of school s,[82] X_{sgt} is a set of school-grade-year covariates, and η_s, θ_t and δ_g are school, time and grade fixed effects, respectively.

The triple-differences coefficient of interest is β_4. To identify it, we assume that the difference in the change in demographic share between CSR and non-CSR grades would have been the same for public schools within x km of a private school and those farther away in the absence of the reform. Results from this regression are reported in Table A.5.

Similar to Appendix B, we provide support for the 'parallel trends' assumption by computing difference-in-differences estimates by year, using the treatment of grades (CSR versus non-CSR) and the distance to the nearest private school competitor (within 3 kilometres versus more than 3 kilometres) as the two dimensions of differencing. Figure A.4 does this for two public school demographic variables: percent white and percent Hispanic. In both cases, the point estimates are

[81] Only private schools with ten or more students in kindergarten through third grade are included.
[82] In Table A.5, we report results for buffers of $1.5km$, $3km$ and $5km$.

indistinguishable from zero in the pre-reform years, with the yearly effects becoming statistically and economically significant once CSR is implemented.

Identifying the Indirect Effect: Reduced-form Analog. Identification of the indirect effect in our framework can be thought of in a more reduced-form way. Effectively, we are comparing sixth grade versus fifth grade in K6 schools versus K5 schools for cohorts affected by CSR versus those that were not: this type of comparison is analogous to running a triple-differences regression. Framing our indirect effect in this manner provides a simple check on our identification strategy: differences between K6 schools and K5 schools among cohorts affected and unaffected by CSR should only appear in the sixth grade versus fifth grade comparison and not the other grade comparisons we can also make.

To assess whether that is indeed the case, Figure A.7 shows triple-differences regression coefficients (using grade g versus $g-1$, K6 versus K5 schools and cohorts affected versus unaffected by CSR as the three layers of differencing) for each grade g and $g-1$ combination.[83] As expected, we only find a significant triple-differences estimate between grades 6 and 5, with our triple-differences estimate for the grade 6 versus grade 5 comparison (0.12σ) being close to our structural indirect estimate (0.16σ).[84] The triple-differences between all other grade g and $g-1$ comparisons are statistically indistinguishable from zero.

Appendix D Estimating Equations for Multiple Differencing

This appendix sets out the main equations used in our multiple differencing approach. It first discusses the identification of the main parameters without teacher effects. We then add teacher effects, discussing the method we use to estimate teacher quality before explaining how we incorporate teacher quality into the main estimating equations.

[83] Specifically, we restrict our data to grades g and $g-1$ and schools with K-5 or K-6 configurations. We then use the following regression:

$$\begin{aligned} y_{sgt} = \alpha + \phi_g G_g + \phi_k K6_s + \phi_t post_t + \zeta_{gk} G_g * K6_s + \zeta_{gt} G_g * post_t + \zeta_{kt} K6_s * post_t \\ + \Phi_{K6-K5,g-(g-1),post-pre} G_g * K6_s * post_t + \phi X_{sgt} + \epsilon_{sgt} \,, \end{aligned} \tag{C.2}$$

where y_{sgt} is the test score in school s in grade g at time t, G_g is an indicator for grade g, $K6_s$ is an indicator for the K-6 grade span configuration (i.e. $K6 = 1$ denotes the K-6 configuration), $post_t$ refers to the 2001-02 school year and later, and X_{sgt} represent school-grade-year characteristics. Our coefficient of interest is $\Phi_{K6-K5,g-(g-1),post-pre}$, which compares g and $g-1$ grade scores between K-5 and K-6 schools before and after the 2001-02 school year (when the first CSR cohort entered sixth grade). Since 2001-02 represents the first sixth grade cohort that experienced CSR, we expect that $\Phi_{K6-K5,6-5,post-pre}$ will be positive and (roughly) similar in magnitude to γ_X. All other triple-differences between adjacent grades are placebo tests.

[84] The difference between the triple-differences estimate and our structural estimate can be attributed to the fact that our structural estimate only uses one pre- year and one post- year of data, while the triple-differences regression uses multiple pre- and post- years.

D.1 Without Teacher Effects

We take each of the key parameters of the technology in turn:

γ_R: Identification of γ_R comes from equation (5.3) in the main text. Here, we derive that equation, which subtracts $\widehat{\Delta y}_{4,01-02}$ from $\widehat{\Delta y}_{3,01-02}$:

$$\begin{aligned}\widehat{\Delta y}_{3,01-02} - \widehat{\Delta y}_{4,01-02} = &(\delta_R)^3\gamma_R\Delta R_{0,98-99} + (\delta_R)^2\gamma_R\Delta R_{1,99-00} + \delta_R\gamma_R\Delta R_{2,00-01} + \gamma_R\Delta R_{3,01-02} \\ &+ (\delta_X)^3\gamma_X\Delta X_{0,98-99} + (\delta_X)^2\gamma_X\Delta X_{1,99-00} + \delta_X\gamma_X\Delta X_{2,00-01} + \gamma_X\Delta X_{3,01-02} + \Delta\epsilon_{3,01-02} \\ &- ((\delta_R)^3\gamma_R\Delta R_{1,98-99} + (\delta_R)^2\gamma_R\Delta R_{2,99-00} + \delta_R\gamma_R\Delta R_{3,00-01} \\ &+ (\delta_X)^3\gamma_X\Delta X_{1,98-99} + (\delta_X)^2\gamma_X\Delta X_{2,99-00} + \delta_X\gamma_X\Delta X_{3,00-01} + \gamma_X\Delta X_{4,01-02} + \Delta\epsilon_{4,01-02}) \\ &= \gamma_R\Delta R_{01-02} + (\Delta\epsilon_{3,01\text{-}02} - \Delta\epsilon_{4,01\text{-}02}), \end{aligned} \tag{D.1}$$

where the final equality comes the fact that CSR affected all grades equally once it was implemented, so that $\Delta R_{gt} = \Delta R_{g't}$ and $\Delta X_{gt} = \Delta X_{g't}\ \forall g, g'$ (Assumption 3). We then invoke the parallel trends assumption (Assumption 4) and use test score differences between third and fourth grades *before* the reform to act as a counterfactual for test score differences after the reform. Using this, we have that:

$$\gamma_R\Delta R_{01-02} = y_{3,01-02} - y_{4,01-02} - (y_{3,97-98} - y_{4,97-98}). \tag{D.2}$$

γ_X: Identification of γ_X comes from equation (5.5), which is derived fully in the main text. We relax the parallel trends assumption (Assumption 4), using two levels of differencing to act as the counterfactual. First, to account for systematic differences between K-6 and K-5 schools, we use fifth grade test scores in K-5 ($y_{5,01-02,K6}$) and K-6 schools ($y_{5,01-02,K5}$) as our first level of differencing. Then we use the pre-reform test scores for both fifth and sixth grades, $y_{5,97-98}$ and $y_{6,97-98}$, in K-5 and K-6 schools as counterfactuals for the observed test scores in fifth and sixth grades in the 2001-02 school year. Therefore, we have:[85]

$$\begin{aligned}\psi\gamma_X\Delta X_{6,01-02} = &[y_{6,01-02,K6} - y_{5,01-02,K6} - (y_{6,97-98,K6} - y_{5,97-98,K6})] \\ &- [y_{6,01-02,K5} - y_{5,01-02,K5} - (y_{6,97-98,K5} - y_{5,97-98,K5})]. \end{aligned} \tag{D.3}$$

(δ_R, δ_X): Identification of δ_R and δ_X takes the parameters γ_R and γ_X to be known and differences the test scores in fourth grade and third grade in the 2000-01 school year, which yields:[86]

[85] Here, we are over-identified since we could use 1997-98, 1998-99 and 1999-00 as counterfactuals: those cohorts in fifth and sixth grades were not subject to CSR in those three years. In practice, we use all three and take an average of the estimates, although estimates are quantitatively similar regardless which counterfactual year is used.

