Edinburgh School of Economics

Discussion Paper Series

Number 224

Class-size Reduction Policies and the Quality of Entering Teachers

Steven Dieterle (University of Edinburgh)

Date
February 2013

Published by
School of Economics
University of Edinburgh
30 -31 Buccleuch Place
Edinburgh EH8 9JT
+44 (0)131 650 8361
http://edin.ac/16ja6A6

THE UNIVERSITY *of* EDINBURGH

Class-size Reduction Policies and the Quality of Entering Teachers

Steven Dieterle*
Email: steven.dieterle@ed.ac.uk
University of Edinburgh

February 21, 2013

Abstract

State-wide class-size reduction (CSR) policies have typically failed to produce large achievement gains. One explanation is that the introduction of such policies forces schools to hire relatively low-quality teachers. This paper uses data from an anonymous state to explore whether teacher quality suffered from the introduction of CSR. We find that it did, but not nearly enough to explain the small achievement effects of CSR. The combined fall in achievement due to hiring lower quality teachers and more inexperienced teachers is small relative to the unrealized gains. Furthermore, between-school differences in the quality of incoming teachers cannot explain the poor estimated CSR performance from previous quasi-experimental treatment-control comparisons.

*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 and Notre Dame University for helpful comments. 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 over the last two decades (Krueger 1999; Krueger & Whitmore 2001; Angrist & Lavy 1999). 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. Due to the high costs of implementation, $21 billion over nine years in Florida (Florida Department of Education) and of $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 new teachers of lower quality in order to meet the class-size requirements. 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, 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 is 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 California's introduction of CSR, this data still does not include any pre-policy test score measures. Identification using the school average data also relies on observed teacher characteristics in order to estimate changes in teacher quality (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)[1] covering 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 the CSR-induced demand increase did in fact lead to schools hiring

[1]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.

and retaining lower quality teachers, here measured by value-added to student mathematics achievement.[2] In doing so, we exploit this sudden increase in teacher demand to inform the literature on an understudied, yet important, feature of teacher labor markets, the value-added elasticity of teacher supply. Finally, we consider the implications of our findings for interpreting prior quasi-experimental estimates of CSR achievement performance in terms of the teacher quality hypothesis.

The value-added estimates of cohort performance found here 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 post-CSR hiring cohorts ranges from 0.0033 to 0.0277 test score standard deviations lower than the smaller pre-CSR cohorts in each cohort's first year. These differences in cohort performance 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 0.0078 to 0.0192 standard deviations. However, there is evidence that further attrition for post-CSR hiring cohorts may lead to negligible differences among the remaining teachers after three to four years, implying an even smaller long-run CSR hiring effect on achievement. It is important to note that these results are robust to several estimation approaches.

Even if the average quality of cohorts had not changed, there may have been an additional short-run effect of CSR hiring on student performance due to hiring more teachers with less experience. The fall in average achievement attributable to the change in both average quality and experience is less than one-fiftieth of the test score standard deviation. This fall in achievement is driven primarily by changes in cohort quality, rather than experience. Importantly for the teacher quality hypothesis, this drop in quality was generally faced by all schools. In fact, schools classified as treated (those for which CSR was binding) in previous quasi-experimental estimates of CSR policy effects in State X experience a slightly smaller drop in achievement attributable to the stock of teachers than those considered untreated. This difference is of the opposite sign needed to support the teacher quality hypothesis. Further, it suggests a role for competition for teacher candidates pushing all schools along the effective teacher supply curve in connected labor markets.

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.

[2]Similar results obtained using reading test scores are available upon request from the author. Generally, the reading results are slightly smaller in magnitude and were slightly more sensitive to the specification and estimator chosen. However, these differences do not change the conclusions drawn. The decision to focus on mathematics scores only was made for the sake of brevity and due to the fact that it is common in the education production function literature for mathematics scores to be more responsive to inputs than reading.

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 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 provides a review of the relevant literature and background information, section 3 discusses the institutional details of the policy, section 4 discusses the data used, section 5 looks at the market for teachers in State X around CSR implementation, section 6 discusses the empirical strategy used throughout, section 7 confirms prior CSR effect estimates for State X, section 8 gives and discusses the baseline results and checks the sensitivity of these results, section 9 gives the preferred estimates that account for teacher attrition, section 10 considers the implications of our findings for interpreting prior quasi-experimental CSR effect estimates, and section 11 concludes.

