Class-size Reduction Policies and the Quality of Entering Teachers

Steven Dieterle [*]
University of Edinburgh
steven.dieterle@ed.ac.uk

June 19, 2015

Abstract

Class-size reduction (CSR) policies have typically failed to produce large achievement gains. One common explanation is that CSR forces schools to hire low-quality teachers. Prior studies of this hypothesis have been hindered by poor data. Using different data, we find that hiring quality did fall with state-wide CSR. However, this drop was temporary due to attrition by the lowest performers. Furthermore, the drop was similar for schools classified as treated and control for prior evaluations of CSR. Therefore, differences in the quality of incoming teachers cannot explain the estimated performance of CSR. This is consistent with hiring spillovers in connected markets.

Keywords: Teacher Labor Markets, Teacher Quality, Class Size

[*] I would like to thank Gary Solon, Todd Elder, Jeff Wooldridge, Cassie Guarino, Steven Haider, Mike Elsby, Otavio Bartalotti, and Quentin Brummet, as well as seminar participants at Michigan State University, Notre Dame University, the University of Stirling, and the University of Aberdeen for helpful comments. I would also like to thank two anonymous referees and the editor, Helena Skyt Nielsen. All errors are the responsibility of the author. The work reported here was supported in part by a Pre-Doctoral Training Grant from the Institute of Education Sciences, U.S. Department of Education (Award # R305B090011) to Michigan State University.

1 Introduction

The potential for student achievement gains from smaller classes has been well documented in experimental and quasi-experimental research (Krueger 1999; Krueger & Whitmore 2001; Angrist & Lavy 1999). For instance, analysis of the Tennessee STAR class-size experiment has found that being randomly assigned to a small (13-17 students) class as opposed to a larger class (22-25 students) in early elementary school has both short and long run effects on students. In particular, students in smaller classes had test scores roughly one-fifth of a standard deviation better on average (Krueger 1999), better long run educational attainment (Krueger & Whitmore 2001), and better labor market outcomes (Chetty et al. 2011). As of 2005, this potential led to the adoption of class-size reduction (CSR) measures in thirty-two states (Council for Education Policy, Research and Improvement (CEPRI) 2005).

To date, studies of CSR policies find only mixed evidence of achievement effects, with estimates consistently falling short of what might be expected from the experimental research (Bohrnstedt & Stecher 2002, Chingos 2012).[1] Due to the high costs of implementation, $21 billion over nine years in Florida (Florida Department of Education) and $1.5 billion a year in California (Bohrnstedt & Stecher 1999), the efficacy of CSR policies has been called into question. One common explanation for the under performance of CSR is that it forces schools to hire and retain teachers of lower quality in order to meet the class-size requirements (Stecher & Bohrnstedt 2000; Imazeki n.d.; Buckingham 2003; CEPRI 2005, Chingos 2012). The gains from having smaller classes are thought to be offset by having teachers of lower quality in the classroom.

Previous studies of this hypothesis have focused on evidence from California's CSR program (Kane & Staiger 2005, Jepsen & Rivkin 2009). However,

[1] Note that not all experimental and quasi-experimental studies find significant class-size effects (Hoxby 2000). A recent paper by Rockoff (2009) discusses the results of several class-size experiments from the beginning of the twentieth century and concludes that the balance of these early class-size experiments suggest there was little achievement benefit to attending smaller classes. This conclusion comes with several caveats. Most importantly, it seems plausible that changes in the educational environment since the early twentieth century may have changed the role of class size in affecting achievement.

studies of California CSR are limited by the available data. Chief among these limitations is a lack of linked student-teacher test score data until several years after CSR's introduction (Kane & Staiger 2005). Due to differential teacher attrition and human capital accumulation, this leaves the short-run implications of CSR induced hiring unanswered. Furthermore, the linked data that are available covers only a single district, prohibiting an analysis of heterogeneity across districts or the potential for across-district hiring spillovers. While school aggregated data is available for the period around the introduction of CSR, these data still do not include any pre-policy test measures. Identification using the school average data also relies on observed teacher characteristics in order to estimate changes in teacher quality as the data do not identify new teachers or link students to specific teachers (Jepsen & Rivkin 2009). However, much of the education production function literature finds that these characteristics play only a small role in explaining the variation in student achievement (Goldhaber 2008).

Using administrative data on individual students and teachers in grades four through six from an anonymous state (subsequently referred to as State X)[2] before and after the introduction of a state-wide CSR program, this paper explores the teacher quality hypothesis in detail, while overcoming the limitations of the prior work. As a starting point, we consider whether there is any evidence that a CSR-induced decline in teacher quality can explain the lack of an *estimated* effect from prior quasi-experimental evaluations of CSR performance.

We find little evidence to support the idea that the small CSR effects estimated using treatment-control comparisons are due to the quality of incoming teachers. Comparing schools categorized as treated (those for which CSR was binding) and control in prior quasi-experimental studies of CSR before and after the introduction of the policy, we find only a very small difference ($\approx 0.15\%$ of a test score standard deviation) in average student performance attributable to the quality of hired teachers. In fact, schools classified as treated experi-

[2] The State X Department of Education has requested the state be kept anonymous for all publications and presentations as a condition of data access and use.

ence a slightly smaller drop in achievement attributable to newly hired teachers than those in the control group. This difference is of the opposite sign needed to support the teacher quality hypothesis. The estimates account for both the underlying quality of hiring cohorts and the potential short-run effect of hiring more teachers with less experience. It is important to note that these results are quite robust to several estimation approaches that account for many possible confounding influences including state-wide policy changes or general state-wide trends in the quality of hired teachers.

Interestingly, the small difference between treated and control groups masks a flat profile in cohort value-added before CSR followed by a sudden decline in the quality of teacher hiring cohorts in *both* treatment and control schools with the introduction of CSR. While the quality of cohorts may not explain the small treatment effect estimates of CSR, due to the strong possibility of treatment spillovers in this setting it may still be the case that teacher quality and student performance did suffer from CSR, a possibility missed by any treatment-control evaluation of CSR. Namely, with treatment and control schools operating in the same labor markets, the increase in teacher demand from the introduction of CSR may have reduced the quality of new hires even in the control schools. That is, schools not directly affected by CSR may have nevertheless hired lower quality teachers due to CSR as potential candidates were hired by schools forced to reduce class size.

While the general equilibrium nature of these potential hiring spillovers makes it difficult to completely rule out other possible explanations for the sudden decline in value-added associated with the CSR induced hiring increase, it is a potentially important effect of CSR that would go unnoticed in treatment-control comparisons. Further, we do provide some suggestive evidence that the drop in quality with increased hiring was not likely driven by changes in certification policies or in the financial attractiveness of teaching in State X that could possible alter the selection into teaching over this time period.

To examine and quantify the possible state-wide CSR hiring effects, we trace the evolution of cohort mathematics value-added over time for three pre- and five post-policy hiring cohorts.[3] The estimates of cohort performance indicate a modest reduction in the average quality of both newly hired teachers and teachers who are retained after their first year. In terms of student achievement, the estimated conditional mean performance of the larger (up to 62% larger) post-CSR hiring cohorts ranges from 0.33% to 2.55% test score standard deviations lower than the smaller pre-CSR cohorts in each cohort's first year. This difference is equivalent to 10-15% of the standard deviation in teacher quality found in our sample.

Furthermore, the impact on individual students assigned to the marginal teachers may be quite large. Back of the envelope calculations based on individual teacher value-added suggest that more students, roughly 7% of all students assigned to a new cohort teacher, were assigned a teacher in the lowest quintile of the value-added distribution during CSR compared to before CSR. Given the large differences in mean value-added by quintile of between 25% and 73% of a test score standard deviation, this represents a potentially large effect for this subset of affected students.

However, the differences in cohort performance only persist partially over time as the composition of each cohort changes, with the differences in pre- and post-CSR second year cohort effects ranging from 1.09% to 1.98% standard deviations. However, there is evidence that further attrition leads to negligible differences among the remaining teachers after three to four years, implying a very small long-run CSR hiring effect on achievement. Importantly, the short-run CSR hiring effects identified here were missed completely by prior studies.

The results are informative beyond providing a better understanding of CSR programs. The results help fill a gap in the prior literature on the quality elasticity of teacher supply. Namely, the intervention studied here provides a rare opportunity to observe a substantial increase in the number of teachers hired for the same schools in a short time period. This sort of variation is preferred to relying on cross-sectional or longer run differences in teacher hiring

[3] Similar results obtained using reading test scores are presented in an appendix. The decision to focus on mathematics scores only was made for the sake of brevity.

to identify this elasticity. An understanding of the nature of the underlying teacher labor supply is useful for predicting the impact of any intervention that results in a sudden change in teacher demand. For instance, short-run increases in teacher demand associated with retirement buyout plans or changes in curriculum are often met with concerns over the quality of the new teachers hired (Center for Local State and Urban Policy 2010). Additionally, recent papers have simulated the achievement effects of value-added based retention policies, the results of which depend critically on the assumptions regarding the quality elasticity of teacher supply (Goldhaber & Theobald 2011, Boyd et al. 2011). The results found here are informative in predicting the fall in quality associated with such policies.

The paper proceeds as follows: section 2 discusses the data used; section 3 discusses the institutional details of the policy and concurrent teacher labor market conditions; section 4 gives the empirical strategy used and provides the baseline results and sensitivity checks; section 5 presents further analysis assessing the implications of the baseline estimates for CSR policy performance, tracing out the long run hiring effects, and characterizing the magnitude of the effects for students; finally, section 6 concludes.

2 Data

The data used for the following analysis will be a combination of restricted-use state administrative data and State X's published class-size averages. The extract of the administrative data available for this study links students in grades one through six to teachers and schools from the 2000-2001 to the 2007-2008 school year. Importantly, the students are linked directly to their math teacher. In other prominent administrative data sets, the student/teacher match is less clean with students linked to all teachers at the grade level or to end-of-year exam proctors. In addition to basic student demographics, the data include mathematics scores for State X's criterion-referenced high-stakes test for students from third to sixth grade. These test score data enable the estimation of teacher value-added for teachers in grades four through six over a seven-year period starting with the 2001-2002 school year.

The data track teachers over the same period as students and include an additional year, beginning with the 1999-2000 school year. This allows teachers to be followed as long as they stay in the state's elementary school education system. For instance, it is possible to identify when teachers enter or exit the public elementary school system over time. Given that CSR began in the 2003-2004 school year, this allows us to identify three pre-policy hiring cohorts, five post policy cohorts, as well as a set of "baseline" teachers hired four or more years before CSR. The teacher information includes relevant variables such as a teacher's experience and degree level. The experience measure used is the sum of four separate categories that are recorded for each teacher capturing all prior experience in public and private schools both within State X and in other states. This encompassing experience measure will be important when distinguishing between teacher quality and experience effects due to the CSR-induced hiring.

Finally, State X has made each district/school's average class size publicly available since the beginning of the CSR program. These class-size averages allow for the identification of districts and schools that needed to reduce class size in order to stay compliant. Importantly, this allows us to match the categorization of treatment and control groups used in prior prior quasi-experimental CSR policy estimates (Chingos 2012). Descriptive statistics for the key variables used in this study are presented in Appendix Table 1. Notably, nearly 70% of the student-year observations in the data are linked to a teacher observed entering at some point in the sample period allowing for comparisons across cohorts.

3 CSR & the Teacher Labor Market in State X

3.1 Institutional Details

In November of 2002, State X voters approved a constitutional amendment that created a state wide CSR program. The program was set to begin in the 2003-2004 school year. Separate class-size maximums were set for different grade levels, as shown in Table 1. The law established per-pupil allocations from the state government for each year a district or school was in compliance. There is anecdotal evidence from board of education meeting transcripts that this was not enough to cover the full costs of CSR for some districts. This anecdote suggests that a reallocation of other resources may partially explain CSR performance. This possibility will be explored in the results section.

The law allowed for a gradual phase-in of the new class sizes. A district or school was in compliance if it had lowered the average class size by two students from the previous year or if it was already below the maximum. For the first three years of the program, the compliance was based on the district average, while the next three years it was based on a school-level average. Non-compliance by districts or schools initially resulted in a portion of the CSR allocation being directed toward capital outlays aimed at reducing class size. Beginning in the third year of the program, the threatened sanctions for non-compliance became more severe. According to the law, districts not in compliance were to be forced to implement one of the following four policies: having year-round schools, having double sessions in schools, changing school attendance zones, or altering the use of instructional staff.

As seen in Table 1, the new maximums were binding for most districts at implementation with only 12% and 42% of districts below the required average class size in kindergarten through third grade and fourth grade through eighth grade, respectively. With district-level average class size dropping from 23 to 16 for the earliest grades and from 24 to 19 in the middle grades, it is clear that the program did achieve the stated goal of reducing class size.

Table 1: CSR in State X

Grades	*Maximum*	*Percent Below Max Yr 1*	*Average CS Yr 1*	*Average CS Yr 8*
KG-G3	18	12%	23	16
G4-G8	22	42%	24	19
G9-G12	25	91%	24	22

3.2 Market for Teachers in State X During CSR

Before analyzing the achievement outcomes associated with CSR and the subsequent teacher hiring in State X, we consider the general state of the teacher labor market, as well as factors that may have led to changes in the supply or demand for teachers over the same time period. This analysis is important for interpreting the results that follow and helps to tie the current work to the previous CSR literature on changes in the teacher workforce. We begin with a discussion of trends in teacher numbers and characteristics over the introduction of CSR.

Figure 1 displays trends in the stock and flow in the number of teachers, percent with an advanced degree, average experience, and percent with three or fewer years of experience. Here, we focus on teachers teaching a core course (those that fall under CSR requirements) in grades four through six (those for which value-added estimation is possible with our data). Recall that the data follow all first through sixth grade teachers in public schools in State X. Therefore, a teacher will be considered part of the flow if they are new to teaching, returning to teaching, transferring from a public middle or high school, moving from a private school within the state, or moving from a public or private school in another state.