[86] $\Delta y_{3,99-00}$–$\Delta y_{4,99-00}$ yields the same structural equation as $\Delta y_{3,00-01}$–$\Delta y_{4,00-01}$ and $\Delta y_{5,01-02}$–$\Delta y_{4,01-02}$ yields the same structural equation as $\Delta y_{5,00-01} - \Delta y_{4,00-01}$. This equation is therefore over-identified. Once again, we use all both equations and take an average of the estimates, although estimates are quantitatively similar regardless which structural equation is used.

$$\begin{aligned}\widehat{\Delta y}_{4,00-01} - \widehat{\Delta y}_{3,00-01} &= (\delta_R)^3\gamma_R\Delta R_{1,97-98} + (\delta_R)^2\gamma_R\Delta R_{2,98-99} + \delta_R\gamma_R\Delta R_{3,99-00} \\ &+ (\delta_X)^3\gamma_X\Delta X_{1,97-98} + (\delta_X)^2\gamma_X\Delta X_{2,98-99} + \delta_X\gamma_X\Delta X_{3,99-00} + \gamma_X\Delta X_{4,00-01} + \Delta\epsilon_{4,01-02} \\ &- ((\delta_R)^2\gamma_R\Delta R_{1,98-99} + \delta_R\gamma_R\Delta R_{2,99-00} + \gamma_R\Delta R_{3,00-01} \\ &+ (\delta_X)^2\gamma_X\Delta X_{1,98-99} + \delta_X\gamma_X\Delta X_{2,99-00} + \gamma_X\Delta X_{3,00-01} + \Delta\epsilon_{3,01-02}) \\ &= (\delta_R)^3\gamma_R\Delta R_{1,97-98} - \gamma_R\Delta R_{3,00-01} + (\delta_X)^3\gamma_X\Delta X_{1,97-98} + (\Delta\epsilon_{4,01-02} - \Delta\epsilon_{3,01-02})\,. \quad \text{(D.4)}\end{aligned}$$

Similarly, comparing test scores between fourth and fifth grade in the 2000-01 school year yields:

$$\begin{aligned}\widehat{\Delta y}_{5,00-01} - \widehat{\Delta y}_{4,00-01} &= (\delta_R)^4\gamma_R\Delta R_{1,96-97} + (\delta_R)^3\gamma_R\Delta R_{2,97-98} + (\delta_R)^2\gamma_R\Delta R_{3,98-99} + (\delta_X)^4\gamma_X\Delta X_{1,96-97} \\ &+ (\delta_X)^3\gamma_X\Delta X_{2,97-98} + (\delta_X)^2\gamma_X\Delta X_{3,98-99} + \delta_X\gamma_X\Delta X_{4,99-00} + \gamma_X\Delta X_{5,00-01} + \Delta\epsilon_{5,00-01} \\ &- ((\delta_R)^3\gamma_R\Delta R_{1,97-98} + (\delta_R)^2\gamma_R\Delta R_{2,98-99} + \delta_R\gamma_R\Delta R_{3,99-00} \\ &+ (\delta_X)^3\gamma_X\Delta X_{1,97-98} + (\delta_X)^2\gamma_X\Delta X_{2,98-99} + \delta_X\gamma_X\Delta X_{3,99-00} + \gamma_X\Delta X_{4,00-01} + \Delta\epsilon_{4,00-01}) \\ &= (\delta_R)^4\gamma_R\Delta R_{1,96-97} - \delta_R\gamma_R\Delta R_{3,99-00} + (\delta_X)^4\gamma_X\Delta X_{1,96-97} + (\Delta\epsilon_{5,00-01} - \Delta\epsilon_{4,00-01}). \quad \text{(D.5)}\end{aligned}$$

Since CSR affected all grades equally, we have that $\Delta R_{1,96-97} = \Delta R_{1,97-98} = \Delta R_{3,99-00} = \Delta R_{3,00-01}$ and $\Delta X_{1,96-97} = \Delta X_{3,97-98}$. This is effectively Assumption 3 (grade-invariant input levels), although there is an additional component here that the input levels were also time-invariant once CSR was implemented. Suppressing the grade and year notation on the ΔR_{gt} and ΔX_{gt} variables and invoking Assumption 4 (parallel trends) yields the following two equations with two unknowns (δ_R, δ_X):

$$y_{4,00-01} - y_{3,00-01} - (y_{4,97-98} - y_{3,97-98}) = \gamma_R\Delta R((\delta_R)^3 - 1) + (\delta_X)^3\gamma_X\Delta X \quad \text{(D.6)}$$

$$y_{5,00-01} - y_{4,00-01} - (y_{5,97-98} - y_{4,97-98}) = \delta_R\gamma_R\Delta R((\delta_R)^3 - 1) + (\delta_X)^4\gamma_X\Delta X\,. \quad \text{(D.7)}$$

D.2 Estimating Teacher Quality

This subsection explains in detail how we incorporate the estimation of teacher quality into our multiple differencing approach.

Let $Q^l_{CSR,t}$ and $Q^l_{non,t}$ denote the effect of teacher quality in year t for students in a CSR and non-CSR grade, respectively. We allow these effects to persist by using the l superscript, which represents the effect of being treated to a CSR or non-CSR teacher $l \geq 0$ periods ago (where 0 is the contemporaneous effect). Note that we do not look at teacher quality at the grade level, but rather distinguish between CSR and non-CSR grades, since CSR should affect teachers across all CSR grades equally.

Our data begin in 1997-98, after the initial increase in the share of inexperienced teachers due to CSR's sudden introduction. The proportion of inexperienced teachers is similar across CSR (second and third) and non-CSR (fourth) grades for that first year.[87] An interesting pattern emerges over the next three years once the CSR program expands to kindergarten and third grade: teacher inexperience falls substantially for CSR grades and rises for non-CSR grades. Inexperience then falls for all grades thereafter.[88]

We incorporate variation in teacher inexperience into our strategy by estimating the teacher quality parameters $Q^l_{CSR,t}$ and $Q^l_{non,t}$ for each lag l according to the following two-step procedure. First, we regress test scores in 1997-98 + l ($y_{s,g,\text{97-98}+l}$) on the share of teacher inexperience in 1997-98 ($X_{s,g,\text{97-98}}$), including grade fixed effects (ϕ_g):

$$y_{s,g,\text{ 97-98}+l} = \kappa_l X_{s,g,\text{97-98}} + \phi_g + \epsilon_{s,g,\text{97-98}}\,.$$

Second, we use the resulting estimate $\hat{\kappa}_l$ to compute teacher quality relative to the 1997-98 baseline:[89]

$$\begin{aligned} Q^l_{CSR,t} &= \hat{\kappa}_l \times (X_{3,t} - X_{3,\text{97-98}}) \\ Q^l_{non,t} &= \hat{\kappa}_l \times (X_{4,t} - X_{4,\text{97-98}})\,, \end{aligned}$$

where CSR and non-CSR values of Q use variation in third- and fourth-grade inexperience, respectively. Thus, the relevant parameters to compute $\hat{\kappa}_R$, are $Q^0_{CSR,\text{01-02}} = \hat{\kappa}_0 \times (X_{3,\text{2001-02}} - X_{3,\text{97-98}})$, $Q^0_{non,\text{01-02}} = \hat{\kappa}_0 \times (X_{4,\text{01-02}} - X_{4,\text{97-98}})$ and $Q^4_{CSR,\text{97-98}} = \hat{\kappa}_4 \times (X_{4,\text{97-98}} - X_{4,\text{97-98}}) = 0$. The necessary parameters to compute $\hat{\gamma}_X$ are estimated analogously.[90]

D.3 Estimating Equations with Teacher Effects

We incorporate general equilibrium teacher effects by controlling for differences in observed teacher quality proxies. Given the teacher effects (as defined in Appendix D.2), we express differences between observed and counterfactual test scores allowing for differences in teacher quality according

[87] Inexperience in fifth and sixth grades is close to but slightly lower than for second through fourth grades.

[88] It may seem puzzling why schools would maintain teacher quality for CSR grades at the expense of non-CSR grades, since formal incentives under the 1999 Public Schools Accountability Act were not provided differentially by grade. Schools perhaps believed policymakers were paying closer attention to CSR grades or schools may have worked to ensure the success of a promising reform.

[89] Defining teacher quality relative to 1997-98 controls for preexisting differences between grades that are unrelated to the implementation of the CSR program. Using 1997-98 as a baseline is justified given that CSR had yet to apply to third grade in that year. Indeed, Table A.12 shows that the share of teacher inexperience is essentially identical across third and fourth grades in 1997-98.

[90] We estimate the parameters $Q^2_{CSR,\text{00-01},K6}$, $Q^2_{non,\text{00-01},K6}$, $Q^2_{CSR,\text{00-01},K5}$ and $Q^2_{non,\text{00-01},K5}$. Due to a lack of test score data in 1996-97, the parameters $Q^5_{CSR,\text{96-97},K5/K6}$ and $Q^5_{non,\text{96-97},K5/K6}$ cannot be estimated and are thus omitted from our estimating equations. However, as with $Q^4_{CSR,97-98}$, we can assume that they are negligible since teacher quality across grades is likely to be similar in 1996-97 and 1997-98 across K5 and K6 schools.

to whether students were in a CSR or non-CSR grade. For example, the difference between observed and counterfactual third grade test scores in 2001-02 can be expressed as:

$$\begin{aligned}\widehat{\Delta y}_{3,01-02} &= (\delta_R)^3\gamma_R\Delta R_{0,98-99} + (\delta_R)^2\gamma_R\Delta R_{1,99-00} + \delta_R\gamma_R\Delta R_{2,00-01} + \gamma_R\Delta R_{3,01-02} \\ &+ (\delta_X)^3\gamma_X\Delta X_{0,98-99} + (\delta_X)^2\gamma_X\Delta X_{1,99-00} + \delta_X\gamma_X\Delta X_{2,00-01} + \gamma_X\Delta X_{3,01-02} \\ &+ \gamma_Q(Q^3_{CSR,98-99} + Q^2_{CSR,99-00} + Q^1_{CSR,00-01} + Q^0_{CSR,01-02}) + \Delta\epsilon_{3,01-02}. \qquad \text{(D.8)}\end{aligned}$$

γ_R: Incorporating general equilibrium teacher effects, the differences between observed and counterfactual test scores that yield γ_R can be expressed in terms of the parameters as followings:

$$\begin{aligned}y_{3,01-02} - y_{4,01-02} - (y_{3,97-98} - y_{4,97-98}) &= \gamma_R\Delta R_{3,01-02} \\ &+ \gamma_Q(Q^3_{CSR,98-99} + Q^2_{CSR,99-00} + Q^1_{CSR,00-01} + Q^0_{CSR,01-02}) \\ &- \gamma_Q(Q^4_{non,97-98} + Q^3_{CSR,98-99} + Q^2_{CSR,99-00} + Q^1_{CSR,00-01} + Q^0_{non,01-02}) \\ &= \gamma_R\Delta R_{3,01-02} + \gamma_Q(Q^0_{CSR,01-02} - Q^0_{non,01-02} - Q^4_{non,97-98})\,. \qquad \text{(D.9)}\end{aligned}$$

γ_X: Similarly, the differences between observed and counterfactual test scores that yield γ_X can be expressed in terms of the parameters in the following way:

$$\begin{aligned}&[y_{6,01-02,K6} - y_{5,01-02,K6} - (y_{6,97-98,K6} - y_{5,97-98,K6})] \\ &- [y_{6,01-02,K5} - y_{5,01-02,K5} - (y_{6,97-98,K5} - y_{5,97-98,K5})] = \psi\gamma_X\Delta X_{6,01-02,K6} \\ &+\gamma_Q(Q^5_{CSR,96-97,K6}+Q^4_{CSR,97-98,K6}+Q^3_{CSR,98-99,K6}+Q^2_{non,99-00,K6}+Q^1_{non,00-01,K6}+Q^0_{non,01-02,K6}) \\ &-\gamma_Q(Q^4_{CSR,97-98,K6}+Q^3_{CSR,98-99,K6}+Q^2_{CSR,99-00,K6}+Q^1_{non,00-01,K6}+Q^0_{non,01-02,K6}) \\ &-[\gamma_Q(Q^5_{CSR,96-97,K5}+Q^4_{CSR,97-98,K5}+Q^3_{CSR,98-99,K5}+Q^2_{non,99-00,K5}+Q^1_{non,00-01,K5}+Q^0_{non,01-02,K5}) \\ &-\gamma_Q(Q^4_{CSR,97-98,K5}+Q^3_{CSR,98-99,K5}+Q^2_{CSR,99-00,K5}+Q^1_{non,00-01,K5}+Q^0_{non,01-02,K5})] \\ &= \psi\gamma_X\Delta X_{6,01-02,K6} + \gamma_Q(Q^5_{CSR,96-97,K6} + Q^2_{non,99-00,K6} - Q^2_{CSR,99-00,K6}) \\ &- [\gamma_Q(Q^5_{CSR,96-97,K5} + Q^2_{non,99-00,K5} - Q^2_{CSR,99-00,K5})]\,. \qquad \text{(D.10)}\end{aligned}$$

(δ_R, δ_X): Finally, to solve for δ_R and δ_X, we incorporate teacher effects into the final two regressions:

$$\begin{aligned}y_{4,00-01} - y_{3,00-01} - (y_{4,97-98} - y_{3,97-98}) &= \gamma_R\Delta R((\delta_R)^3 - 1) + (\delta_X)^3\gamma_X\Delta X \\ &+ \gamma_Q(Q^4_{non,96-97} + Q^3_{CSR,97-98} + Q^2_{CSR,98-99} + Q^1_{CSR,99-00} + Q^0_{non,00-01}) \\ &- \gamma_Q(Q^3_{non,97-98} + Q^2_{CSR,98-99} + Q^1_{CSR,99-00} + Q^0_{CSR,00-01}) \\ &= \gamma_R\Delta R((\delta_R)^3 - 1) + (\delta_X)^3\gamma_X\Delta X \\ &+ \gamma_Q(Q^4_{non,96-97} + Q^3_{CSR,97-98} - Q^3_{non,97-98} + Q^0_{non,00-01} - Q^0_{CSR,00-01})\,. \qquad \text{(D.11)}\end{aligned}$$

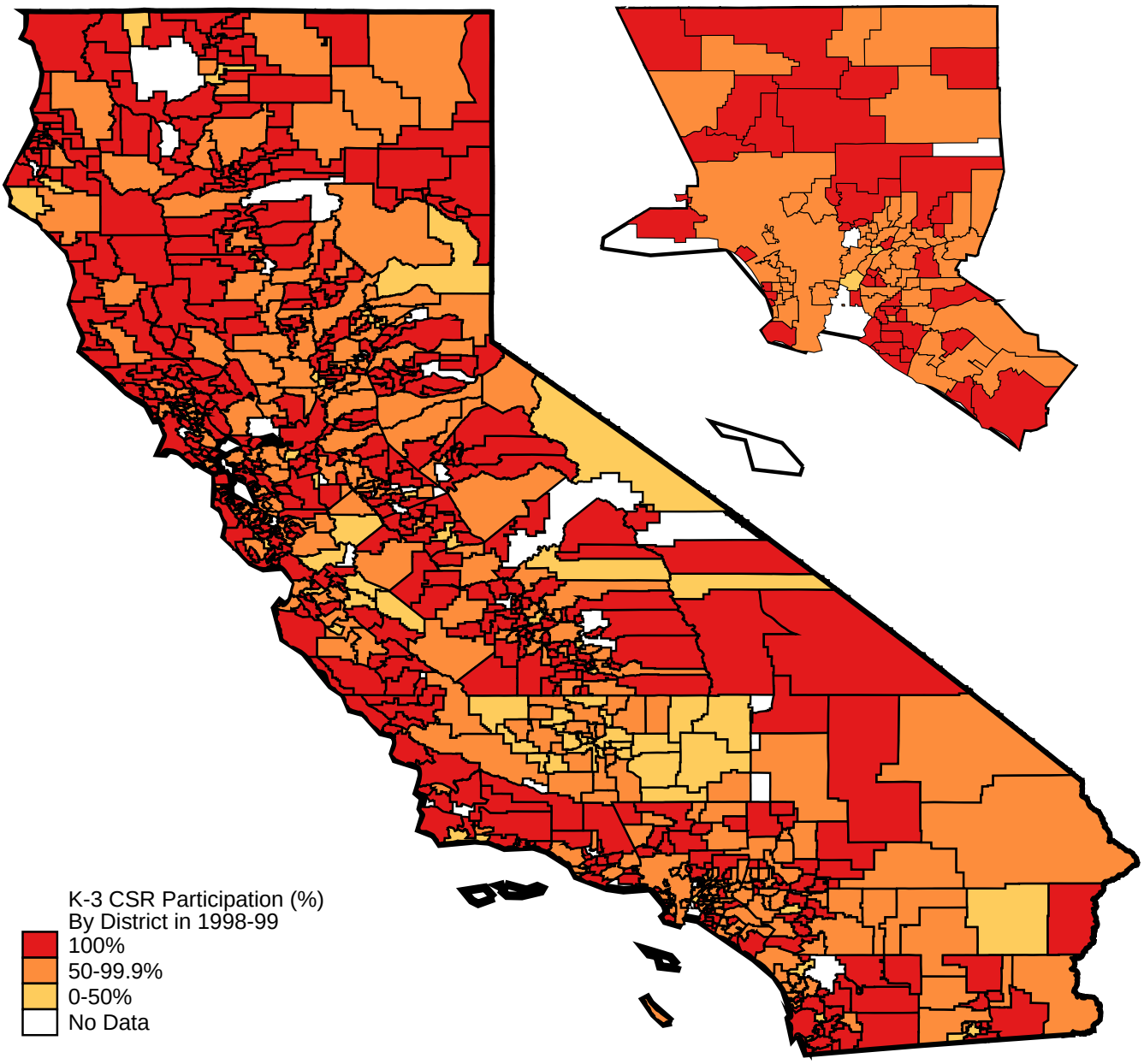
$$
\begin{aligned}
y_{5,00-01} - y_{4,00-01} - (y_{5,97-98} - y_{4,97-98}) &= \delta_R \gamma_R \Delta R((\delta_R)^3 - 1) + (\delta_X)^4 \gamma_X \Delta X \\
&+ \gamma_Q (Q^4_{CSR,96-97} + Q^3_{CSR,97-98} + Q^2_{CSR,98-99} + Q^1_{non,99-00} + Q^0_{non,00-01}) \\
&- \gamma_Q (Q^4_{non,96-97} + Q^3_{CSR,97-98} + Q^2_{CSR,98-99} + Q^1_{CSR,99-00} + Q^0_{non,00-01}) \\
&= \delta_R \gamma_R \Delta R((\delta_R)^3 - 1) + (\delta_X)^4 \gamma_X \Delta X \\
&+ \gamma_Q (Q^4_{CSR,96-97} - Q^4_{non,96-97} + Q^1_{non,99-00} - Q^1_{CSR,99-00}) \,, \qquad \text{(D.12)}
\end{aligned}
$$

where the time and grade subscripts on ΔX and ΔR have been dropped (as in equation (D.6)).