In panel A, we see a steady rise in teacher numbers over the introduction of CSR from under 19,500 before CSR to nearly 24,500 after five years. This rise is driven in part by an increase in the number of entering teachers from a pre-CSR average of roughly 4,600 each year to 6,100 during district-level enforcement and 6,700 during school-level enforcement of CSR. We also see that the percentage with an advanced degree among both the stock and inflow falls with the introduction of CSR and the change to school-level enforcement, while

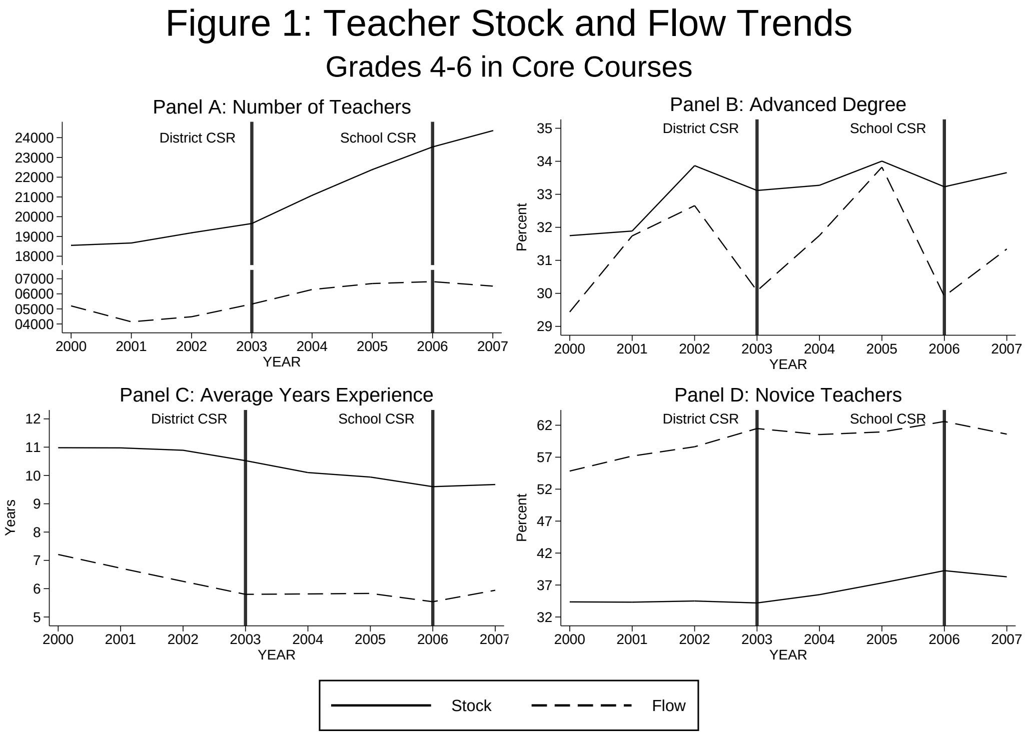


Figure 1: Teacher Stock and Flow Trends

increasing in the other years.[4] Average experience of all teachers drops from a pre-CSR level of roughly eleven years to nearly 9.5 years by the introduction of school-level enforcement four years later. Not surprisingly, the percentage of teachers considered novices, with three or fewer years of experience, also increased over the implementation of CSR.[5]

Note in panel C, we see a decline in average experience before the introduction of CSR. This drop is unlikely to be due to hiring in anticipation of CSR. Importantly, the amendment was voted on and passed in November of 2002, after the majority of hiring for the 2002-03 school year would have been completed. One may worry, however, that this signals a decline in teacher quality even before the policy is introduced. Importantly, in section 5 we will show that there is no pre-reform trend in cohort value-added, the main measure of quality used in this paper. Given the weaker connection between teacher observables and student achievement in the literature, it is the value-added trends that are more important as they will capture the unobservable factors that contribute to student achievement. That said, in the analysis that follows, we will explicitly control for teacher experience. If experience is a proxy for other unobserved factors, then this is accounted for and would alter the interpretation of the estimated experience effects and not the cohort effects. If, however, the concern is that the slight drop in experience beforehand was a signal that schools would be forced to move along several margins, including those not typically correlated with experience, to hire lower quality cohorts regardless of increased hiring, then our estimated cohort effects are an upper bound on the magnitude of CSR induced hiring effects on student achievement. Of course, this story requires the response on these other margins to

[4] For Advanced Degrees, this occurs during the adjustment to the Federal No Child Left Behind Act (NCLB) of 2001. NCLB placed more emphasis on having teachers with advanced degrees. As there is a lag for degree completion and it represents a discrete change in status for individual teachers, we might expect this to be less smooth than measures of experience that evolve more-or-less continuously over time for individual teachers. Schools may also be more willing to make trade-offs on the degree level dimension as teachers can be hired with the expectation that they will complete their masters in a specified time period.

[5] A more detailed discussion of changes in the observable characteristics of teachers in State X over CSR implementation based on publicly available school-level data is available upon request.

lag behind the experience drop to explain why there is no pre-policy drop in cohort value-added. We will return to this in discussing the results.

While this descriptive analysis has established a clear link between the timing of the CSR policy and both an increase in hiring and a drop in average experience of teachers, there are other concurrent factors worth mentioning. In terms of the demand for teachers, State X faced a growing student population that, irrespective of CSR, would require additional teachers. Soon after CSR adoption, the state projected the hiring needs across all grades and subjects from CSR and student enrollment growth, as shown in Table 2A. Hiring needs driven by enrollment growth were projected to be fairly steady, at just over 3,000 each year. At the change to school-level enforcement in 2006-07, the number of new teachers needed due to CSR was projected to be nearly three times that from enrollment growth. The difference for the grades studied here is likely to be even more stark, as the numbers in Table 2A include high school grades that were relatively unaffected by CSR.

In Table 2B, we show enrollment numbers for grades 4 through 6 (the grades studied here) in State X from the 1996-97 school year until 2006-07. While we see steady growth of around 2.5% several years before CSR, there is actually a decline in enrollment just before CSR and much smaller growth, around 0.25%, for the first few years of CSR. While the underlying growth of the student population certainly implies that the stock of teachers was likely to grow regardless of CSR, due to the relatively flat profile for projected enrollment growth based hiring in Table 2A and the slowdown in growth for the grades studied here in Table 2B, it is likely that the sudden increase in the number of teachers hired shown in Figure 1 was in fact largely due to CSR.[6] In the analysis that follows, it is best to think of the results coming from a situation where CSR has been implemented in a state of growing enrollment and that CSR policies implemented in times of falling or roughly stable student numbers may lead to different results. However, it is important to note that rising student numbers is the reality in many cases and, as such, is not unique

[6] Note that the trend in actual hiring may have been smoother than the projected numbers due to preemptive hiring.

to State X.

Table 2A: Projected Hiring in State X

	Year			
Hiring Need	*2004-05*	*2005-06*	*2006-07*	*2007-08*
CSR	4,324	2,378	11,821	974
Enrollment Growth	3,297	3,024	3,134	3,451

Table 2B: Grade 4-6 Enrollment in State X

CSR	*Year*	*G4-G6 Enrollment*	*Percentage Change*
None	1996-97	529815	
None	1997-98	539832	1.89%
None	1998-99	552636	2.37%
None	1999-00	567904	2.76%
None	2000-01	583434	2.73%
None	2001-02	597991	2.50%
None	2002-03	585174	-2.14%
District	2003-04	586840	0.28%
District	2004-05	588082	0.21%
District	2005-06	589620	0.26%
School	2006-07	599815	1.73%

Over this time period, the state commonly recruited teachers from other states to fill teaching needs. If out-of-state teachers are less familiar with the curriculum and the marginal teachers hired due to CSR were from out-of-state, any fall in teacher quality may partially reflect this. Once more, this does not invalidate the results to follow, as such a strategy may be pursued by any state facing an increase in teacher demand. Simply put, hiring more out-of-state teachers is one of the margins schools can move along when faced with CSR. Nevertheless, the administrative data can be used to help assess the importance of this hypothesis for interpreting the results. While the data do not include indicators for where a teacher completed their initial educator training, separate experience measures are recorded for time spent in State X and in other states. Recall from Panel C of Figure 1 that many entering teachers in our data have some previous experience, therefore we can look at entrants separately by the type of experience.

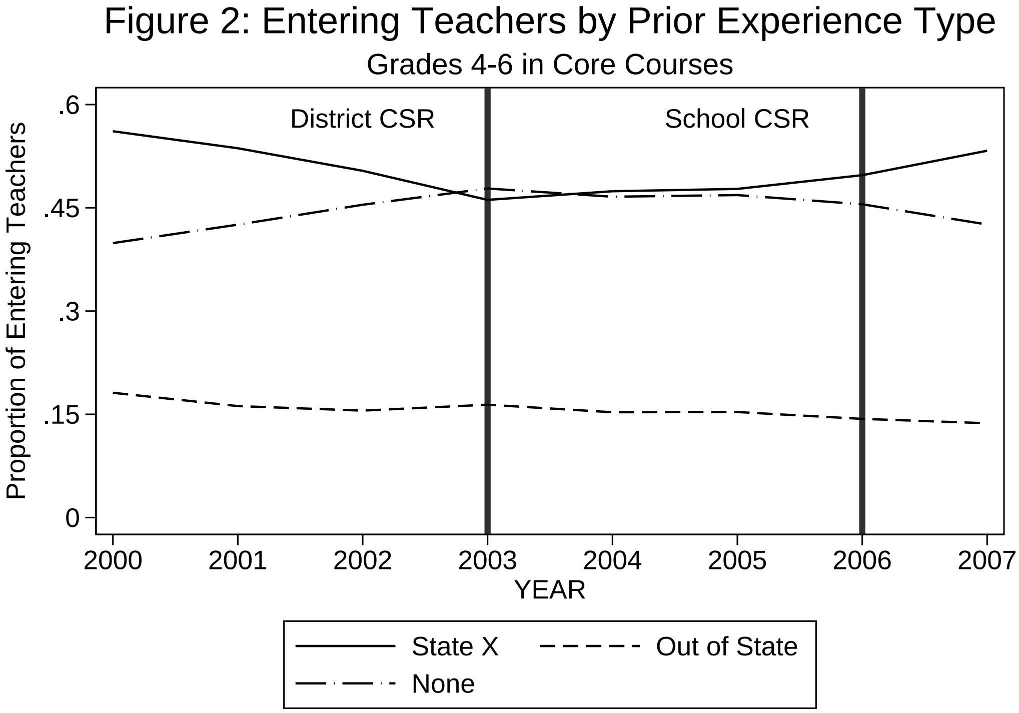


Figure 2: Entering Teachers by Prior Experience Type

Figure 2 plots the proportion of entering teachers in our sample that have no experience, experience in State X, and experience outside of State X.[7] We see the proportion of entrants with out-of-state experience stay roughly level at about 15%. We do see a rise in the proportion with no experience and a fall in those with experience in the state. The result is an eroding of a pre-policy gap of nearly 15 percentage points in favor of hiring teachers with prior in-state experience. This gap begins to reappear after the 2005-06 school year.

This analysis cannot capture changes in the composition of newly hired teachers without prior experience. While complete records covering this period are not available, one report from the state suggests that of the newly certified teachers whose certification was based on completion of an approved preparation program, roughly a quarter were from an out of state program in the first year of CSR, 2003-04, and another report puts the number at 29% the following year. As the majority of new hires entered with either prior experience in State X or were trained in State X, an increase in hiring out-of-state teachers can play only a small role in interpreting the main results of this paper.

While other changes in demand serve to inform the interpretation of the main analysis of this paper, it is concurrent changes in teacher supply that may directly affect the performance of hiring cohorts that pose the biggest threat to validity. In particular, we are concerned with potential changes in the selection into the teaching profession in State X, as well as changes in the training received by new cohorts. To be clear, these concerns are less important for our ability to assess the teacher quality hypothesis within the quasi-experimental setup used to estimate CSR policy effects. As will become apparent later, any state-wide trends in hiring cohort quality will be explicitly accounted for. However, when we allow for hiring effects to "spillover" to schools not directly under CSR pressure, any concurrent changes in the selection into or training of new cohorts of teachers will affect our ability to attribute changes in quality

[7] Note that some teachers identified as entering the data will have both in- and out-of-state experience so the sum across categories in each year can be greater than one.

to the CSR induced hiring increase.

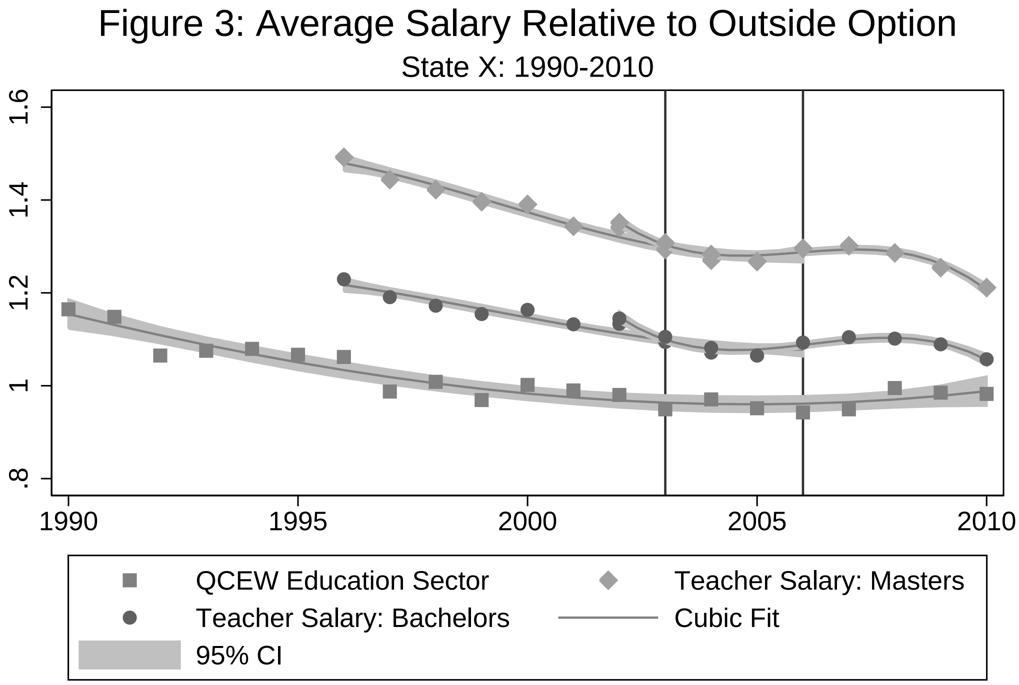
One concern would be that general labor market changes over time may alter the choice of entry into teaching. In particular, we consider the financial attractiveness of the teaching profession relative to alternative occupations in State X. Following Feng (2009), we use the Quarterly Census of Employment and Wages (QCEW) to calculate the average annual salary in sectors that teachers are most likely to enter upon leaving teaching.[8] Feng finds that the average salary in these sectors is predictive of teacher exit in State X. Under the assumption that these opportunity salaries are also important when making an initial occupation or college major choice, we document the long run changes in the salary of teachers relative to these outside options in State X. We would be particularly concerned with a sudden change in the relative attractiveness of teaching that coincides with either the introduction of CSR or a few years earlier when new teachers hired during CSR made degree choices.