APPENDIX FIGURES AND TABLES

Figure A.1: K-3 CSR Participation by District in 1998-99 ('CSR Intensity' Measure)

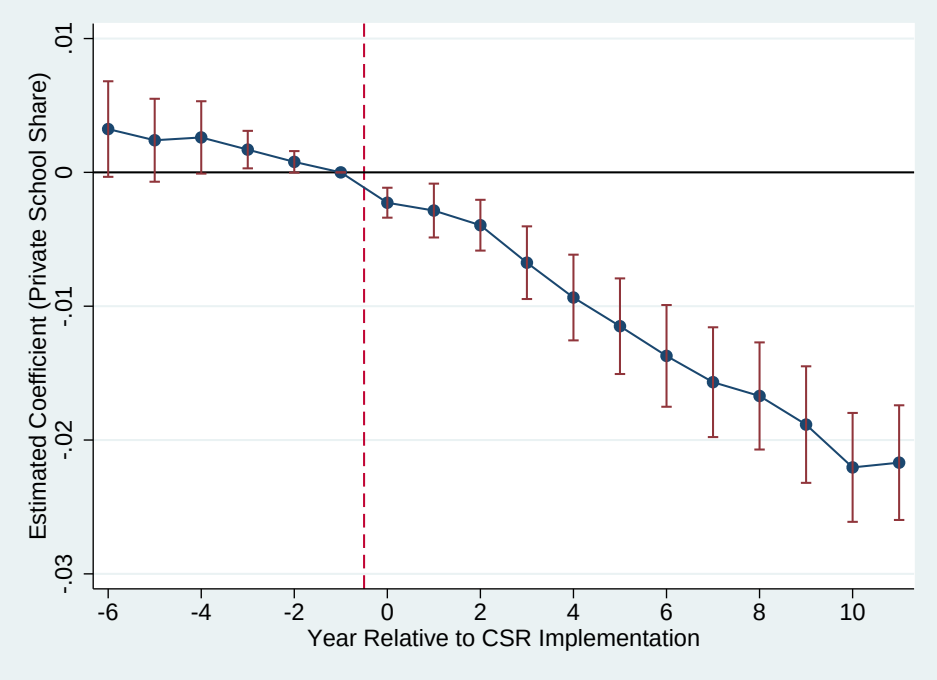
(a) California **(b)** Los Angeles and Orange Counties



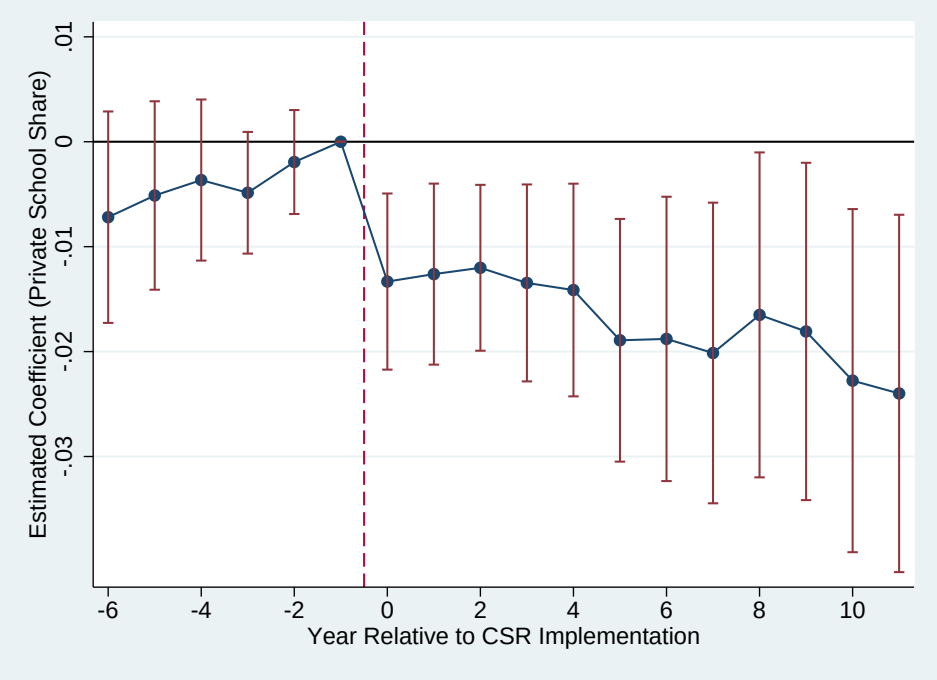
Notes: The above figure shows the percentage of district-level K-3 enrollment in a CSR-participating school-grade for the 1998-99 school year. Los Angeles and Orange Counties combined are shown separately for better visualization of that region. White areas denote regions that cannot be assigned to a school district.

Figure A.2: The Effect of CSR on Private School Share by Year

(a) Difference-in-Differences



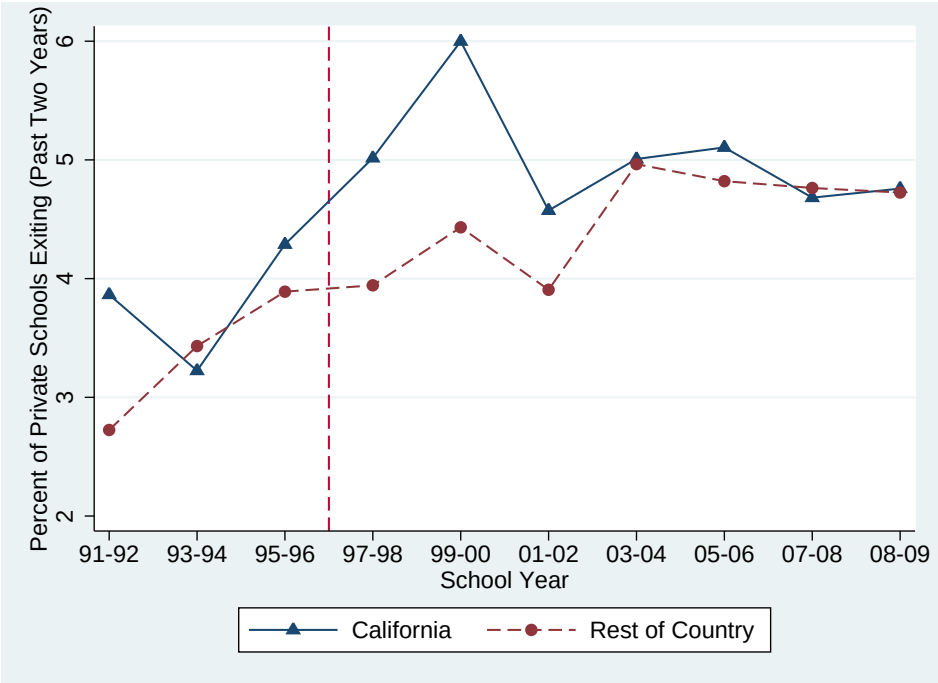
(b) Triple-Differences



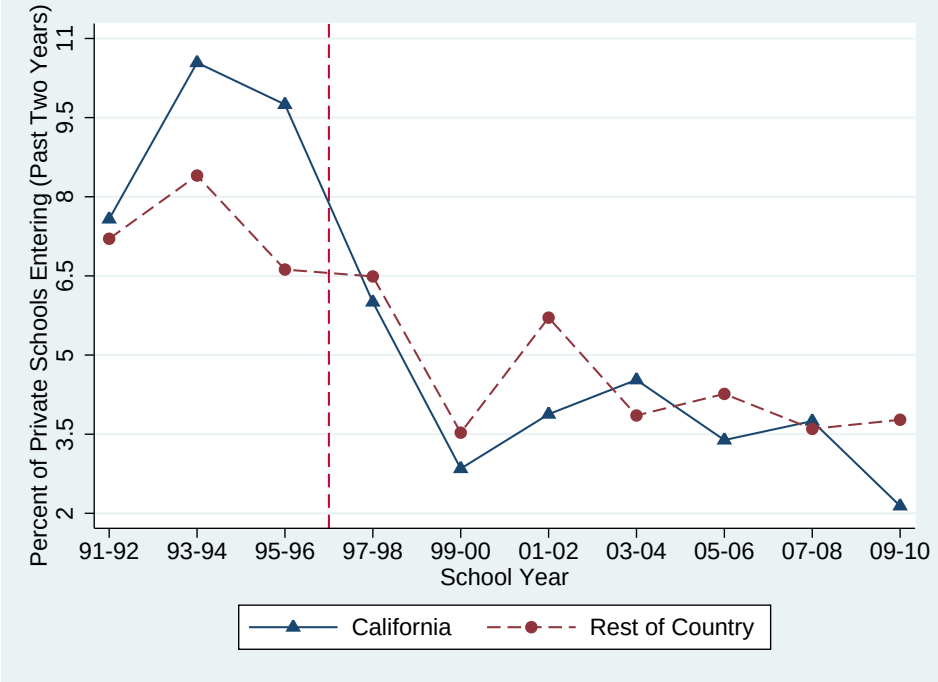
Notes: Figure A.2(a) shows the estimated change in private school share by year in 'treated' CSR grades (K-3) relative to 'untreated' non-CSR grades (4-12). Figure A.2(b) adds district-level CSR participation intensity as an additional layer of differencing. In both figures, the dashed vertical line represents the start of CSR implementation while the horizontal line indicates an estimate of zero. The estimated coefficient for the year prior to the start of CSR implementation is normalized to zero. Vertical bands represent 95% confidence intervals for each point estimate. Covariates and grade, year and district fixed effects are included. Standard errors are clustered at the district level.

Figure A.3: Biennial Private School Entry and Exit Rates

(a) Private School Exit Rates



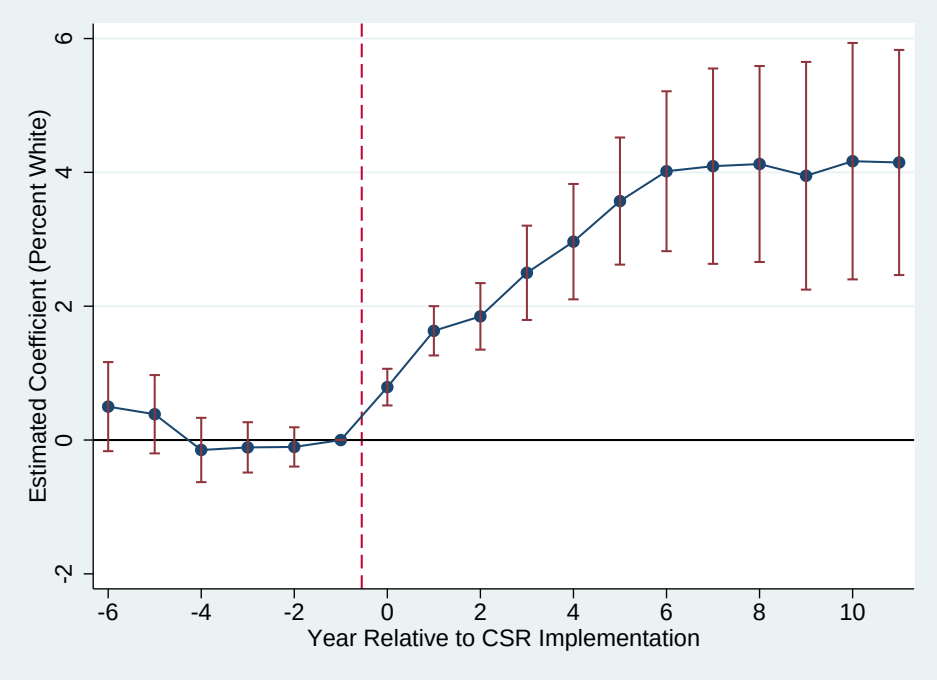
(b) Private School Entry Rates



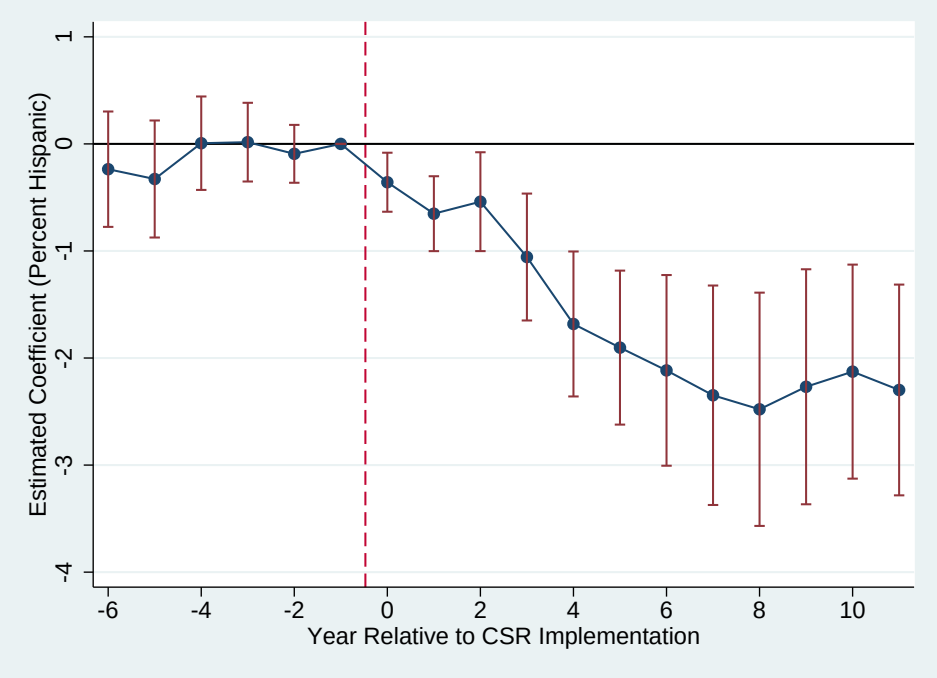
Notes: The above figures display the percent of private schools that have exited or entered the private school market within the last two years (data are available only every two years). The dashed vertical lines indicate the 1996-97 introduction of the CSR reform. Figures only include private schools that serve CSR grades – that is, if (on average) the school consists of twenty percent or more students in K-3 in the 1989-90 through 2009-10 school years.