Figure 3 presents the evolution of the ratio of average teacher salary to the average opportunity sector salaries. We use the QCEW measured average annual salary in the outside option sectors as the denominator and consider three different measures of teacher salaries as the numerator. We use the average annual salary for the Education Sector from the QCEW and the average salary for teachers with either a bachelors or masters degree as reported by State X.[9] For all three measures, we see a general decline in the relative attractiveness of teaching. However, there is no evidence of a sudden, large change in any of the salary ratios either just before or at the introduction of CSR.[10] Obviously, this is only suggestive evidence that there were no concurrent changes in the labor

[8] We use the same set of sectors as Feng: Retail Trade, Information, Finance and Insurance, Services, and Public Administration. Feng selects these sectors based on survey responses of exiting teachers in the Schools and Staffing Survey.

[9] Note that State X has published average salaries for both degree levels in two separate reports: one covering the period from 1996 to 2006 (the beginning of School-level CSR enforcement) and one starting in 2002 to the present. The two series differ slightly for the period of overlap, so we plot both. The QCEW series has the advantage of covering a longer period with no change in reporting, while the State X reports are a better reflection of what a teacher might expect, rather than the average for the entire Education sector in State X.

[10] Formal tests at CSR introduction or up to four years prior do not reject the null of no jump in relative salary.



Sources: Quarterly Census of Employment and Wages; State X School Profile Reports

Figure 3: Average Salary Relative to Outside Option

market that would alter the composition of entry cohorts. Two important caveats are that this does not include other forms of non-salary compensation and that the observed average salaries are the equilibrium outcome of union bargaining and individual labor supply decisions.

A further concern is that State X introduced measures to reduce the costs of entering the teaching profession through alternative certification pathways. These changes included the authorization of school districts (rather than just colleges and universities) to provide professional preparation programs for certification beginning in the 2002-2003 school year and a law in 2004 allowing for the creation of teacher preparation institutes for college graduates with a non-education degree to receive certification (Feistritzer 2007). These policies may alter both the selection into entry cohorts (lowering entry costs) and the quality of the training received by entry cohorts (changing the required training). If these measures led to a change in the labor supply of teachers in CSR years, part of what is estimated as changes in cohort quality in this paper may be capturing these changes as well. Fortunately, the uptake of these alternative pathways was quite low over the period of our data. Sass (2011) documents the number of teachers in grades three through ten from 2000-2001 to 2006-2007 certified by these two pathways at only 1,679. Clearly, the number of these alternatively certified teachers in grades four through six will be much lower and, in the longer run, some substitution from traditional certification may be expected, suggesting little role for the introduction of these two programs to be driving the results that follow.

4 Empirical Approach and Baseline Estimates

Our two main goals are to investigate whether there was a drop in entering teacher quality associated with the increased hiring at the introduction of CSR and whether any such drop in quality can explain the lack of an estimated effect of the CSR policy. We start by considering the first question. For now, this requires us to identify the effect of particular hiring cohorts on student achievement. Later we will consider the implications for CSR effect estimates and our ability to tie any such changes to CSR hiring.

The methodology used here follows from the standard value-added approach to education production function estimation. For the purposes of this paper, teacher quality will be defined as the contribution teachers make to student mathematics achievement growth. While it is clear that test scores are only one facet of a student's academic growth and that a good teacher may contribute to other areas such as a child's social development, the advent of school accountability programs has positioned test scores as the key measure used to assess teachers and schools. Indeed, value-added to test scores is a particularly appropriate metric for assessing why test scores did not increase more with CSR.

Here, we outline the basic strategy for identifying changes in teacher quality. These baseline estimates are presented along with several sensitivity checks and then our preferred estimates that account for teacher attrition are presented in section 5. The baseline specification discussed here estimates a single cohort value-added effect for each entry cohort. This provides for a more tractable comparison among the several estimators considered and can be considered a weighted average of the results that follow and is, therefore, a good summary measure.[11] The intuition presented here for interpreting the results broadly applies to the other estimates as well. The main strategies used are based on OLS estimation using student level observations of what will be referred to as a lag score specification due to the presence of the student's prior test score as an explanatory variable:[12]

$$\begin{aligned} A_{igjst} =& \zeta_t + \lambda A_{igjst-1} + X_{igjst}\beta + Cohort_j\gamma_1 + \gamma_2\overline{A}_{-igjst-1} + f(Exp_{jt}) \\ & + \gamma_3 CS_{igjst} + \phi_g + c_i + \delta_s + e_{igjst} \end{aligned} \tag{4.1}$$

where
i, g, j, s, t index student, grade, teacher, school, and year
A_{igjst} is student *i's* test score
ζ_t are year fixed effects
$A_{igjst-1}$ is student *i's* prior test score
X_{igjst} are student demographics[13]
$Cohort_j$ are teacher cohort indicators
$\overline{A}_{-igjst-1}$ is the average prior test score of student *i's* classmates

[11] Consider the simple case of estimating a single cohort effect with no other covariates. the estimating equation is given by $A_i = \gamma D_i + u_i$ where D_i is an indicator for having a teacher from a particular cohort. In this simple setting the OLS estimate will be the Wald Estimator: $\hat{\gamma} = E[A_i|D = 1] - E[Y|D = 0]$. Later we will allow for separate cohort effects by year, effectively splitting the $D = 1$ group into supgroups denoted by d_1 and d_2 yielding the following estimating equation: $A_i = \gamma_1 d_{1i} + \gamma_2 d_{2i} + u_i$. It is straightforward to show that our original estimate will be a weighted average of the subgroup effects: $\hat{\gamma} = \gamma_1 Pr(d_1 = 1|D = 1) + \gamma_2 Pr(d_2 = 1|D = 1)$

[12] See Appendix B: Measuring Teacher Quality for a discussion of value-added estimation.

[13] The student controls include indicators for race, gender, disability status, free or reduced price lunch status, limited English proficiency, being foreign born, as well as the student's age and the number of days present and absent the prior year.

$f(Exp_{jt})$ is a cubic in teacher experience
CS_{igjst} is a proxy measure of class size [14]
ϕ_g are grade fixed effects
c_i is an unobserved student heterogeneity term
δ_s are school fixed effects

Note that OLS estimation of (4.1) (our preferred strategy) ignores c_i. While this assumption may appear strong, there is evidence that OLS estimation of the lag score specification typically performs well. Using simulated data, Guarino et al. (2015) find that the lag score specification estimated by OLS is fairly robust, compared to other common value-added estimators, to different teacher and student sorting mechanisms. Kane & Staiger (2008) find that this method does the best at estimating a teacher's value-added in non-experimental settings by comparing estimates for the same teachers both with and without random assignment to students. The intuition for this result is that assignment is driven more by dynamic (i.e. changes in test performance), rather than static, characteristics of students. Estimators that attempt to eliminate unobserved student heterogeneity introduce additional assumptions and greatly reduce the identifying variation, while failing to capture much of the assignment mechanism that threatens the validity of the estimates. The presence of c_i only threatens the consistency of our results if student-teacher assignment decisions are made in such a way to induce a correlation between the time-constant student heterogeneity and the hiring year of a student's teacher. In exploring the sensitivity of the results below, we will argue that

[14] Class size is measured by the number of students linked to a teacher in a given year in the test data. While this serves as a reasonable proxy in fourth and fifth grade, it is less reliable in sixth grade when many schools have teachers teaching multiple classes. In estimating (4.1) we allow for different effects of class size for each grade. The proxy measure of class size is important for separating out the quality of newly hired teachers from any effect the reduced class sizes may have had on achievement under CSR. Importantly, there is sufficient within cohort and within year variation in class size to separately identify the class size and cohort effects. Due to the potential for biased class size effects from measurement error, we conducted an exercise in which we estimated our main specification constraining the effect of class size at different plausible values yielding qualitatively similar results for our cohort effect estimates. Results are available upon request.

such assignment policies are unlikely in practice.

The main coefficients of interest are the estimates of γ_1, the average quality of entry cohorts of teachers. Specifically, interest lies in comparing the average quality of cohorts hired before and after the introduction of CSR. The teacher-quality explanation for the poor performance of CSR would be consistent with smaller gains associated with cohorts entering the data after CSR was implemented compared to earlier cohorts.

The inclusion of δ_s, the school fixed effects, is important for two reasons. First, it helps to control for differences across schools in student ability. The school fixed effects are also critical to identify whether schools hired teachers of lower quality in CSR years. Given evidence that there is strong sorting of teachers into geographically small markets (Boyd et al. 2005; Lankford et al. 2002), schools may face different levels of average quality. For now, assume there was no change in the quality of teachers hired by particular schools, but that CSR disproportionately induced hiring in schools that faced supplies of lower quality teachers. In this scenario, without controlling for these school level differences we would identify a negative relationship between CSR years and the average quality of new entrants. The school fixed effects control for the time-invariant quality level of teacher supply that different schools face by relying on within school comparisons of teachers. We also consider an alternative approach that relies on within school-grade-year variation.

The experience profile captures three distinct factors: teaching-specific human capital, non-random assignment of students to teachers based on experience, and non-random attrition of teachers. Focusing on the human capital piece, the possible effect of CSR on short-run achievement is better captured when the experience of the teacher is not controlled for. However, controlling for experience allows for a more direct comparison of teacher quality throughout the sample period. If experience is not controlled for, teachers from earlier cohorts may look better than later cohorts simply because the estimates are partially based on years in which these teachers have more experience than later cohorts. The joint contribution of both cohort quality and experience to student achievement is considered in Appendix F.

As we noted at the end of section 3, care should be taken in interpreting the estimates of equation (4.1) as identifying CSR induced hiring effects. Note, however, that changes that affect *all* students or teachers in a particular year, such as changes in curriculum, will be controlled for by the inclusion of the year fixed effects, ζ_t. Here the main concerns are factors that alter the performance of students with teachers hired in a particular year (i.e. that affect only the students with a teacher hired in 2003-04 but not 2002-03) and are therefore captured in the estimates of γ_1. As mentioned in the previous section, there is suggestive evidence that two first order concerns, the expansion of alternative certification pathways and concurrent changes in the financial attractiveness of teaching relative to other occupations in State X, are not driving changes in estimated cohort quality.

The approach used here captures potential CSR effects that would be difficult to identify given the available data. For example, the school-level class-size averages within enforcement grade groups are only available starting with the year directly before school-level enforcement.[15] This data limitation makes it difficult to identify schools that may have hired teachers during district-level enforcement years in order to preempt the switch to school-level enforcement. The estimates of γ_1 for the 2005-2006 hiring cohort will include the effect of schools hiring additional teachers because of the switch in enforcement the following year.

Note that these value-added measures may also capture changes in resources that complement a teacher's ability to raise achievement. If CSR led to a reduction in these resources available to newly hired teachers (relative to teachers from earlier hiring cohorts), then part of the change in measured cohort effectiveness over time may be capturing these changes as well. We suspect, however, that many of these changes in resources will apply to all

[15] While the state does have records of average class size at the school level for several years prior to CSR, these are not separated by the enforcement grades. Since many of the schools studied here include grades in both the K-3 and 4-8 enforcement groupings, it is difficult to create a comparable measure of average class-size that is directly related to CSR enforcement. Furthermore, these other class-size records are based on student counts in October, while the CSR enforcement averages are based on counts made in February.

teachers in a particular school and year, both in new and old cohorts. Therefore, we can compare the performance for teachers hired in different years, but teaching in the same schools at the same time to test whether such changes are driving the results. In the end we find evidence that this is not the case.

Table 3 presents the baseline estimates of the cohort effects (γ_1) from equation (4.1) in the first column.[16] The policy-relevant comparison is between pre-CSR and post-CSR cohorts. We use the convention of shading district CSR enforcement years in light gray and school CSR enforcement years in dark gray. For reference, the initial cohort size is also presented. All specifications are estimated using developmental scale test scores that have been standardized within grade and year.

The results show that students with teachers who entered during CSR perform worse on average. For instance, students of teachers from the 2006-2007 cohort are estimated to score, on average, over one-fiftieth of a standard deviation (0.0317-0.0088=0.0229; p-value=0.000)[17] worse than students with a 2002-2003 cohort teacher. Note, in each case, the estimated cohort effects are relative to the set of teachers already teaching in State X in the 1999-2000 school year (hired in or before 1999) and represent the conditional mean performance of students across all years for a given cohort. We will return to this in more detail in section 5, however, the slight drop in cohort value-added before the introduction of CSR is driven by having fewer years of data for each successive cohort coupled with nonrandom attrition over time of the lowest performers. When we account for attrition in Section 5, the first year effects for pre-CSR cohorts are nearly identical. Overall, the estimated post-CSR cohort effects range from 0.0069 (p-value=0.147) to 0.0360 (p-value=0.000) standard deviations lower than the two pre-CSR cohorts.[18]

In addition to the baseline estimates, we consider three main sensitivity

[16] See Appendix Table 2 for other estimates from this regression and Appendix Table 3 for results using Reading test scores.

[17] Throughout we will present p-values for tests that two particular cohort values are the same.

[18] All pre- post-CSR cohort comparisons are statistically significant at the 5% level except the comparison between the 2002-2003 cohort and the 2003-2004 cohort.