Figure A.4: The Effect of CSR on Public School Composition by Year

(a) Percent White

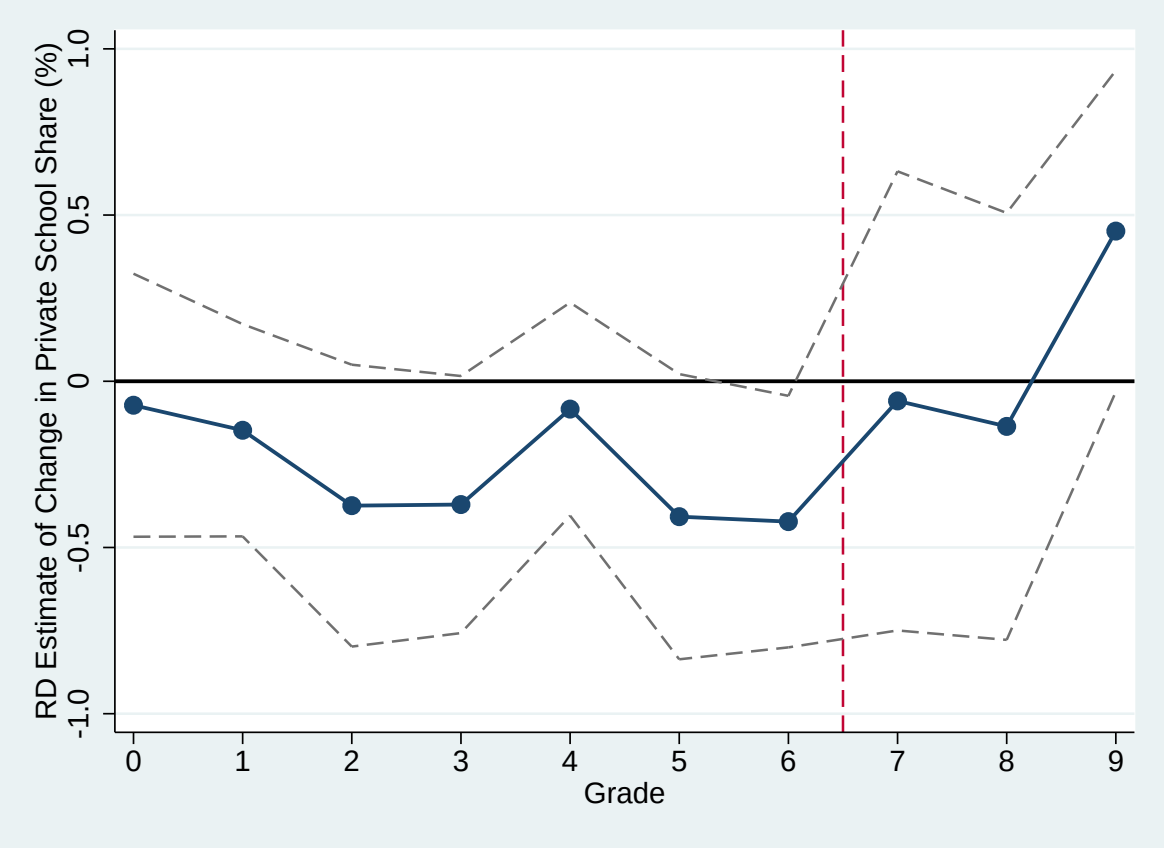


(b) Percent Hispanic



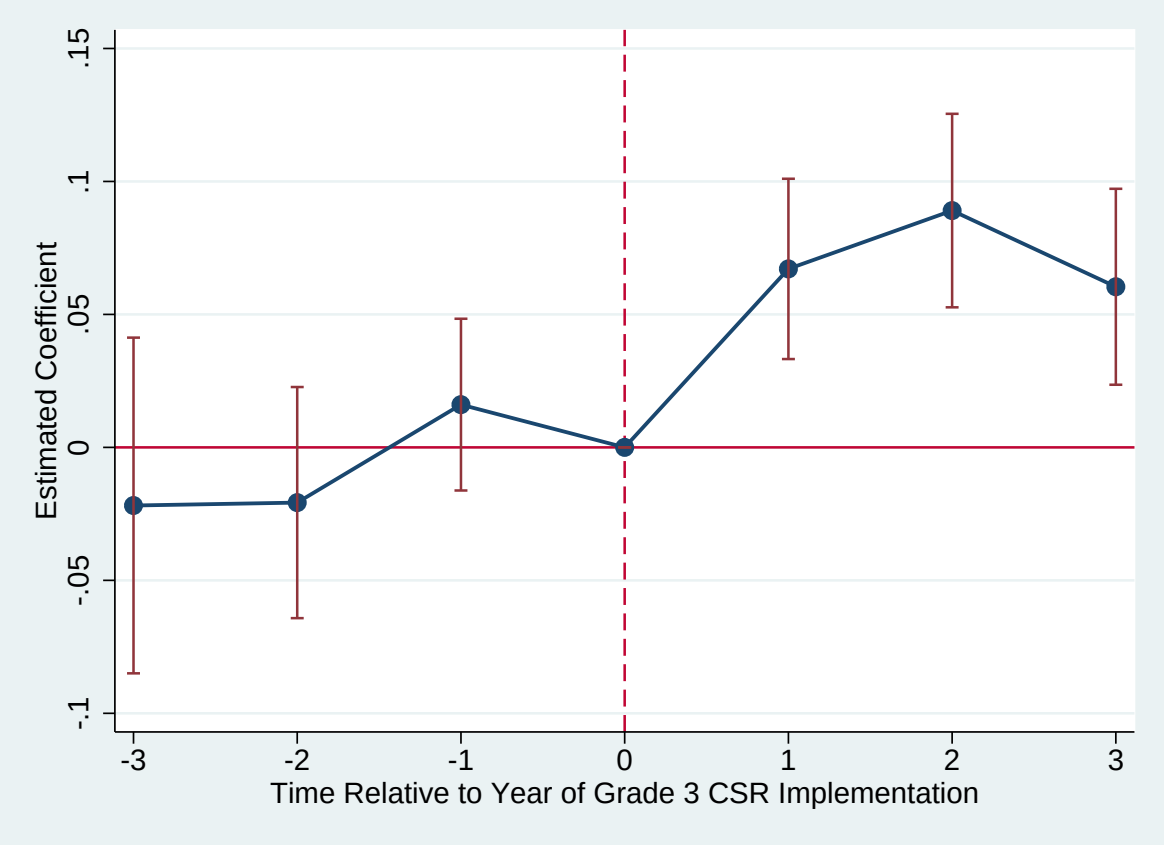
Notes: Figures show the estimated change in public school demographics by year using grade (CSR vs. non-CSR) and closeness to a private school (close vs. far) as the two layers of differencing. 'Close' is defined as any public school within $3km$ of a private school serving ten or more students in grades K-3, while all other public schools are categorized as being 'far.' The dashed vertical line represents the start of CSR implementation while the horizontal line indicates an estimate of zero. The estimated coefficient for the year prior to the start of CSR implementation is normalized to zero. Vertical bands represent 95% confidence intervals for each point estimate. Covariates and grade, year and school fixed effects are included. Standard errors are clustered at the district level.

Figure A.5: The Effect of CSR on Private School Share by Grade



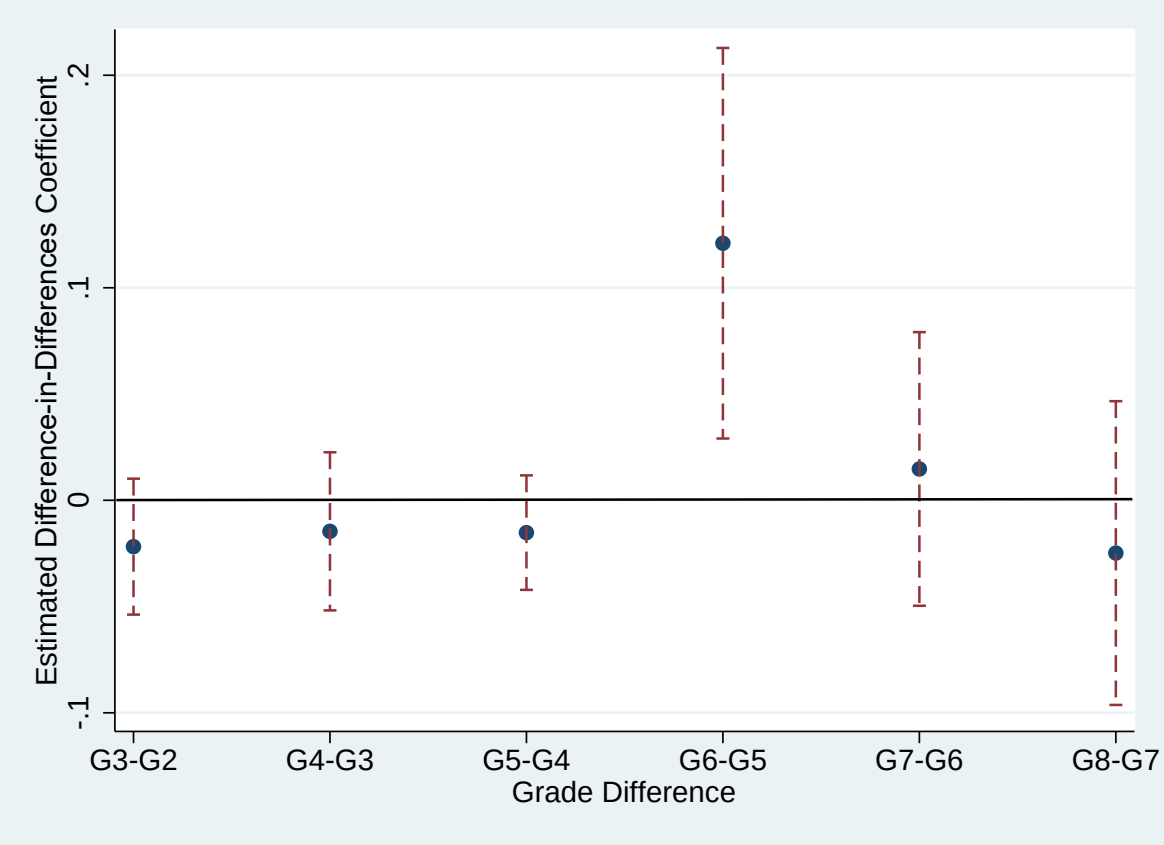
Notes: This figure shows the estimated effect of CSR on private school share for each grade using the RD design described in Section 4.3. The vertical line between sixth and seventh grade indicates the first grade where (almost) all students have transitioned to middle school from their elementary school. The horizontal line represents an estimate of zero. The kindergarten effect here represents a placebo test as kindergarten (grade '0') was not a CSR grade for the cohorts around the discontinuity. The effect for each grade is estimated using a local linear regression allowing for a different functional form on either side of the cutoff. District fixed effects and demographic controls are included in all regressions. The bandwidth used is three. Standards errors are clustered at the district level.

Figure A.6: The Effects of CSR on Test Scores by Event Time



Notes: The above figure shows the estimated effects of CSR by year on public school mathematics test scores using a difference-in-differences design (see footnote 59). In essence, this figure provides support the 'parallel trends' assumption by showing that there is no pre-trend for this difference-in-differences design, whose results are reported in Table A.9. The dashed vertical line represents the last year before CSR was implemented and is normalized to zero. The horizontal line represents an estimate of zero. Vertical bands represent 95% confidence intervals for each point estimate. Demographic controls and grade, year and school fixed effects are included. Standard errors are clustered at the school level.

Figure A.7: Reduced-Form Identification of Indirect Effect (Mathematics Scores)



Notes: This figure shows point estimates of a triple-differences regression using grade g vs. $g-1$, K6 vs. K5 schools and cohorts affected vs. unaffected by CSR as the three layers of differencing. The figure highlights the identification of the indirect effect in our structural model in a reduced-form way. Specifically, we expect that the only grade g vs. $g-1$ differences in mathematics test scores that should appear are for the grade 6 vs. grade 5 comparison, as that is when students induced into the public system due to CSR return to the private system due to the transition to middle school. The outcome variable is the mathematics test score, normalized by grade-year to have mean zero and standard deviation one. Demographic controls include student race, enrollment and enrollment squared. Vertical dashed bands represent 95% confidence intervals for each point estimate, while the horizontal line indicates an estimate of zero. Standard errors are clustered at the district level. These results are the same as those reported in column (2) of Table A.10.

Table A.1: Data Sources and Availability

Data	Observation Level (1)	Years Covered (2)	Number of Observations (3)	Data Source (4)
Data Type: California Department of Education Data (publicly available)				
Public School Enrollment Data (includes race)	School-Grade-Year	1990-91 to 2008-09	914,514[a]	www.cde.ca.gov/ds/sd/sd/filesenr.asp
Private School Enrollment Data[b]	District-Grade-Year	1990-91 to 2008-09	261,573	www.cde.ca.gov/ds/si/ps/index.asp
Public School ESL Data[c]	School-Grade-Year	1990-91 to 2008-09	914,514	www.cde.ca.gov/ds/sd/sd/fileselsch.asp
Public School Free or Reduced-Price Meal Data	School-Year	1990-91 to 2008-09	200,848	www.cde.ca.gov/ds/sh/cw/filesafdc.asp
CSR Implementation Data	School-Grade-Year (grades K-3 only)	1998-99 to 2003-04	130,011	www.cde.ca.gov/ds/si/ps/index.asp
Standardized Testing and Reporting Data	School-Grade-Year (grades 2-11 only)	1997-98 to 2001-02	231,129[d]	star.cde.ca.gov
Teacher Assignment and Demographic Data	School-Grade-Year	1997-98 to 2001-02	222,626[d]	www.cde.ca.gov/ds/sd/df/filesassign.asp
Teacher Demographic and Experience Data	School-Year	1994-95 to 2008-09	136,935	www.cde.ca.gov/ds/sd/df/filescertstaff.asp
Data Type: Other Data				
Private School Universe Survey (State-level)	State-Year (biannual)	1989-90 to 2009-10	561	nces.ed.gov/surveys/pss/
U.S. Population Data (State-level)	State-Year	1989-2009	1,122	seer.cancer.gov/popdata/download.html

Notes: All data can be aggregated to higher levels. For instance, 'school-grade-year' observations can be aggregated into 'district-grade-year' or 'school-year' observations.

[a] Only non-zero grade-level observations are included in this observation count.

[b] Private school enrollment data for 1990-91 through 1998-99 inclusive are not available on the CDE website. They were provided upon request by the CDE.

[c] California divides ESL students into English Learners and Fluent English Proficient. Since schools can alter students' ESL designations, we combine these two categories at the observation level into an ESL control, to avoid picking up any endogenous responses in ESL designations following CSR.

[d] Data are available up to 2008-09, but we only use observations from 1997-98 to 2001-02 due to the switch from the Stanford Achievement Test to the California Achievement Test in the 2002-03 academic year.

Table A.2: Mathematics Test Score Summary Statistics

School Year	Grade 2	Grade 3	Grade 4	Grade 5	Grade 6
1997-98	44.6	43.6	41.4	43.3	50.6
	(19.2)	(19.5)	(19.1)	(19.7)	(19.0)
1998-99	44.6	43.6	41.4	43.4	50.6
	(19.2)	(19.5)	(19.1)	(19.7)	(18.9)
1999-00	58.5	58.2	52.4	52.2	58.8
	(18.6)	(18.1)	(18.6)	(19.4)	(18.2)
2000-01	59.8	61.1	55.4	55.8	61.4
	(18.0)	(17.5)	(18.2)	(18.9)	(17.8)
2001-02	62.6	63.5	58.1	58.2	63.1
	(16.9)	(16.8)	(17.5)	(18.1)	(17.3)
Total Observations (School-Grade-Year)	33,044	33,209	32,678	32,111	16,498

Notes: Test scores are from the Stanford 9 test and report the mean percentile ranking of students relative to a nationally representative reference group. The increases in test scores from the 1998-99 school year to the 1999-00 school year were caused by the addition of several test items intended to cover material in California's content standards that were not previously addressed by the Stanford 9. This led causing California students to score higher relative to the norm-referencing group.

Table A.3: Triple-Differences Estimates of CSR on Private School Share

Outcome Variable: Private School Share (%)

	(1)	(2)	(3)	(4)
Treatment*Post*CSR	-1.34**	-1.34**	-1.35**	-1.29**
	(0.55)	(0.55)	(0.67)	(0.60)
Treatment*Post	0.13	0.24	0.12	-0.29
	(0.47)	(0.47)	(0.53)	(0.46)
Treatment*CSR	2.44**	2.47**	2.35*	2.86***
	(1.09)	(1.09)	(1.32)	(1.00)
Post*CSR	2.00***	1.97***	1.60**	1.20**
	(0.60)	(0.60)	(0.64)	(0.54)
Post	-2.32***	-1.54***	-1.27**	-0.63
	(0.52)	(0.57)	(0.59)	(0.49)
CSR	5.67**	5.65**	1.90	-
	(2.25)	(2.25)	(1.96)	
Treatment	0.00	-	-	-
	(1.00)			
Year/Grade FE	No	Yes	Yes	Yes
Demographic Controls	No	No	Yes	Yes
District FE	No	No	No	Yes
Number of Observations	192,848	192,848	161,967	161,967

Notes: This table shows results from the triple-differences regression described by equation (B.2) with varying levels of controls. Observations are at the district-grade-year level and cover the 1990-91 through 2008-09 school years. Demographic controls include student race, gender, English second language, enrollment and enrollment squared. The 'treatment' variable is omitted for columns (2)-(4) since it is collinear with the grade fixed effects, and CSR is omitted in column (4) as it is collinear with district fixed effects. All regressions are weighted by district-grade-year enrollment. Standard errors are clustered at the district level. ***,** and * denote significance at the 1%, 5% and 10% levels, respectively.

Table A.4: Triple-Differences Estimates of CSR on Private School Numbers

Outcome Variable: Private Schools per 1000 School-Aged Children

	D-in-D (CSR Schools) (1)	D-in-D (non-CSR Schools) (2)	Triple Differences (3)
Estimate	-0.070***	-0.011**	0.059***
	(0.016)	(0.005)	(0.015)
State and Year FE	Yes	Yes	Yes
Observations	561	561	1,122

Notes: This table shows results from difference-in-differences regressions using time (pre- vs. post-CSR) and state (California vs. rest-of-country) as the two layers of differencing and restricts to private schools that do primarily serve CSR grades in column (1) and private schools that do not in column (2). Schools are defined as primarily serving CSR grades if more than twenty percent of their student body is in grades K-3 in the 1995-96 school year. Column (3) then runs the triple-differences regression described by equation (B.3) by adding whether the private school primarily serves CSR grades as an additional layer of differencing. Observations are at the state-by-biennial year level and cover 1989-90 through 2009-10 school years. The number of school-aged children by state is measured as the number of 5-17 year old children in the state according to data given to the National Cancer Institute by the U.S. Census Bureau (available at https://seer.cancer.gov/popdata/download.html). Standard errors are clustered at the state level. *,** and *** denote significance at the 10%, 5% and 1% levels, respectively.

Table A.5: Triple-Differences Estimates of School Compositional Changes

Outcome Variable: Public School Student Demographic Compositions (%)

	Percent White	Percent Hispanic	Percent Black	Percent Asian
	(1)	(2)	(3)	(4)
Treatment*Post*$\mathbb{1}\{Buffer < 1.5\ km\}$	2.94***	-1.52***	-0.68***	0.03
	(0.61)	(0.37)	(0.18)	(0.21)
Treatment*Post*$\mathbb{1}\{Buffer < 3\ km\}$	2.92***	-1.47***	-0.71***	0.05
	(0.62)	(0.39)	(0.18)	(0.23)
Treatment*Post*$\mathbb{1}\{Buffer < 5\ km\}$	2.94***	-1.45***	-0.73***	0.03
	(0.62)	(0.39)	(0.18)	(0.23)
% Share in Private School (1997-98)	52.9	17.2	7.1	12.3
% Share in Public School (1997-98)	38.8	40.5	8.8	11.1
School/Grade/Year FE	Yes	Yes	Yes	Yes

Notes: This table shows results from the triple-differences regression using time (pre- vs. post-CSR), grades (CSR vs. non-CSR) and closeness to a private school ('near' vs. 'far') as the three layers of differencing as described in equation (C.1). $\mathbb{1}\{Buffer < x\ km\}$ is the distance from a private school that a public school must be to be considered 'treated'. Three alternative buffers are provided for robustness. Observations are at the school-grade-year level, and cover 1990-91 through 2008-09 school years. There are 914,514 observations. Enrollment and enrollment squared are included as controls. Private and public school demographic shares from the National Center for Education Statistics for the 1997-98 school year are provided in the penultimate two rows for reference. All regressions are weighted by school-grade-year enrollment and standard errors are clustered at the district level. ***,** and * denote significance at the 1%, 5% and 10% levels, respectively.