Table 3: Baseline Cohort Effect Estimates and Sensitivity

	(1)	(2)	(3)	(4)	(5)	(6)
Prior Score	*Lag*	*Gain*	*Gain*	*Lag*	*Lag*	*Gain*
Estimator	*OLS*	*FE*	*FE*	*FDIV*	*OLS*	*OLS*
Entry Cohort						
2000-2001	0.0043	0.0061	0.0084	0.0071**	0.0050	0.0102
N=3957	(0.0030)	(0.0057)	(0.0057)	(0.0032)	(0.0031)	(0.0118)
2001-2002	-0.0027	0.0038	0.0003	0.0031	-0.0026	0.0018
N=3023	(0.0033)	(0.0081)	(0.0074)	(0.0040)	(0.0027)	(0.0195)
2002-2003	-0.0088***	-0.0135**	-0.0097	-0.0146***	-0.0096***	-0.0070
N=3171	(0.0025)	(0.0064)	(0.0071)	(0.0035)	(0.0029)	(0.0134)
2003-2004	-0.0157***	-0.0267***	-0.0206**	-0.0241***	-0.0143***	-0.0203
N=3719	(0.0046)	(0.0051)	(0.0082)	(0.0026)	(0.0046)	(0.0127)
2004-2005	-0.0217***	-0.0360***	-0.0319***	-0.0298***	-0.0193***	-0.0265**
N=4497	(0.0045)	(0.0098)	(0.0087)	(0.0049)	(0.0045)	(0.0137)
2005-2006	-0.0301***	-0.0438***	-0.0432***	-0.0382***	-0.0250***	-0.0484***
N=4714	(0.0024)	(0.0066)	(0.0062)	(0.0041)	(0.0023)	(0.0140)
2006-2007	-0.0317***	-0.0431***	-0.0409***	-0.0391***	-0.0290***	-0.0395**
N=4882	(0.0045)	(0.0100)	(0.0082)	(0.0051)	(0.0034)	(0.0150)
2007-2008	-0.0264***	-0.0164*	-0.0151**	-0.0243***	-0.0239***	-0.0244*
N=4463	(0.0047)	(0.0092)	(0.0074)	(0.0049)	(0.0039)	(0.0147)
Fixed Effects						
Student	No	Yes	Yes	Yes	No	No
School	Yes	No	Yes	No	No	No
Grade	Yes	Yes	Yes	Yes	No	Yes
Year	Yes	Yes	Yes	Yes	No	Yes
School-Grade-Year	No	No	No	No	Yes	No
Student-School	No	No	No	No	No	Yes
Observations	2,752,060	2,752,060	2,752,060	1,329,658	2,752,060	2,752,060
R-Squared	0.653	0.399	0.412	–	0.674	0.668

Standard errors clustered at the District level in parentheses

*** p<0.01, ** p<0.05, * p<0.1

checks.[19] First, we address the unobserved heterogeneity term (c_i) found in equation (4.1). We consider two ways to control for c_i. First, we use the fixed effects (FE) estimator that can be obtained by OLS on the within-student time-demeaned data. Importantly, the FE estimator is inconsistent when lagged dependent variables are included as explanatory variables. Instead we control for prior achievement by using the test score gain as the dependent variable (fixing $\lambda = 1$ in (4.1)).[20] Columns (2) and (3) of Table 3 display cohort effects estimated by FE both excluding and including the school fixed effects, respectively.[21] We also consider a 2SLS version of the Arellano & Bond (1991)

[19] In Appendix E we also disaggregate the cohort effects by CSR pressure, finding no clear pattern consistent with schools facing more pressure hiring lower quality teachers. This is also consistent with the results in section 5.

[20] Note that the choice of the gain score or lag score estimating equation is of little consequence here, with OLS estimates producing nearly identical cohort effect estimates.

[21] Controlling for student and school fixed effects simultaneously relies on the presence of sufficient school-switching among students, as such, we consider estimates both with and

dynamic GMM estimator, referred to as the First Differenced Instrumental Variables (FDIV) estimator, in order to address the presence of c_i while not constraining $\lambda = 1$.[22]

Comparing columns (1) through (4) of Table 3 shows that each estimator leads to the conclusion that the post-CSR cohorts have lower value-added than the pre-CSR cohorts. For instance, comparing the estimated difference between the 2002-2003 and 2006-2007 cohorts, all estimators suggest similar magnitudes of this effect with the largest being in column (3).

Given concerns over the role unobserved student ability may play in estimating education production functions, it may be surprising that the methods used to address unobserved heterogeneity (FE and FDIV) yield similar results to those that do not. As was alluded to before, the unobserved heterogeneity threatens the consistency of the estimates if schools were using a static unobserved characteristic of students to determine whether a student would be taught by a teacher hired in a particular year. It seems reasonable, particularly when controlling for teacher experience, that schools were not engaging in this sort of non-random assignment. While it may certainly be the case that student achievement is affected by a student's innate ability and that this ability is used by schools in making some decisions, it does not appear to be used in a way that would lead to inconsistencies in our main estimates.

For our second sensitivity check, we replace the separate school (δ_s), grade (ϕ_g), and year (ζ_t) effects with a single school-by-grade-by-year fixed effect. As a thought experiment, the baseline estimates identify each cohort effect using within school comparisons of student performance in classes taught by teachers hired in different years while flexibly controlling for state-wide time trends and time constant differences across grades in average achievement. This leaves the potential for other factors particular to a school in a given year (change in leadership) or grade (pedagogical approach) to affect our estimates. To generate problems, these factors must be related to the student-teacher as-

without the school effects.

[22] Note that the sample size is decreased substantially for the FDIV estimator as the requirement of a twice lagged score leaves only students with three consecutive test scores in the estimation sample thereby excluding all fourth grade students.

signment decision in such a way to induce a correlation between the cohort indicators and the unobserved factors even after controlling for the other covariates. In contrast, the estimates when including the school-by-grade-by-year fixed effects effectively control for any unobserved factors particular to a given school-grade-year that may affect student achievement. In fact, this would include school-grade-year factors related to CSR implementation, such as splitting up classes or altering the use of building space, that on average affect classes taught by all cohorts of teachers. This added flexibility comes at the cost of relying on within school-grade-year comparisons in order to identify the cohort effects. That is, school-grade-year observations only contribute to the estimation of a particular cohort effect if there is at least one teacher from that cohort and one from another cohort teaching in that school-grade-year.[23]

Column (5) displays the cohort effect estimates when including school-by-grade-by-year fixed effects. Columns (1) and (5) show very similar pre-CSR cohort effects, while the absolute value of the post-CSR effects are slightly smaller in magnitude. However, this slight change does not alter the conclusion that post-CSR cohorts tend to have lower value-added than pre-CSR cohorts.[24]

One final sensitivity check concerns the movement of students to new schools in response to CSR. If there are student-school match effects (i.e. some schools better suit a particular student's learning needs) and if CSR increased transitions between schools due to capacity constraints, then students may be forced to move to a school with different match quality. One might worry that match quality will tend to be higher on average in the pre-CSR "unconstrained" school choice than in the post-CSR "constrained" match. In this case, the previous estimates may partially be capturing the associated reduction in student-school match quality, rather than a drop in teacher quality.

[23] Omitting the school fixed effects entirely and including school characteristics identifies the cohort effects by comparing teachers across schools as well. While this may increase the number of comparisons that contribute to identification, such an approach is the most susceptible to omitted variables bias as outlined above. Here, this approach leads to a similar conclusion that students in post-CSR cohort classes perform worse.

[24] The results are also invariant to the many potential combinations of year, grade, school, school-year, school-grade, and grade-year effects that could be included in the model. See Appendix Table 4 for these results.

To address this possibility, in column (6) we present results with student-by-school fixed effects. Once again the general trend of lower quality in the larger post-CSR cohorts persists, although the standard errors are noticeably larger.

Motivated by these results and the prior literature discussed above, throughout the remainder of the paper we will estimate variants of (4.1) by OLS controlling for separate grade, year, and school effects.

Note that the comparison among the estimated cohort effects does not fully capture the contribution of these teachers to average state-wide achievement. In particular, this comparison misses the fact that not all students in CSR years are taught by teachers hired in post-CSR cohorts and that the average experience in the state dropped in post-CSR years. In Appendix F, we show that the cohort and experience profile estimates suggest an overall decline in average student performance of 0.0154 standard deviations. The drop is shown to be driven predominately by cohort quality rather than experience effects.

5 Further Analysis: CSR Performance, Long Run Effects, and Distributional Effects

5.1 Entering Teacher Quality and CSR Performance Estimates

Prior research found no evidence of CSR policy effects in State X using treatment-control comparisons with schools with average class size above the new maximums considered treated while those below were the control group. In Appendix C, we confirm these previous results using our sample and methods. We now consider whether the lack of a CSR effect in the prior literature can be attributed to a CSR hiring induced fall in teacher quality. To do so, we modify equation (4.1) by interacting a CSR District treatment dummy variable with all included regressors allowing for separate cohort effects in treated and control schools.

We find little difference in the estimated cohort effects across the two sets of schools. Our focus here is on the implied contribution of new teacher hiring to average student achievement in treatment and control schools. That is, we want to directly address the teacher quality hypothesis in relation to the prior quasi-experimental estimates of CSR performance that have been used to conclude that CSR was ineffective. To do so, we need to consider both the differences in cohort quality and the rise in the number of inexperienced teachers associated with the introduction of CSR documented in section 3.

The estimated contribution to average achievement in both treatment and control schools ($j = T, C$) of the cohort composition are calculated in each year as $\overline{COHORT}_{jt}\hat{\gamma}_{1j}$. This effectively weights the cohort effects by the proportion of students with teachers from each cohort in treatment or control schools in each year. Similarly, we estimate the experience contribution in both sets of schools by $\overline{\hat{f}(EXP_{jt})}$.

Table 4 displays the evolution of the total contribution (cohort composition plus experience) of teachers to average performance separately for schools considered treated and untreated. Table 4 also shows the difference in these changes between treated and untreated schools. Column six is of particular interest as it relates to the type of comparison used to estimate CSR policy effects. Specifically, prior studies rely on treatment-control comparisons (Difference-in-difference (DinD), Comparative Interrupted Time Series, or other related estimators) to estimate CSR effects. Loosely speaking, instead of examining the DinD of student achievement as in the prior work, here we consider the DinD of the portion of student achievement attributable to teacher cohorts and experience. Both treated and untreated schools experience a drop in the teachers' contribution to average achievement. Interestingly, the CSR schools saw a slightly smaller drop, the largest difference being 0.0015 test score standard deviations in 2005-2006, than those schools for which CSR was not binding at introduction. This estimate is small relative to the unrealized CSR achievement gains and, most importantly, is of the opposite sign needed to explain the finding of no achievement gain from CSR.

While we find no evidence of a differential change in teacher quality for the treatment and control schools, the fact that the schools saw similar declines

Table 4: Estimated Total Contribution to Average Achievement: Treatment vs. Control Schools

	Total Achievement Contribution			**Change from 2001-2002**		
Year	Treatment	Control	Difference	Treatment	Control	Difference
2001-2002	0.0380***	0.0388***	-0.0008***	-	-	-
	(0.0081)	(0.0080)	(0.0002)	-	-	-
2002-2003	0.0367***	0.0369***	-0.0002	-0.0013**	-0.0019**	0.0006
	(0.0081)	(0.0084)	(0.0004)	(0.0006)	(0.0008)	(0.0004)
2003-2004	0.0343***	0.0347***	-0.0004*	-0.0038***	-0.0041***	0.0004**
	(0.0087)	(0.0088)	(0.0002)	(0.0010)	(0.0011)	(0.0002)
2004-2005	0.0331***	0.0336***	-0.0005**	-0.0049***	-0.0053***	0.0003***
	(0.0080)	(0.0080)	(0.0002)	(0.0018)	(0.0018)	(0.0001)
2005-2006	0.0293***	0.0282***	0.0011***	-0.0087***	-0.0106***	0.0019***
	(0.0082)	(0.0080)	(0.0003)	(0.0013)	(0.0012)	(0.0003)
2006-2007	0.0234***	0.0227***	0.0007***	-0.0146***	-0.0161***	0.0015***
	(0.0086)	(0.0086)	(0.0002)	(0.0013)	(0.0014)	(0.0002)
2007-2008	0.0209**	0.0202**	0.0007***	-0.0171***	-0.0186***	0.0015***
	(0.0095)	(0.0096)	(0.0002)	(0.0022)	(0.0023)	(0.0002)

Standard errors clustered at the District level in parentheses
*** $p<0.01$, ** $p<0.05$, * $p<0.1$

in achievement attributable to the teacher stock post-CSR is potentially interesting. An important question that emerges is whether this general decline in cohort quality during CSR implementation can be attributed to the CSR induced hiring increase. It seems reasonable that while only some schools faced direct pressure to increase hiring to reduce class size, all schools in an area may be affected. By hiring teachers in the same market, the CSR-induced demand shift would force all schools along the effective labor supply curve to hire lower quality teachers. That is, the effects of CSR on teacher quality may "spillover" to schools that were not under pressure to reduce class size, but were hiring teachers for other reasons. In this way, the general decline could still be the result of CSR. Furthermore, it would go unnoticed in any treatment-control comparison of CSR.

We readily acknowledge that there may be a number of alternative explanations for the concurrent decline in cohort quality, however the set of plausible explanations is limited in a few ways. Alternative explanations should be consistent with the key patterns in Table 4: similar relative new teacher quality in treatment and control schools before CSR followed by a similar decline in new teacher quality for both sets of schools. Again, these patterns are evident after controlling for time constant school factors and statewide trends captured by the school and year fixed effects, respectively, and are robust to a number of alternative estimation approaches. Any alternative explanation must also be related to the quality of teachers hired in post-CSR years *relative* to earlier cohorts and not just the performance of all teachers post-CSR more generally. That is, potential factors must disproportionately impact new hires or the students in classes taught by new hires in order to explain the general decline in new hire quality.

Therefore, our primary concern are changes in the composition of entering teachers that may be driven by changes in both the selection into teaching and the training received. In section 3, we discussed changes in certification policy and the financial attractiveness of teaching in State X. The low take up rates of the new certification policies make it unlikely that the decline comes from either a change in selection (due to lower entry costs) or training (allowing teachers with alternative training backgrounds) from these policies. We also found no direct evidence of a sudden break in trend for the relative financial attractiveness of teaching in State X that could directly spur a change in composition of entry cohorts.

Ultimately, there may be other undocumented changes being captured by the cohort effect estimates. Unfortunately, the "general equilibrium" nature of the potential hiring spillovers due to the CSR induced hiring increase make it difficult to completely account for other factors through common quasi-experimental techniques that estimate treatment effects for only a subset of the population.

Further, any difference-in-difference type approach would require identifying a control state that can be used to approximate the change in teacher quality that State X would have experienced in the absence of CSR. However, given nontrivial differences in test content and scaling, educational institutions, policy adoption, and general economic factors coupled with the finding that teachers tend to focus job search in small geographic areas makes finding a credible control state extremely difficult. This issue is compounded by the practical fact that we are limited to a small set of states that have comparable linked student-teacher data over the same time period. Indeed, an exploratory analysis of publicly available state-wide data from North Carolina, a state that does have similar data to State X over this time period, found large differences in the characteristics of incoming teachers between the two states. For instance, incoming teachers in North Carolina were, on average, nearly twenty percentage points more likely to have no prior experience over this time period than incoming teachers in our sample over the same time period.[25] While this does not tell us about value-added differences between the two states, it does suggest that North Carolina may not be a clean "control" state when considering the labor market for teachers in State X.

Nonetheless, given the plausible connection between the increased hiring due to CSR and the fall in cohort quality, it is important to consider the extent of the potential effect in more detail. Therefore, for the remainder of the paper we explore this general trend toward lower quality cohorts. In particular we wish to document the potential magnitude of the effect by considering the long run effect on teacher quality and to move from cohort effect summary measures to consider the impact for individual students.

5.2 Long Run Hiring Effects

The estimates above combine the initial performance level for a cohort with the longer-term impact of that cohort as the composition changes. With non-random attrition, having a single cohort indicator for the 2001-2002 cohort will disproportionately weight the estimates toward the teachers that stay in the data longer. Conversely, the estimated 2007-2008 cohort effect roughly weights each teacher evenly, regardless of their eventual attachment, giving an estimate of the initial performance.