Table A.6: Average Class Sizes by Grade and Year

	School Year				
Grade	**1997-98**	**1998-99**	**1999-2000**	**2000-01**	**2001-02**
			Average Class Size		
Kindergarten	24.2	**21.0**	**19.9**	**19.6**	**19.5**
Grade 1	**19.2**	**19.2**	**19.2**	**19.2**	**19.2**
Grade 2	**19.4**	**19.2**	**19.1**	**19.0**	**19.0**
Grade 3	22.4	**20.1**	**19.6**	**19.4**	**19.3**
Grade 4	29.1	28.9	28.9	28.7	28.5
Grade 5	29.4	29.3	29.2	29.3	29.0

Notes: The numbers in the table represent average class sizes by grade and year. Grade-year combinations that were affected by CSR are in bold font. Since grade-level class sizes are not observed before 1997-98, grades 1 and 2 have no pre-CSR comparison because those grades implemented CSR during the 1996-97 and 1997-98 school years, respectively. Some pre-kindergarten classes are included in the kindergarten average class size calculation.

Table A.7: Triple-Differences Estimates of Compositional Changes by Grade

Outcome Variable: Public School Student Demographic Compositions (%)

	Percent White	Percent Hispanic	Percent Black	Percent Asian
	(1)	(2)	(3)	(4)
Kindergarten*Post*$\mathbb{1}\{Buffer < 3\ km\}$	3.18***	-1.57***	-0.79***	0.02
	(0.60)	(0.41)	(0.21)	(0.27)
Grade 1*Post*$\mathbb{1}\{Buffer < 3\ km\}$	2.77***	-1.32***	-0.78***	0.04
	(0.67)	(0.41)	(0.20)	(0.23)
Grade 2*Post*$\mathbb{1}\{Buffer < 3\ km\}$	2.69***	-1.32***	-0.67***	-0.01
	(0.65)	(0.40)	(0.18)	(0.22)
Grade 3*Post*$\mathbb{1}\{Buffer < 3\ km\}$	2.82***	-1.35***	-0.64***	-0.12
	(0.63)	(0.41)	(0.18)	(0.21)
School/Grade/Year FE	Yes	Yes	Yes	Yes

Notes: This table shows results from a variant of the triple-differences regression described in equation (C.1). Specifically, a dummy for each CSR grade is used, individually, to see if the composition effect is driven by certain grades. Other CSR grades are not included in the regression. Therefore, the triple-differences regression in the first row (represented by coefficient on Kindergarten*Post*$\mathbb{1}\{Buffer < 3\ km\}$) uses time (pre- vs. post-CSR), grades (Kindergarten vs. non-CSR) and closeness to a private school ('near' vs. 'far') as the three layers of differencing with data for grades 1-3 omitted. Observations are at the school-grade-year level, and cover 1990-91 through 2008-09 school years. Enrollment and enrollment squared are included as controls. All regressions are weighted by school-grade-year level enrollment and standard errors are clustered at the district level. ***,** and * denote significance at the 1%, 5% and 10% levels, respectively.

Table A.8: Parallel Trends in Untreated Cohorts

Outcome Variable: Mathematics Test Scores (Percentile Rank)			
	1997-98	1998-99	1999-00
	(1)	(2)	(3)
Panel A. *Trends in Grade 5 Relative to Grade 4*			
Grade 5	43.60	43.53	-
Grade 4	41.62	41.60	-
Difference	1.99	1.93	-
(Grade 5 - Grade 4)	(0.41)	(0.41)	
Panel B. *Trends in Grade 6 Relative to Grade 5*			
Grade 6	51.61	52.02	60.21
Grade 5	43.60	43.53	52.36
Difference	8.01	8.49	7.86
(Grade 6 - Grade 5)	(0.48)	(0.40)	(0.49)

Notes: This table compares untreated adjacent grades for 'never treated' cohorts to bolster the argument that differences between the grade 4 versus grade 3 comparison identifying the direct effect would be similar in the absence of CSR. Given test scores become available in 1997-98, we have two pre-treatment years for the grade 5 relative to grade 4 comparison (grade 4 was first treated in 1999-00) and three pre-treatment years for the grade 6 relative to grade 5 comparison (grade 5 was first treated in 2000-01). Standard errors of the differences are reported in parentheses. The cells report mathematics test scores in percentile rank relative to a national norming sample, where one percentile rank roughly equates to 0.05σ in the distribution of school-grade level test scores.

Table A.9: Difference-in-Differences Estimates of CSR on Test Scores

Outcome Variable: Mathematics Scores (σ)

	(1)	(2)	(3)
Treat*Post	0.105***	0.070***	0.067***
	(0.016)	(0.011)	(0.011)
Post	0.137***	0.022**	0.028***
	(0.032)	(0.010)	(0.010)
Treat	-0.094***	-0.065***	-0.066***
	(0.015)	(0.011)	(0.011)
Grade/Year/School FE	No	Yes	Yes
Demographic Controls	No	No	Yes
Number of Observations	207,926	207,926	207,523

Notes: This table show the results of the difference-in-differences regression defined in footnote 59, which compares mathematics test scores in CSR grades (2-3) and non CSR grades (4-8) before and after CSR was implemented. Observations are at the school-grade-year level, and cover the 1997-98 through 2003-04 school years. Test scores are normalized by grade-year to have mean zero and standard deviation one. Demographic controls include student race, enrollment and enrollment squared. Standard errors are clustered at the school level. ***,** and * denote significance at the 1%, 5% and 10% levels, respectively.

Table A.10: Triple-Differences Estimates of CSR General Equilibrium Effects on Test Scores

Outcome Variable: Mathematics Scores (σ)

	(1)	(2)	(3)
A. Grade 6 versus Grade 5 (Coefficient of Interest)			
$\Phi_{K6-K5,6-5,post-pre}$	0.128**	0.112**	0.099**
	(0.054)	(0.055)	(0.050)
A. Other Grade Differences (Placebo Tests)			
$\Phi_{K6-K5,7-6,post-pre}$	0.041	0.026	0.028
	(0.032)	(0.028)	(0.026)
$\Phi_{K6-K5,5-4,post-pre}$	-0.017	-0.017	-0.018
	(0.014)	(0.014)	(0.015)
$\Phi_{K6-K5,4-3,post-pre}$	-0.019	-0.018	-0.018
	(0.017)	(0.018)	(0.018)
$\Phi_{K6-K5,3-2,post-pre}$	-0.003	-0.003	-0.004
	(0.016)	(0.016)	(0.016)
Grade/Year/School FE	No	Yes	Yes
Demographic Controls	No	No	Yes

Notes: Observations are at the school-grade-year level, and cover the 1997-98 through 2008-09 school years. Test scores are normalized by grade-year to have mean zero and standard deviation one. Demographic controls include student race, enrollment and enrollment squared. Standard errors are clustered at the district level. ***,** and * denote significance at the 1%, 5% and 10% levels, respectively.

Table A.11: School Statistics by Grade Span Configuration

Grade Span	Number of Schools	% of Schools	% Implementing CSR in First Year
	(1)	(2)	(3)
K-5	2183	44.3	95.9
K-6	1954	39.6	92.5
K-8	455	9.2	90.1
K-12	49	1.0	48.3
Other	289	5.9	90.3
Total	4,930	100	93.2

Notes: This table shows the number and percentage of schools by grade span serving at least two K-3 grades in the 1998-99 school year. The most common 'Other' configuration schools are K-3 or K-4 schools. Given the data limitation that CSR implementation is first observed in 1998-99, the percentage implementing CSR in the first year is calculated as the proportion of schools that had implemented CSR in either Kindergarten or third grade in the 1998-99 school year.

Table A.12: Percentage of Inexperienced Teachers

	Year					
Grade	1997-98	1998-99	1999-00	2000-01	2001-02	2002-03
2	27.3	26.7	21.6	18.8	17.0	15.2
3	26.8	26.3	22.0	17.9	16.4	14.0
4	26.9	33.1	32.9	30.1	27.6	24.5
5	24.0	27.8	28.8	28.4	25.7	22.5
6	23.5	26.7	27.4	26.5	27.0	23.2

Notes: Percent inexperienced is defined as the fraction of full time equivalent teachers with less than three years of experience teaching in the state of California.

Table A.13: Estimates of Teacher Quality

Outcome Variable: Mathematics Test Scores		
	CSR	non-CSR
	(1)	(2)
$Q_{CSR/non,01-02}$	1.123***	-0.089***
	(0.057)	(0.004)
$Q_{CSR/non,00-01}$	0.929***	-0.361***
	(0.047)	(0.018)
$Q_{CSR/non,99-00}$	0.520***	-0.643***
	(0.026)	(0.032)
$Q_{CSR/non,98-99}$	0.041***	-0.678***
	(0.002)	(0.034)
$Q_{CSR/non,99-00,K5}$	0.997***	-1.132***
	(0.078)	(0.089)
$Q_{CSR/non,99-00,K6}$	0.632***	-0.673***
	(0.050)	(0.053)

Notes: This table shows estimates of teacher quality. Observations are at the school-grade-year level, and cover the 1997-98 through 2001-02 school years. Mathematics test scores are shown in percentile ranks relative to a national norming sample, where one percentile rank roughly equates to 0.05σ in the distribution of school-grade level test scores. Standard errors are computed using the delta method and are clustered at the school level. ***,** and * denote significance at the 1%, 5% and 10% levels, respectively.