To address whether the CSR induced demand increase led to both the hiring and retention of lower value-added teachers, as well as the possibility that attrition from teaching led to different long-term cohort effects, the cohort-

[25] See Appendix G

Table 5: Pooled OLS Cohort-by-Year Estimates

Specification				*Cohort-by-Year*			
Equation				(5.1)			
Year	*2001-2002*	*2002-2003*	*2003-2004*	*2004-2005*	*2005-2006*	*2006-2007*	*2007-2008*
Entry Cohort							
2000-2001	-0.0135**	0.0004	0.0009	-0.0039	0.0068	0.0069	0.0044
	(0.0063)	(0.0068)	(0.0053)	(0.0061)	(0.0051)	(0.0056)	(0.0060)
N	2764	2419	2091	1965	1795	1608	1479
2001-2002	-0.0441***	-0.0173***	0.0024	0.0014	0.0103*	0.0102	0.0112*
	(0.0055)	(0.0054)	(0.0064)	(0.0061)	(0.0057)	(0.0065)	(0.0057)
N	3023	2119	1741	1645	1452	1308	1179
2002-2003		-0.0455***	-0.0117***	-0.0126*	0.0015	0.0113	-0.0047
		(0.0053)	(0.0039)	(0.0064)	(0.0064)	(0.0079)	(0.0057)
N		3171	2131	1858	1636	1440	1323
2003-2004			-0.0488***	-0.0315***	-0.0010	-0.0022	-0.0079
			(0.0064)	(0.0057)	(0.0065)	(0.0096)	(0.0066)
N			3719	2635	2219	2002	1817
2004-2005				-0.0696***	-0.0244***	-0.0081	0.00034
				(0.0072)	(0.0055)	(0.0058)	(0.0055)
N				4497	3132	2626	2261
2005-2006					-0.0632***	-0.0291***	-0.0196***
					(0.0047)	(0.0038)	(0.0047)
N					4714	3188	2684
2006-2007						-0.0670***	-0.0260***
						(0.0039)	(0.0058)
N						4882	3340
2007-2008							-0.0580***
							(0.0054)
N							4463
Observations				2,752,060			
R-squared				0.653			

District Cluster Robust standard errors in parentheses: *** p<0.01, ** p<0.05, * p<0.1

specific indicators in (4.1) are replaced with cohort-by-year indicators:

$$\begin{aligned} A_{igjst} =& \zeta_t + \lambda A_{igjst-1} + X_{igjst}\beta + Cohort \times Year_{jt}\gamma_1 + \gamma_2 \overline{A}_{-igjst-1} + \\ & f(Exp_{jt}) + \gamma_3 CS_{igjst} + \phi_g + c_i + \delta_s + e_{igjst} \end{aligned} \tag{5.1}$$

Table 5 displays the estimates of equation (5.1). To interpret the table, we begin along the diagonal with each cohort's first year effect, following year-by-year along the row. For instance, the 2005-2006 cohort has an estimated effect of -6.32% of a test score standard deviation in their first year (cohort size=4,714), -2.91% in their second year (cohort size=3188), and -1.96% in their third year (cohort size=2684). While the initial productivity of the earlier cohorts is lower than the previous estimates would suggest, the relative performance of cohorts in their first years is essentially unchanged from the previous estimates with post-CSR cohorts having average achievement 0.0033 (p-value=0.558) to 0.0255 (p-value=0.004) standard deviations below the pre-CSR cohorts.[26] Note that for the 2000-01 hiring cohort, the first estimate shown is from 2001-02, their second year. The point estimates suggest the relative performance gap between pre-CSR and post-CSR cohorts is between 0.0109 (p-value=.330) and 0.0198 (p-value=0.003) standard deviations in each cohort's second year. Importantly some, but not all, second year cohort effects are statistically different at conventional levels.[27]

Also note that pre-CSR cohorts become comparable to the baseline teachers after three or four years with year-specific cohort effects statistically indistinguishable from zero. The two post-CSR cohorts observed for at least four years, 2003-2004 and 2004-2005, also appear to level off to be roughly comparable to the baseline after four years. This result suggests that the potential long-run CSR hiring effects may be even smaller than those initially observed. However, the largest post-CSR hiring cohorts are not observed long enough to make a complete comparison across all cohorts. In particular, the estimated third-year effect for the 2005-2006 cohort is still statistically different from zero, at nearly one-fiftieth of a standard deviation.

It is important to note that there is sufficient within cohort variation in initial experience (particularly for the baseline group) to still control for teacher experience. This implies that the observed improvement for cohorts is not reflecting human capital accumulation that is common to all cohorts. However, it may still be the case that the cohort-by-year effects capture deviations from the average experience profile that are unique to each cohort. That is, post-CSR cohorts may be initially lower performing and have smaller than average improvements with experience.

[26] First year cohort differences that are not statistically significant at the 10% level include 2001-2002 to 2003-2004 (p-value=0.539), 2002-2003 to 2003-2004 (p-value=0.558), and 2002-2003 to 2007-2008 (p-value=0.104).

[27] Second year cohort differences that are not statistically significant at the 10% level include 2000-2001 to 2004-2005 (p-value=0.250), 2000-2001 to 2006-2007 (p-value=0.136), 2001-2002 to 2004-2005 (p-value=0.330), and 2001-2002 to 2006-2007 (p-value=0.227)

The fact that post-CSR cohorts perform worse in their second years suggests that not only may schools be initially hiring lower value-added teachers due to the CSR-induced demand increase, but the schools may be retaining more low value-added teachers longer in order to meet CSR requirements. State X is notable for dismissing teachers within their first three years for poor performance at a much higher rate than the nation as a whole, with the state's ninety-seven day probationary rule cited as a possible explanation. However, these results suggest that the short run CSR demand increase may have weakened this mechanism for ensuring quality instruction. Both phenomenon, the hiring and retention of lower value-added teachers, fit nicely within the framework of a simple search model of teacher hiring in which teachers are effectively viewed as experience goods (see Rockoff & Staiger 2010). However, it appears that the long-run achievement effect of these changes may be relatively small.

A comparison across cohorts within the same year lends some insight into the role other inputs into the education process may have had in affecting student performance over this time. In particular, the effect of unmeasured changes in classroom inputs directly complementary to teaching may be included in the cohort effect estimates. Recall that there is some anecdotal evidence that State X's CSR program was not fully funded, raising the possibility that a reallocation of other inputs may have coincided with the hiring increase studied here. However, since all teachers likely face similar resources within schools in a given year, the fact that the earlier cohorts perform noticeably better in each year suggests that it is not changes in these other complementary inputs driving the results. For instance, in the 2004-2005 school year the 2002-2003 cohort has an estimated cohort effect over one-twentieth (0.0696-0.0126=0.0570; p-value=0.000) of a standard deviation better than the 2004-2005 cohort. This is a practically and statistically significant difference in performance that is likely not due exclusively to differences in other classroom-level inputs.

The results found here are consistent with the notion that increased teacher hiring is associated with a modest short-run decrease in quality. For instance, the introduction of a value-added based retention policy may indeed increase hiring and reduce new hire quality, partially offsetting the potential gains from the policy. However, it is possible that there remain uncontrolled for trends in cohort quality driving the result. This serves as an important caveat for considering the implications for teacher hiring more generally. Recall from section 3 however, that the most obvious policy reason for such a trend, the introduction of new certification pathways, likely had little affect on the composition of teachers during this time. We also documented the fact that there was no sudden contemporaneous, or appropriately lagged, change in the relative financial attractiveness of teaching in State X over this time period.

Importantly, while we saw a pre-policy decline in teacher experience in section 3, Table 5 shows no pre-policy trend in cohort value-added with pre-policy cohorts performing similarly in the first and second years with first year effects by cohort of -0.0441 [2001-02] and -0.0455 [2002-03] and second year effects of -0.0135 [2000-01], -0.0173 [2001-2002], and -0.0117 [2002-03]. This suggests that there was no pre-policy trend toward hiring lower quality cohorts, lending some support for the idea that the drop in quality we document post-CSR might be due to the associated hiring increase. Recall from section 3, if the pre-CSR drop in experience was an indication that schools would be forced to move along several unobserved margins typically not correlated with experience to hire lower quality teachers in the coming years, than our estimates serve as an upperbound for the CSR induced hiring effect.

Furthermore, as we showed earlier, a general trend toward lower quality cohorts does not affect the evaluation of CSR in light of the teacher quality hypothesis. In short, any general trend in cohort quality differenced out when comparing schools under different CSR pressure. Finally, note once more that any policy change that affects *all* students or teachers in a given year, grade, or school are controlled for by the included fixed effects.

5.3 Discussion of Hiring Effects

The cohort effect estimates presented above suggest a short run drop in student performance associated with the larger post-CSR hiring cohorts. It is important to consider the magnitude of these effects in interpreting the potential achievement impact on students of increased hiring. Our estimates suggest that the post-CSR hiring cohorts have students who perform as much as 2.55% of a test score standard deviation worse *on average* in the cohort's first year compared to the smaller pre-CSR cohorts. Relative to many other education production function estimates, these average effects may seem small. Indeed, they are much smaller than the Tennessee STAR class size effect estimates mentioned earlier of roughly one-fifth of a standard deviation for an average reduction of eight students.

To provide a benchmark for evaluating the size of the cohort effect differences, we estimate individual teacher value-added by replacing the cohort indicators with dummy variables for each teacher yielding a distribution of teacher value-added.[28] We find that the standard deviation of teacher value-added in our sample is 27.61% of a test score standard deviation. Therefore, a difference in mean cohort quality of 2.55% of a test score standard deviation is nearly 10% of the standard deviation of the teacher quality distribution. Note, the standard deviation of 27.61% has not been adjusted for noise due to small samples for some teachers. As an alternative, we also estimate the standard deviation of the teacher effects using a mixed effects framework with teacher and school random effects replacing the fixed effects in order to account for sampling noise in the individual teacher effect estimates.[29] Using this approach, the estimated standard deviation is 17.14%, implying an even larger relative effect of the post-CSR hiring cohorts of nearly 15% Viewed in this light, the estimated cohort effects represent a modest decline in performance.

While focusing on mean cohort effects was instructive for considering the role hiring quality played in prior quasi-experimental estimates and for providing a clear summary of the overall effects, it may miss large effects for those

[28] Appendix D provides a more detailed analysis of the individual value-added results that closely mirrors the results for the mean cohort effects.

[29] This is an alternative, but related, approach for handling sampling noise to the "adjusted" standard deviations used in Rothstein (2010) or Koedel & Betts (2011). Here, the adjustment procedure found in these prior papers is computationally intense given our sample size as it requires standard errors for the individual teacher effects.

students assigned to the marginal teachers hired due to CSR. To provide a better sense of the effect CSR hiring may have had on individual students, we divide the new hire teachers into quintiles based on the estimated value-added distribution in pre-CSR years for new hires in grades four and five.[30] To provide a reference for the importance of being assigned a teacher in different quintiles, Table 6 displays the mean value-added across all entering (both pre- and post-CSR) teachers by quintile in column (2). We see a large difference across groups, with a mean effect at the low end of -40.74% of a test score standard deviation while the highest quintile is 32.75%. This represents a large difference in teacher quality for students assigned to teachers in different quintiles.[31]

To provide a measure of the distributional effects, we calculate the number of students assigned to a teacher in each quintile for each hiring cohort in their first year. We focus on the first year for each cohort to capture the number of affected students before any teacher attrition occurs. Note, this means we are focusing only on the sub-population of students who receive a teacher that was part of an incoming cohort in that particular year.

Table 6 shows the average across years by quintile separately for the pre- and post-CSR cohorts, as well as the difference between the two groups. To account for the fact that post-CSR cohorts taught more students in their first year, we present a counterfactual distribution of student counts for the pre-CSR cohorts by scaling up the student numbers to match the post-CSR total (multiplying by 53381/39797=1.34). The counterfactual numbers roughly reflect the predicted number of students exposed to incoming teachers from each quintile in post-CSR years if the teachers had been drawn from the pre-CSR new hire quality distribution. Of particular interest is the difference between the actual post-CSR numbers and the counterfactual pre-CSR numbers. To

[30] We focus on grades four and five here since comparisons based on teacher value-added and student numbers are complicated in grade six by the fact that some schools maintain a system with a single teacher for all subjects (the norm in grades four and five) while others have subject specific teachers.

[31] Interestingly the unadjusted standard deviation of first year teacher value added is remarkably stable by cohort, always between 0.28 and 0.30.

help provide a sense of scale, we also present the same information as a percentage of the students who were assigned an incoming teacher. For reference, Table 6 also presents similar statistics for the average number and percentage of teachers from each quintile before and after CSR with the counterfactual numbers scaled by the rise in the average fourth and fifth grade teacher cohort (3109/2174=1.43).

Table 6: Average Number of Students and Teachers by Pre-CSR Value-added Quintile: G4-G5

		Pre-CSR Cohorts			**Post-CSR Cohorts**		**Difference**		
Student Numbers	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Quintile	*Mean Value-added*	*Actual Count*	*Actual Percent*	*Counterfactual Count*	*Actual Count*	*Actual Percent*	*Actual Count (5)-(2)*	*Actual Percent (6)-(3)*	*Counterfactual Count (5)-(4)*
1	-0.4074	7294	18.33%	9784	13652	25.57%	6358	7.25%	3868
2	-0.1507	8041	20.20%	10785	11236	21.05%	3195	0.84%	451
3	-0.0261	8129	20.43%	10904	9917	18.58%	1788	-1.85%	-987
4	0.0966	8259	20.75%	11078	9576	17.94%	1317	-2.81%	-1502
5	0.3275	8074	20.29%	10830	9000	16.86%	926	-3.43%	-1830
Total		39797	100.00%	53381	53381	100.00%	13584	0.00%	0

		Pre-CSR Cohorts			**Post-CSR Cohorts**		**Difference**		
Teacher Numbers	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Quintile	*Mean Value-added*	*Actual Count*	*Actual Percent*	*Counterfactual Count*	*Actual Count*	*Actual Percent*	*Actual Count (5)-(2)*	*Actual Percent (6)-(3)*	*Counterfactual Count (5)-(4)*
1	-0.4074	435	20.01%	622	831	26.73%	396	6.72%	209
2	-0.1507	435	20.01%	622	648	20.84%	213	0.83%	26
3	-0.0261	435	20.01%	622	560	18.01%	125	-2.00%	-62
4	0.0966	435	20.01%	622	542	17.43%	107	-2.58%	-80
5	0.3275	434	19.96%	621	528	16.98%	94	-2.98%	-93
Total		2174	100.00%	3109	3109	100.00%	935	0.00%	0

Focusing on the counterfactual difference, we see 4,319 students (8.09% of all students assigned to incoming teachers) "shifted" from a teacher in the top three quintiles to one in the bottom bottom two due to the lower quality of post-CSR cohorts, with 3,868 (7.25%) being assigned to teachers in the bottom twenty percent. Even if the additional students assigned to a teacher in the lowest quintile would have only had a second quintile teacher in the absence of a fall in quality, the average difference in quality across these groups of 25.67% (40.74-15.07) of a test score standard deviation represents a large difference in teacher quality. The effect on some of these students will be even larger if they had been displaced from having a higher quintile teacher. Whether this represents a large number of affected students is subjective, however it does seem clear that for particular students the shift in quality was potentially quite large.

Finally, note that in the absence of CSR, there would have been fewer teachers hired in post-CSR cohorts implying fewer students assigned to new hires in those years. Rather, they would have received a more tenured teacher who, due to experience and selective attrition, would be of higher quality on average. Therefore, the actual number of affected students is likely larger.

6 Conclusion

The results presented above provide little support for the conclusion that a drop in the quality of newly hired teachers explains the lack of an *estimated* achievement gain from CSR in State X. While there was a modest decrease in student performance attributable to teachers (due to quality and inexperience) with the policy, this decrease was experienced by both treated and untreated schools alike. These spillovers imply that the disappointing CSR effects found in quasi-experimental research cannot be explained by differential changes in the quality of newly hired teachers.

That said, the general fall in quality suggests there may have been a negative effect of CSR on achievement not captured by the quasi-experimental estimates. We do find this effect to be of modest size and only short-term as the lowest performing teachers in each post-policy cohort were the most likely to leave in subsequent years. However, for students taught by the lowest quality marginally hired teachers, the effect was potentially large.

Given that entering teacher quality does not play a large role in the failure of State X's CSR program to achieve expected gains, exploring alternative mechanisms is an important next step. One possibility is that other input levels may have changed, especially in cases in which CSR was implemented without full funding, as was the case in State X. As noted above, however, differences in resources directly used by teachers after CSR may also have a limited scope for explaining CSR performance. Finally, in this paper we focus on the inflow of teachers into the state public elementary school system that accompanied CSR. However, exploratory analysis of the movement of teachers across schools in response to CSR reveals no clear evidence that schools forced to reduce class size fared worse in this regard either. Understanding the mechanisms at play will help to determine whether popular CSR policies can be designed to promote achievement gains.

These conclusions should be interpreted with caution, as our findings reflect the experience of a single state for teachers in grades four to six. In other states or grades, the quality of incoming teachers may fall more dramatically in response to changes in teacher hiring.

References

Angrist, J. D. & Lavy, V. (1999). Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement. *Quarterly Journal of Economics, 114(2)*, 533-575.

Bohrnstedt, G. W. & Stecher, B. M. (1999). *Class-size Reduction in California 1996-1998: Early Findings Signal Promise and Concerns.* Palo Alto, CA.: CSR Research Consortium, EdSource, Inc.

Bohrnstedt, G.W. & Stecher, B.M. (2002). *What We Have Learned about Class-Size Reduction in California.* Sacramento: California Department of Education.

Boyd, D., Lankford, H., Loeb, S., & Wyckoff, J. (2005). The Draw of Home: How Teachers' Preferences for Proximity Disadvantage Urban Schools. *Journal of Policy Analysis and Management, 24(1),* 113-132.

Boyd, D., Lankford, H., Loeb, S., & Wyckoff, J. (2011). Teacher Layoffs: An Empirical Illustration of Seniority versus Measures of Effectiveness. *Education Finance and Policy, 6(3),* 439-454.

Buckingham, J. (2003). Class Size and Teacher Quality. *Educational Research for Policy and Practice, 2,* 71-86.

Center for Local State and Urban Policy (2010). Mandating Merit: Assessing the Implementation of the Michigan Merit Curriculum. `http://closup.umich.edu/files/pr-13-michigan-merit-curriculum.pdf`

Chetty, R., Friedman, J., Hilger, N., Saez, E., Schanzenbach, D., & Yagan, D. (2011). How Does your Kindergarten Classroom Affect your Earnings? Evidence from project STAR. *Quarterly Journal of Economics, 126(4),* 1596-1660.

Council for Education Policy, Research and Improvement (2005). *Impact of the Class-size Amendment on the Quality of Education in Florida.*

Chingos, M. M. (2012).The Impact of a Universal Class-size Reduction Policy: Evidence from Floridas Statewide Mandate. *Economics of Education Review, 31(5),* 543-562.

Feistritzer, C. E. (2007). *Alternative Teacher Certification 2007.* Washington D.C.: National Center for Education Information.

Feng, L. (2009). Wages, Classroom Characteristics, and Teacher Mobility. *Southern Economic Journal, 75(4),* 1165-1190.

Florida Department of Education (n.d.). Class size reduction amendment. Retrieved from `http://www.fldoe.org/ClassSize/`.

Goldhaber, D. (2008). Teachers Matter, But Effective Teacher Quality Policies are Elusive. In Ladd, H. F. & Fiske, E. B. (ed.) *Handbook of Research in Education Finance and Policy.* New York, NY : Routledge, 146-165.

Goldhaber, D. & Theobald, R. (2011). Managing the Teacher Workforce in Austere Times: The Determinants and Implications of Teacher Layoffs. CEDR WP 2011-1.3.

Guarino, C. M., Reckase, M. D., & Wooldridge, J. M. (2015). Can Value-Added Measures of Teacher Performance be Trusted? *Education Finance and Policy, 10(1),* 117-156.

Harris, D., Sass, T., & Semykina, A. (2011). Value-added Models and the Measurement of Teacher Quality. Unpublished draft.

Hoxby, C. M. (2000). The Effects of Class Size on Student Achievement: New Evidence from Population Variation. *Quarterly Journal of Economics, 115(4),* 1239-1285.

Imazeki, J. (n. d.). Class-size Reduction and Teacher Quality: Evidence from California. Working paper.

Jepsen, C. & Rivkin, S. (2009). Class Size Reduction and Student Achievement: The Potential Tradeoff between Teacher Quality and Class Size. *Journal of Human Resources, 44(1),* 223-250.

Kane, T. J. & Staiger, D. O. (2005). Using Imperfect Information to Identify Effective Teachers. Unpublished manuscript.

Kane, T. & Staiger, D. (2008) Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation. Working Paper 14607, National Bureau of Economic Research.

Koedel, C. & Betts J. R. (2011). Does Student Sorting Invalidate Value-added Models of Teacher Effectiveness? An Extended Analysis of the Rothstein Critique. *Education Finance and Policy, 6(1),* 18-42.

Krueger, A. B. (1999). Experimental Estimates of Education Production Functions. *Quarterly Journal of Economics, 114(2),* 497-532.

Krueger, A. B. & Whitmore, D. M. (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.

Lankford, H., Loeb, S., & Wyckoff, J. (2002). Teacher Sorting and the Plight of Urban Schools: A Descriptive Analysis. *Educational Evaluation and*

Policy Analysis, 24(1), 37-62.

McCaffrey, D., Lockwood, J.R., Koretz, D., Louis, T., & Hamilton, L. (2004) Models for Value-added Modeling of Teacher Effects. *Journal of Educational and Behavioral Statistics, 29(1),* 67-101.

Rivkin, S., Hanushek, E. A., & Kain, J. F. (2005). Teachers, Schools, and Academic Achievement. *Econometrica, 73(2),* 417-458.

Rockoff, J. (2009). Field Experiments in Class Size from the Early Twentieth Century. *Journal of Economic Perspectives, 23(4),* 211-230.

Rothstein, J. (2009). Student Sorting and Bias in Value-added Estimation: Selection on Observables and Unobservables. *Education Finance and Policy, 4(4),* 537-571.

Rothstein. J. (2010). Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement. *Quarterly Journal of Economics, 125(1),* 175-214.

Sass, T.R. (2011). Certification Requirements and Teacher Quality: A Comparison of Alternative Routes to Teaching. Working paper.

Staiger, D. & Rockoff, J. (2010). Searching for Effective Teachers with Imperfect Information. *Journal of Economic Perspectives, 24(3),* 97-118.

Stecher, B. & Bohrnstedt G., eds. (2000). *Class-size Reduction in California: Summary of the 1998-1999 Evaluation Findings.*

Todd, P. & Wolpin, K. (2003). On the Specification and Estimation of the Production Function for Cognitive Achievement. *Economic Journal, 113(485),* 3-33.

Appendix

A Additional Tables

Appendix Table 1: Descriptive Statistics

	Mean	**Std. Dev.**		**Mean**	**Std. Dev.**
Test Score	1625.46	246.90	**District CSR**		
Asian	0.02	0.14	*G4-G8 Average Class-size*	24.27	2.86
Black	0.23	0.42	*Below Max*	0.26	0.44
Hispanic	0.23	0.42	*Q1*	0.20	0.40
Other Race	0.03	0.18	*Q2*	0.23	0.42
Female	0.50	0.50	*Q3*	0.17	0.37
Disabled	0.12	0.33	*Q4*	0.14	0.35
Free or Reduced Lunch	0.50	0.50	**School CSR**		
Limited English	0.04	0.20	*G4-G8 AverageClass-size*	20.83	3.15
Age	10.67	1.00	*Below Max*	0.71	0.45
Foreign Born	0.09	0.28	*Q1*	0.07	0.26
Days Present	166.75	21.04	*Q2*	0.07	0.26
Days Absent	7.72	7.70	*Q3*	0.07	0.26
Lagged Peer Score	1515.01	169.72	*Q4*	0.07	0.26
Class-size G4	20.86	8.70	**Entry Cohorts**		
Class-size G5	22.49	11.07	*2001-2002*	0.10	0.30
Class-size G6	82.46	35.32	*2002-2003*	0.09	0.29
Teacher Experience	10.77	10.35	*2003-2004*	0.10	0.30
			2004-2005	0.11	0.31
			2005-2006	0.10	0.30
			2006-2007	0.09	0.29
			2007-2008	0.07	0.25

Source: State X Administrative Data

Appendix Table 2: Estimates from Pooled OLS Regressions

Specification	*Cohort*	*Cohort-by-Year*
Equation	(4.1)	(5.1)
Prior Math Score	0.706***	0.706***
	(0.00564)	(0.00564)
Asian	0.0947***	0.0947***
	(0.00515)	(0.00511)
Black	-0.137***	-0.137***
	(0.00347)	(0.00347)
Hispanic	-0.0273***	-0.0273***
	(0.00242)	(0.00244)

Other Race	-0.0239***	-0.0240***
	(0.00229)	(0.00231)
Female	-0.0160***	-0.0160***
	(0.00148)	(0.00148)
Disabled	-0.185***	-0.185***
	(0.0124)	(0.0125)
Free or Reduced Lunch	-0.0585***	-0.0584***
	(0.00141)	(0.00140)
Limited English	-0.0738***	-0.0742***
	(0.01000)	(0.0100)
Age	-0.0555***	-0.0554***
	(0.00322)	(0.00322)
Foreign Born	0.0706***	0.0706***
	(0.00354)	(0.00356)
Days Present	0.00109***	0.00108***
	(3.58e-05)	(3.56e-05)
Days Absent	-0.00500***	-0.00500***
	(0.000293)	(0.000293)
Experience	0.00731***	0.00502***
	(0.000890)	(0.000699)
Experience Sq	-0.000341***	-0.000231***
	(4.72e-05)	(3.40e-05)
Experience Cu	4.23e-06***	2.76e-06***
	(6.92e-07)	(4.39e-07)
Lagged Peer Score	0.0799***	0.0789***
	(0.0131)	(0.0131)
Class Size	8.97e-05	5.00e-06
	(0.000252)	(0.000258)
*Class Size*G5*	-7.95e-05	-2.58e-05
	(0.000412)	(0.000429)
*Class Size*G6*	-0.000535	-0.000540*
	(0.000328)	(0.000320)
Observations	2,752,060	2,752,060
R-squared	0.653	0.653

Robust standard errors in parentheses:
*** p<0.01, ** p<0.05, * p<0.1

Appendix Table 3: Cohort Effect Estimates for Reading

	(1)	**(2)**	**(3)**	**(4)**	**(5)**
Prior Score	*Lag*	*Gain*	*Gain*	*Lag*	*Lag*
Estimator	*OLS*	*FE*	*FE*	*FDIV*	*OLS*
Entry Cohort					
2000-2001	0.0016	-0.0002	0.0007	-0.0004	0.0017
N=4051	(0.0026)	(0.0048)	(0.0046)	(0.0022)	(0.0022)
2001-2002	-0.0054***	-0.0053	-0.0046	-0.0036	-0.0035*
N=3217	(0.0018)	(0.0052)	(0.0057)	(0.0025)	(0.0020)
2002-2003	-0.0042	-0.0036	-0.0026	-0.0057**	-0.0032
N=3314	(0.0032)	(0.0065)	(0.0088)	(0.0027)	(0.0020)
2003-2004	-0.0043	-0.0139***	-0.0113*	-0.0119***	-0.0032*
N=3985	(0.0031)	(0.0050)	(0.0058)	(0.0026)	(0.0019)
2004-2005	-0.0123***	-0.0194***	-0.0202***	-0.0141***	-0.0107***
N=4791	(0.0020)	(0.0045)	(0.0048)	(0.0022)	(0.0021)
2005-2006	-0.0179***	-0.0296***	-0.0287***	-0.0211***	-0.0172***
N=5167	(0.0022)	(0.0063)	(0.0053)	(0.0028)	(0.0028)
2006-2007	-0.0191***	-0.0255***	-0.0242***	-0.0220***	-0.0177***
N=5260	(0.0021)	(0.0036)	(0.0050)	(0.0025)	(0.0023)
2007-2008	-0.0136***	-0.0068	-0.0065	-0.0077**	-0.0157***
N=4829	(0.0029)	(0.0069)	(0.0058)	(0.0032)	(0.0038)
Fixed Effects					
Student	No	Yes	Yes	Yes	No
School	Yes	No	Yes	No	No
Grade	Yes	Yes	Yes	Yes	No
Year	Yes	Yes	Yes	Yes	No
School-Grade-Year	No	No	No	No	Yes
Observations	2,761,971	2,761,971	2,761,971	1,336,719	2,761,971
R-Squared	0.613	0.377	0.381	-	0.624

Standard errors clustered at the District level in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Appendix Table 4: Alternative Fixed Effect Combinations

	(1)	**(2)**	**(3)**
Prior Score	*Lag*	*Lag*	*Lag*
Estimator	*OLS*	*OLS*	*OLS*
Entry Cohort			
2000-2001	0.00501*	0.00425	0.00456
N=3957	(0.00286)	(0.00300)	(0.00292)
2001-2002	-0.00202	-0.00320	-0.00177
N=3023	(0.00361)	(0.00291)	(0.00336)
2002-2003	-0.00765***	-0.0106***	-0.00764***
N=3171	(0.00214)	(0.00276)	(0.00269)
2003-2004	-0.0154***	-0.0156***	-0.0157***
N=3719	(0.00508)	(0.00432)	(0.00476)
2004-2005	-0.0206***	-0.0216***	-0.0215***
N=4497	(0.00431)	(0.00415)	(0.00448)
2005-2006	-0.0278***	-0.0281***	-0.0302***
N=4714	(0.00257)	(0.00234)	(0.00237)
2006-2007	-0.0308***	-0.0315***	-0.0321***
N=4882	(0.00403)	(0.00378)	(0.00405)
2007-2008	-0.0279***	-0.0250***	-0.0271***
N=4463	(0.00406)	(0.00438)	(0.00444)
Fixed Effects			
School	No	No	Yes
Grade	Yes	No	No
Year	No	Yes	No
School-Year	Yes	No	No
School-Grade	No	Yes	No
Grade-Year	No	No	Yes
Observations	2,752,060	2,752,060	2,752,060
R-Squared	0.663	0.658	0.653

Standard errors clustered at the District level in parentheses

*** p<0.01, ** p<0.05, * p<0.1

B Measuring Teacher Quality

The purpose of value-added models (VAMs) is to separate the portion of student growth attributable to particular teachers from the many other possible sources of growth. Viewed in this light, the challenges of VAM estimation are those faced in identifying causal relationships with panel data more generally. VAM estimation has proven to be difficult in non-experimental settings and there is no consensus on what the best model of student achievement is or the best approach to estimating the portion attributable to teachers (McCaffrey et al. 2004; Kane & Staiger 2008, Rothstein 2009, 2010; Koedel & Betts 2011). Much of this difficulty stems from the non-random assignment of students to teachers both within and across schools.

The following discussion draws heavily from prior work on the assumptions applied to the education production function underlying VAM estimation (Todd & Wolpin 2003; Harris, Sass, & Semykina 2011; Guarino, Reckase, & Wooldridge 2015). This discussion should be thought of as a guide for considering the issues that arise in VAM estimation, rather than outlining a more formal structural model of education production to be estimated. The starting point for the value-added framework is a very general model that specifies a student's achievement in a particular year as a function of both current and past inputs to the education process and the student's unobserved ability:

$$A_{it} = f_t(X_{it}, \dots, X_{i0}, E_{it}, \dots, E_{i0}, c_i, u_{it}) \tag{B.1}$$

where

A_{it} is the achievement of student i in year t

X_{it} is a vector of family and student characteristics for student i in year t

E_{it} is a vector of education inputs for student i in year t

c_i is unobserved student ability

u_{it} is an idiosyncratic shock to student *i's* achievement in year t

Here, the vector E_{it} can be thought to include indicators for individual teachers or groups of teachers. Given computational and data constraints, several assumptions are typically made to yield a tractable estimating equation. First it is assumed that f_t is linear and constant across years:

$$A_{it} = \alpha_t + X_{it}\beta_0 +, \ldots, + X_{i0}\beta_t + E_{it}\gamma_0 + \ldots + E_{i0}\gamma_t + \eta_t c_i + u_{it} \tag{B.2}$$

Typically, researchers do not have complete data on all prior inputs. To address the lack of prior inputs, it is common to add and subtract λA_{it-1} to the right hand side of (B.2). Assuming that the effect of the inputs decays at a geometric rate equal to λ and that $\eta_t - \lambda\eta_{t-1}$ is a constant (set to equal one without loss of generality) allows us to eliminate the lagged inputs and rewrite equation (B.2) as a function of current inputs and lagged achievement only:

$$\begin{aligned} A_{it} &= \zeta_t + \lambda A_{it-1} + X_{it}\beta_0 + E_{it}\gamma_0 + c_i + e_{it} \\ e_{it} &= u_{it} - \lambda u_{it-1} \end{aligned} \tag{B.3}$$

Up to now, the assumptions made on the original model in equation (B.1) have been primarily data-driven. At this point, there is some choice over further assumptions imposed on the model. Under the assumptions that e_{it} is serially uncorrelated and that c_i is uncorrelated with the included inputs (or equal to zero),[32] equation (B.3), referred to as the lag score equation from here on, could be reasonably estimated by OLS.[33] While the no-serial-correlation assumption is by no means trivial, the assumption that c_i is uncorrelated with the inputs is perhaps the most questionable. It seems possible, given non-random sorting of students and teachers into schools, as well as non-random assignment of students to teachers within schools, that the student unobserved ability may be correlated with teacher assignment. Despite these concerns, there is evidence that this approach may be preferred and so it will serve as the basis for the main analysis in this paper.

As a sensitivity check, we also consider other value-added models and es-

[32] This condition would hold if $\lambda \approx 1$ and $\eta_t \approx \eta_{t-1}$

[33] Note that prior achievement is also a function of the unobserved student heterogeneity term, and is therefore endogenous in (7.3) when c_i is not zero and ignored. This certainly leads to inconsistent estimates of λ, but the extent to which this bias is propagated in the estimated teacher effects is unclear.

timators. Briefly, it is also common to assume that $\lambda = 1$, and to subtract A_{it-1} from both sides of equation (B.3), yielding a gain score model of student achievement:

$$\begin{aligned} \Delta A_{it} &= \zeta_t + X_{it}\beta_0 + E_{it}\gamma_0 + c_i + \nu_{it} \\ \nu_{it} &= u_{it} - u_{it-1} \end{aligned} \tag{B.4}$$

Equation (B.4) could then be estimated by OLS or fixed effects (FE).[34] OLS estimation of (B.4) relaxes the need for no serial correlation in the errors at the cost of assuming the prior achievement persists completely in determining current achievement. If $\lambda \neq 1$, then this approach effectively introduces an additional term, $(\lambda - 1)A_{it-1}$, on the right hand side of equation (B.4), which may lead to an omitted variables bias. Importantly, OLS on (B.4) does not control for the unobserved student heterogeneity in any way.

FE estimation is particularly appealing, as it relaxes the assumption that c_i is uncorrelated with the inputs. However, FE requires the additional assumption that X_{it} and E_{it} are strictly exogenous conditional on c_i in (B.4) for consistent estimation. The strict exogeneity assumption essentially implies that the inputs in time t are uncorrelated with the unobserved error terms in every time period.[35] Practically speaking, the strict exogeneity assumption precludes any feedback from realized achievement shocks to future inputs. For instance, if a principal reacts to a randomly good or bad test score in one year when determining a future teacher assignment, this would violate strict exogeneity. As noted by Rothstein (2009, 2010), the fixed effects approach is useful when assignment to teachers is made based on a static characteristic of

[34] In the panel data context, the gain score equation is also commonly estimated using an Empirical Bayes shrinkage estimator (Kane & Staiger, 2008). Note that the shrinkage factor is determined by the number of observations per group and tends toward one as the group size becomes large. Since in our preferred specification the groups size is quite large and is similar across all groups, the Empirical Bayes estimator will yield results very similar to OLS.

[35] Note that the strict exogeneity assumption is what precludes the use of fixed effects on the lag score equation as well. The lag score equation necessarily violates strict exogeneity by including the lagged dependent variable as a regressor since A_{it-1} must be correlated with the error term in period *t-1*.

the student. The usefulness of FE estimation breaks down some when assignment decisions are made dynamically based on new information gathered over time by the relevant decision makers, be it principals, parents, or the students.

Finally, it has become more common to estimate teacher value-added using approaches based on the dynamic GMM estimator found in Arellano & Bond (1991) (see Koedel & Betts 2011). Researchers taking this approach either use the Arellano & Bond GMM estimator, or a 2SLS version based on identical moment conditions, here referred to as the First-Differenced Instrumental Variables (FDIV) estimator.[36] Specifically, a first-differenced version of the lag score equation (B.3) is estimated using twice-lagged test scores as an instrument for the lagged gain score. This estimator directly addresses the presence of c_i in (B.3) through the first-differencing while also avoiding the problem that including lagged achievement violates strict exogeneity with the use of instrumental variables. Importantly, this approach still requires strict exogeneity of the other regressors. While this assumption could be relaxed by using lagged regressors as instruments, as is done for prior achievement, this has not been common in the value-added literature. Most importantly, the Arellano & Bond-inspired approach requires that the errors in (B.3) not be serially correlated for twice lagged achievement to be a valid instrument. Finally, these approaches require an additional year of data for each student, thereby reducing the sample with which teacher value-added can be calculated.

C CSR Effect Estimates

Here, we estimate the CSR policy effect within the framework discussed in section 4. These results will complement a prior paper on CSR effects in State X to confirm that it fell short of the potential experimental gains from reducing class size for the sample and model used here. Specifically, equation (4.1) is adapted by replacing the cohort indicators, teacher experience, and class size

[36] The GMM and FDIV approaches are identical if the optimal GMM weighting matrix is replaced by an identity matrix.

variables with CSR treatment-by-year indicators:

$$\begin{aligned} A_{igst} =& \zeta_t + \lambda A_{igst-1} + X_{igst}\beta + (T \times Year_{st})\gamma_1 + \gamma_2 \overline{A}_{-igst-1} + \phi_g \\ &+ c_i + \delta_s + e_{igst} \end{aligned} \tag{C.1}$$

Two separate regressions are estimated based on school- or district- level CSR enforcement. For the district-level enforcement, treatment T equals 1 for districts that were above the new class-size maximum in the year before CSR, and 0 otherwise. The school-level treatment status is similarly determined by the school average class size the year prior to school-level enforcement. It is important to note that the regressions include year and school dummy variables and the omitted treatment category is for the 2001-2002 school year.[37] Table C1 presents the estimates of equation (C.1) for district- and school-level CSR with district-enforcement years shaded light gray and school-enforcement years in dark gray. Note that these regressions use test scores standardized within grade and year as the dependent variable. Beginning with the district-CSR results, most of the estimated CSR achievement effects are small and not statistically different from either zero or the estimated pre-CSR treatment-year interaction coefficient (*T x 2002-2003*). The one exception is the 2004-2005 effect, estimated to be a statistically significant 0.0264 standard deviations. While statistically significant, the point estimate is practically small. As a rough point of comparison, a simple prediction of the potential effect of CSR based on the STAR estimates of Krueger (1999) would be on the order of one-eighth of a standard deviation.[38] Even the ninety-five percent confidence

[37] While the data includes two pre-policy years, perfect collinearity between the treatment-year interactions and the school fixed effects requires omitting the 2001-2002 school year treatment interaction. Importantly, when we shift our focus to estimating cohort effects, we can identify three pre-policy cohorts.

[38] Krueger estimates the small class effect in third grade (the closest grade to those considered here) to be roughly one-fifth of a standard deviation. This corresponds to an average difference in class-size of eight students, from 24 to 16. State X's average class-size change in fourth through eighth grade was five students, from 24 to 19. Assuming a linear effect of class-size, the Krueger estimates from Tennessee suggest an effect of one-fortieth of a standard deviation per student which gives the simple prediction of one-eighth. This Tennessee STAR Benchmark can be thought of as a rough guide for assessing CSR and cohort performance. While it is not clear what magnitude of achievement effects would constitute

Table C1: Estimated CSR Mathematics Achievement Effects for State X

CSR Level	*District*	*School*
Tx2002-2003	-0.0170	-0.0323
	(0.0180)	(0.0244)
Tx2003-2004	0.0163	-0.0284*
	(0.0152)	(0.0143)
*Tx2004-200*5	0.0264**	-0.00604
	(0.0125)	(0.0102)
Tx2005-2006	0.00902	-0.0459***
	(0.0183)	(0.0164)
Tx2006-2007	-0.00522	-0.0410*
	(0.0186)	(0.0231)
Tx2007-2008	0.00915	-0.0273
	(0.0156)	(0.0216)
Observations	2,752,060	2,716,399
R-squared	0.653	0.653

Cluster robust standard errors in parentheses;
District (school) level for district (school) CSR
*** p<0.01, ** p<0.05, * p<0.1

intervals for these estimates fall short of half of the rough Tennessee STAR benchmark.

As shown by the results in the last column of Table C1, the treatment-by-year effects after the switch to school-level enforcement during the 2006-2007 school year are negative. The interpretation of these results is made more difficult by the fact that there are also statistically significant negative CSR achievement effects estimated prior to the switch to school-level enforcement. One potential explanation is that those schools farthest from meeting the class-size requirements in 2006-2007 were forced to allocate more resources to class-size reduction in anticipation of the switch in enforcement.

The results found in Table C1 generally concur with those found in State X in a prior paper using similar data and treatment definitions, but employing a Comparative Interrupted Time Series estimation approach. Both suggest,

a successful CSR policy, having an external, experimental comparison is preferred to simply testing for statistically significant estimates.

at most, small positive effects of CSR when treatment is defined by pre-CSR district level class-size averages and potentially negative effects for estimates based on school-level treatment status. A full investigation of the potential issues in estimating CSR effects in State X is beyond the scope of this paper. It is reassuring that the approach adopted here yields roughly similar results to the previous paper on CSR achievement effects in State X. Importantly the evidence here and in the prior work allow for the possibility that the average quality of the newly hired teachers may have affected the performance of the policy compared to the experimental results.

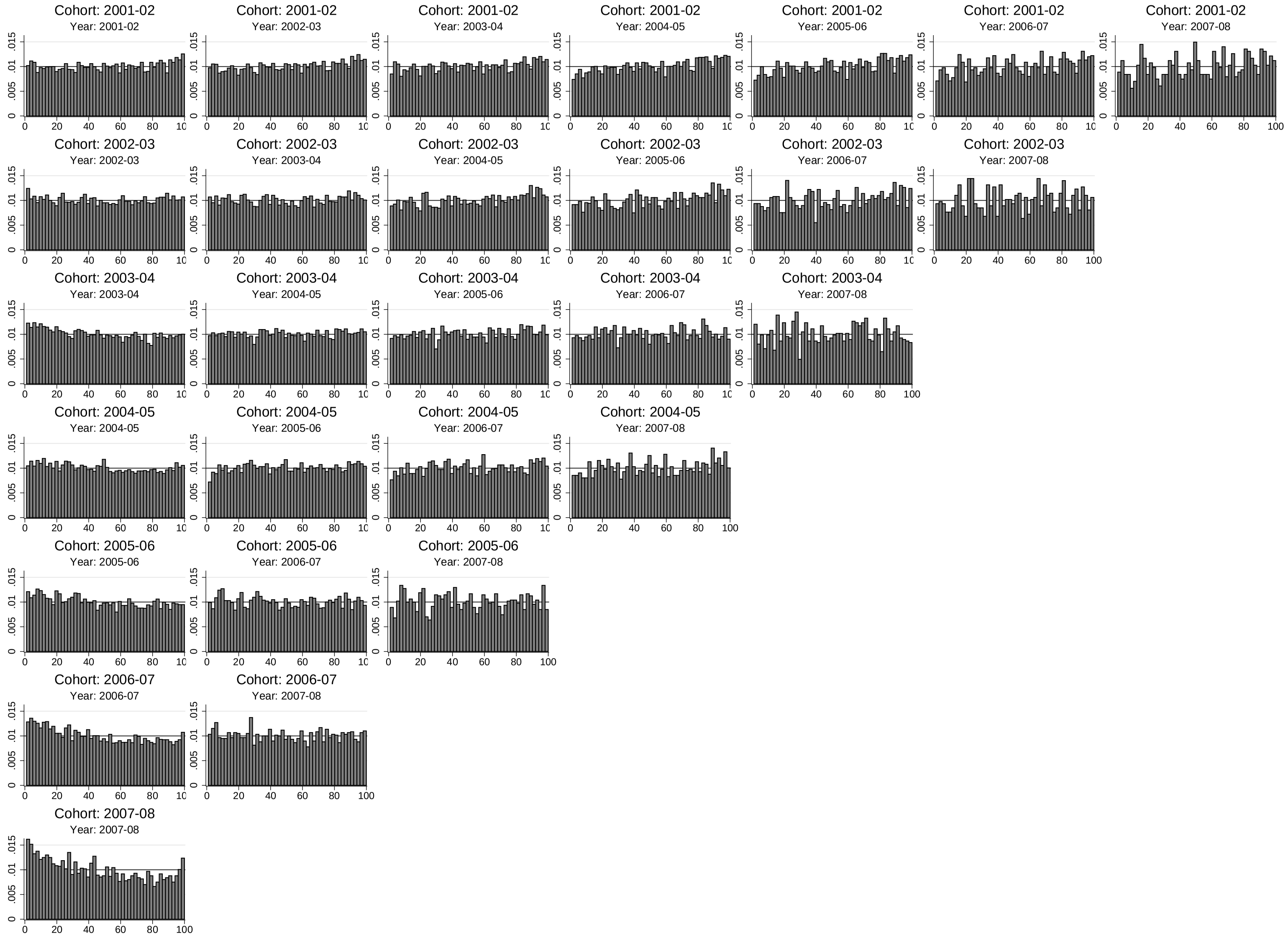
D Individual Teacher Value-added

The main estimates found in the paper identify changes in mean cohort performance. To allow for a comparison of the entire distribution of teacher quality over time, individual teacher value-added is also estimated. We do this two ways. First estimating constant effects for each teacher over time by replacing the cohort indicators in (4.1) with indicators for each teacher. We similarly estimate teacher-by-year effects to capture potentially different patterns of human capital growth. We present results for the teacher-by-year effects here as they more directly relate to the year-to-year effect on students by each cohort. Teachers are given a percentile rank based on their estimated value-added relative to all the teachers in the sample. Figure D1 displays histograms of the distribution of teacher percentile ranks for each entry cohort by year. The solid line on each graph represents a uniform distribution of percentile ranks (i.e., the distribution for a cohort if a given teacher from that cohort was equally likely to be ranked anywhere in the overall distribution). Starting in each cohorts first year, the percentile rank distribution of the post-CSR cohorts (2003-04 to 2007-08) show many more teachers at the low end of the distribution than the pre-CSR cohorts (2001-02 and 2002-03).[39] Indeed, the pre-CSR cohort first year distributions are nearly uniform while the post-CSR cohorts

[39] The 2000-01 cohort has been excluded since we can only estimate value-added starting with the second year for this cohort.

show a higher probability of being in the lowest quintile. As the post-CSR cohorts evolve over time, due to human capital growth and attrition we see that these cohort differences weaken, reflecting the mean cohort effect results.

Figure D1: Value-added Percentile Distributions



E Cohort Effects by CSR Pressure

The estimates of equations (4.1) can be thought of as identifying the statewide general equilibrium relationship between hiring cohorts and student performance. However, it is possible that the CSR policy had more bite in schools farther away from the new class-size maximums. In fact, the hypothesis that changes in teacher quality can explain CSR performance is based on this notion. To be consistent with the teacher quality hypothesis, we would need to see teacher quality fall more for those districts and schools which were considered treated in the prior CSR effect estimates based on pre-policy class-sizes. Table E1 shows the estimates from specifications in which the entry cohorts are further divided based on the amount of CSR pressure the school was under. This grouping is done based on both the district averages prior to CSR and the school averages prior to the change to school-level enforcement. Those schools already below the maximums are included in the None group while the remaining schools are divided into quartiles based on average class size. Starting with the district groupings, the estimates show that across the board all schools saw a decline in the performance of new teachers over the implementation of CSR. Importantly, it is not the case that the estimated effects are monotonically increasing in magnitude with increases in CSR pressure. Taken together, it appears that CSR-induced hiring did not just impact the quality of new teachers for schools originally above the new class-size maximums. Rather it suggests either a more general trend in cohort quality or that the untreated schools were still forced to move along the effective teacher supply curve as candidates they may have otherwise hired to fill openings created by turnover and enrollment growth were hired by nearby schools facing CSR pressure.

Similarly, the results for the school-level disaggregation do not consistently tell a story that CSR lowered incoming teacher quality disproportionately for treated schools. One exception, however, is in the year before school-level enforcement for those schools farthest from reaching the new maximums (Q4). These schools, which were likely pre-empting the switch to school-level enforcement in the following year, had a hiring cohort estimated to be 0.0617

Table E1: Estimates of New Cohort Effects by CSR Intensity

CSR Intensity	*None*	*Q1*	*Q2*	*Q3*	*Q4*
Entry Cohort			*District Enforcement*		
2000-2001	-0.0020	0.0042	0.0122*	0.0038	0.0051***
	(0.0069)	(0.0076)	(0.0062)	(0.0071)	(0.0008)
2001-2002	-0.0052	-0.0030	0.0022	-0.0159	0.0062***
	(0.0077)	(0.0041)	(0.0091)	(0.0106)	(0.0013)
2002-2003	-0.0153***	0.0045	-0.0001	-0.0183***	-0.0192***
	(0.0055)	(0.0086)	(0.0049)	(0.0044)	(0.0028)
2003-2004	-0.0252***	-0.0191	-0.0151***	-0.0168*	0.0049**
	(0.0063)	(0.0121)	(0.0045)	(0.0086)	(0.0020)
2004-2005	-0.0229***	-0.0285***	-0.0159**	-0.0372***	-0.0056***
	(0.0049)	(0.0056)	(0.0064)	(0.0096)	(0.0015)
2005-2006	-0.0322***	-0.0235***	-0.0329***	-0.0273***	-0.0334***
	(0.0050)	(0.0074)	(0.0067)	(0.0039)	(0.0028)
2006-2007	-0.0390***	-0.0172**	-0.0212***	-0.0667***	-0.0226***
	(0.0071)	(0.0078)	(0.0076)	(0.0076)	(0.0041)
2007-2008	-0.0358***	-0.0389***	-0.0247***	-0.0162	-0.0076***
	(0.0083)	(0.0060)	(0.0066)	(0.0126)	(0.0024)
Observations			2,752,060		
R-Squared			0.0653		
Entry Cohort			*School Enforcement*		
2000-2001	0.0061	-0.0143*	0.0046	0.0005	0.0065
	(0.0039)	(0.0075)	(0.0115)	(0.0121)	(0.0075)
2001-2002	-0.0078*	-0.0137	-0.0150	0.0056	0.0498***
	(0.0046)	(0.0181)	(0.0107)	(0.0113)	(0.0056)
2002-2003	-0.0067*	-0.0214*	-0.0197*	-0.0072	-0.0122
	(0.0036)	(0.0113)	(0.0114)	(0.0140)	(0.0078)
2003-2004	-0.0220***	-0.0159	-0.0046	0.0043	0.0166*
	(0.0044)	(0.0100)	(0.0125)	(0.0164)	(0.0087)
2004-2005	-0.0219***	-0.0107	-0.0373	-0.0178	-0.0200*
	(0.0045)	(0.0113)	(0.0231)	(0.0126)	(0.0113)
2005-2006	-0.0273***	-0.0339**	-0.0261**	-0.0219**	-0.0615***
	(0.0036)	(0.0135)	(0.0117)	(0.0085)	(0.0037)
2006-2007	-0.0302***	-0.0345***	-0.0501***	-0.0196*	-0.0373***
	(0.0050)	(0.0065)	(0.0118)	(0.0103)	(0.0077)
2007-2008	-0.0306***	-0.0320*	-0.0203	0.0040	-0.0158*
	(0.0048)	(0.0184)	(0.0166)	(0.0144)	(0.0080)
Observations			2,752,060		
R-Squared			0.0653		

Standard errors clustered at the District level in parentheses
*** p<0.01, ** p<0.05, * p<0.1

test score standard deviations worse than the baseline teachers, while the other schools saw cohorts between 0.0219 and 0.0326 standard deviations worse. These results by CSR pressure cast doubt on the teacher quality hypothesis.[40]

F New Teacher Contribution to Mean Achievement

To summarize the contribution of cohort quality and experience on overall student achievement, we calculate $\overline{COHORT}_t\hat{\gamma}_1$ and $\overline{\hat{f}(EXP_t)} = \overline{EXP_t}\hat{\beta}_1 + \overline{EXP_t^2}\hat{\beta}_2 + \overline{EXP_t^3}\hat{\beta}_3$. Both the total contribution and the separate contribution of each component are presented in Table F1, along with the change since 2001. While the contribution attributable to these components falls over the introduction of CSR, even in the worst year this represents only a difference of 0.0154 standard deviations. This difference is driven more by the relative performance of the cohorts than by the drop in teacher experience.[41]

[40] Using a similar approach, disaggregating the entry cohorts by quartiles of school-level mean student characteristics (free or reduced lunch status, Black, or Hispanic) yields similarly mixed results with no clear evidence that schools serving more disadvantaged students saw disproportionately worse hiring cohorts.

[41] Recall that the experience profile can be thought to capture the effects of differential attrition and within school sorting of students to more experienced teachers, in addition to human capital accumulation.

Table F1: Estimated Contribution of Cohort Composition and Experience to Average Achievement

	Achievement Contribution			**Change from 2001-2002**		
Year	$\overline{COHORT_t}\widehat{\gamma}_1$	$\overline{\widehat{f}(EXP_t)}$	*Total*	$\overline{COHORT_t}\widehat{\gamma}_1$	$\overline{\widehat{f}(EXP_t)}$	*Total*
2001-2002	-0.0059***	0.0452***	0.0393***	-	-	-
	(0.0011)	(0.0052)	(0.0051)	-	-	-
2002-2003	-0.0066***	0.0452***	0.0386***	-0.0007	0.0000***	-0.0007*
	(0.0011)	(0.0052)	(0.0051)	(0.0004)	(0.0000)	(0.0004)
2003-2004	-0.0082***	0.0448***	0.0366***	-0.0024***	-0.0004***	-0.0028***
	(0.0015)	(0.0051)	(0.0048)	(0.0008)	(0.0001)	(0.0008)
2004-2005	-0.0107***	0.0443***	0.0335***	-0.0048***	-0.001***	-0.0058***
	(0.0019)	(0.0051)	(0.0045)	(0.0013)	(0.0001)	(0.0013)
2005-2006	-0.0142***	0.0442***	0.0299***	-0.0084***	-0.001***	-0.0094***
	(0.0017)	(0.0051)	(0.0047)	(0.0010)	(0.0001)	(0.001)
2006-2007	-0.0176***	0.0434***	0.0258***	-0.0117***	-0.0018***	-0.0135***
	(0.0021)	(0.0050)	(0.0045)	(0.0013)	(0.0002)	(0.0014)
2007-2008	-0.0195***	0.0434***	0.0239***	-0.0136***	-0.0018***	-0.0154***
	(0.0024)	(0.0050)	(0.0047)	(0.0017)	(0.0002)	(0.0017)

Standard errors clustered at the District level in parentheses
*** p<0.01, ** p<0.05, * p<0.1

G Comparison to North Carolina Teacher Data

Ideally, we would like to have a clean comparison state to use as control group when estimating the potential hiring spillover effects of CSR. Unfortunately, given differences across states in institutional and economic factors coupled with the finding in the literature that teachers tend to focus job search in small geographic areas, such comparisons may not be very clean. To explore this, we use publicly available state-wide data from North Carolina, a state that does have similar data to State X over the time period we study. While we cannot make direct comparisons to all the data series presented in Section 3, there are three comparable series that can be matched in the following table.

Appendix Table G1:State X and North Carolina Teacher Characteristics

	State X Grades 4-6			**North Carolina All Grades**		
	No Prior Experience		*Advanced Degree*	*No Prior Experience*		*Advanced Degree*
	Percent of Stock	Percent of Flow	Percent of Stock	Percent of Stock	Percent of Flow	Percent of Stock
School Year	(1)	(2)	(3)	(4)	(5)	(6)
2000-01	18.15%	39.88%	31.75%	6.67%	67.13%	36.40%
2001-02	17.63%	42.55%	31.89%	6.25%	64.06%	36.21%
2002-03	19.35%	45.45%	33.87%	5.68%	62.62%	36.36%
2003-04	17.44%	47.81%	33.12%	6.05%	62.43%	36.79%
2004-05	18.49%	46.61%	33.27%	6.51%	63.52%	36.00%
2005-06	19.45%	46.85%	34.00%	6.47%	61.79%	35.77%
2006-07	17.87%	45.52%	33.23%	6.11%	60.81%	36.58%
2007-08	16.07%	42.58%		6.40%	62.09%	

As we can see, State X and North Carolina are quite different over this time period, particularly when looking at teacher experience. State X had a much larger proportion of completely new teachers in the stock than North Carolina (column (1) versus column (4)), but a much smaller proportion of the teachers seen entering State X public schools were complete novices than in North Carolina (column (2) versus column (5)). While this suggests that North Carolina may not be a clear control state, it does point to the idea that the size of the year-to-year fluctuations in State X might be in line with the idiosyncratic year-to-year movements in these variables. Of course, this analysis does have not consider any relevant policy changes may have occurred in North Carolina over this period. More generally, this comparison does not tell us how value-added may compare across hiring cohorts int he two states.