2 Literature Review and Background

Based on the random assignment of students and teachers to classrooms of varying sizes, the results of the Tennessee STAR experiment suggested that class-size reduction is a potentially viable tool to promote achievement gains. Krueger (1999) analyzes STAR and finds that being randomly assigned to a small (13-17 students) class as opposed to a larger class (22-25 students) in early elementary school led to roughly one-fifth of a standard deviation increase in average test scores. In a follow-up, Krueger & Whitmore (2001) find that being in a small class also impacted student outcomes well after the experiment, such as increasing the likelihood of taking a college entrance exam. More recent work by Chetty et al. (2011) suggests that the benefits of being assigned to a small class in STAR even persist into the adult labor market

The positive achievement effects from Tennessee STAR led many states to explore the use of CSR to promote student achievement growth. By 2005, thirty-two states had adopted some sort of CSR program (CEPRI 2005). Despite CSR's popularity among teachers and

parents, there is only mixed support for the conclusion that these large-scale programs are effective at helping to raise test scores. In their official report on CSR in California, Bohrnstedt & Stecher (2002) were unable to find conclusive evidence of achievement gains for kindergarten through third grade. In contrast, Jepsen & Rivkin (2009) use class-size variation from California CSR and find that a ten student reduction in class size is associated with an increase in achievement of one-tenth to one-twentieth of a standard deviation in grades two through four. Like Bohrnstedt & Stecher, Chingos (2012) found null effects for fourth through eighth grade of CSR in Florida. We will discuss policy effect estimates from State X in more detail in section 7, however a prior paper has similarly found no evidence of positive achievement effects from the policy.

Assuming that there are potential gains from reducing class size, a leading explanation for the failure of CSR revolves around changes in teacher quality associated with the implementation of the program (Stecher & Bohrnstedt 2000; Imazeki n. d.; Buckingham 2003; CEPRI 2005, Chingos 2012).[3] In interpreting his results, Chingos suggests that factors "such as reduced teacher quality" may help explain his findings. One way in which teacher quality may change is if schools are forced to hire additional teachers from lower on the quality distribution in order to meet the new class-size requirements. Schools may also retain teachers that would otherwise have been dismissed for poor performance to lessen the hiring burden. Gains associated with smaller classes are then offset by having less capable teachers in classrooms, yielding no gains on net.

To support these teacher-quality-based explanations, Stecher & Bohrnstedt (2000) document declines in the percentages of fully certified teachers, teachers with advanced degrees, and experienced teachers in California. While changes in teacher characteristics do indicate changes in the teacher workforce, the link between these characteristics and achievement has often been found to be weak. Goldhaber (2008) provides a detailed review of the education production function literature concluding that teacher quality is not "strongly correlated" with observable teacher characteristics. Therefore, the finding that observable teacher characteristics change after CSR implementation may not adequately explain the lack of test score gains. The more relevant question is whether schools are forced to hire teachers who contribute less to a student's achievement growth.

Jepsen & Rivkin (2009) analyze California's CSR program to estimate the relationship

[3]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.

between teacher cohort size and quality. The authors use data aggregated at the school level, meaning they cannot identify the teachers that make up a hiring cohort or link students to specific teachers. Instead, Jepsen & Rivkin examine whether the estimated effects of school-average teacher experience and certification status differed across years. Intuitively, this approach identifies the quality of new cohorts of teachers because those teachers categorized as inexperienced or uncertified in a given year are more likely to be the new hires. They find no statistically or practically significant differences in the estimated experience or certification effects across years. The finding that much of the variation in teacher quality does not work through observed characteristics makes interpreting these results difficult.

Kane & Staiger (2005) are able to partially address the identification issue discussed above by using individual-level data from Los Angeles to analyze the hiring achievement effect of California's CSR program. They calculate value-added for teachers hired for the 1995-96 school year, just before CSR was introduced, and compare it to value-added for the first CSR cohort hired in 1996-97. They find no differences among the two cohorts of teachers in terms of value-added. However, due to data availability, this comparison can only be made starting with the 1999-2000 school year, four to five years after the teachers were initially hired. Given differential attrition by teachers of varying quality and human capital growth for those teachers that remained, this comparison may miss important short-run effects on student achievement. The analysis may also miss differences between Los Angeles and other districts in the state, making it difficult to conclude how general the results are. Finally, using data from a single district may miss potential hiring spillovers that are important for understanding the unrealized gains from CSR. With data on individual students and teachers for an entire state that spans the introduction of the policy, it is possible to assess the change in teacher quality associated with CSR more directly and overcome many of the limitations of the prior work.

3 Institutional Details: CSR in State X

In November of 2002, State X voters approved a constitutional amendment that created a new 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 found to be in compliance. There is anecdotal evidence from board of education meeting transcripts that the allocation was not enough to cover the full costs of CSR implementation for some districts. This anecdote suggests that a reallocation of other resources may partially explain CSR performance. This possibility will be explored in the

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

Source: State X Department of Education

results section.

The new law allowed for a gradual phase-in of the mandated 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 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.

4 Data

The data used for this 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 time period as the students. 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. 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 by allowing re-entrants to be considered separately from truly new teachers.

Finally, State X has made each district/school's average class size within the three enforcement grade groupings 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. 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.

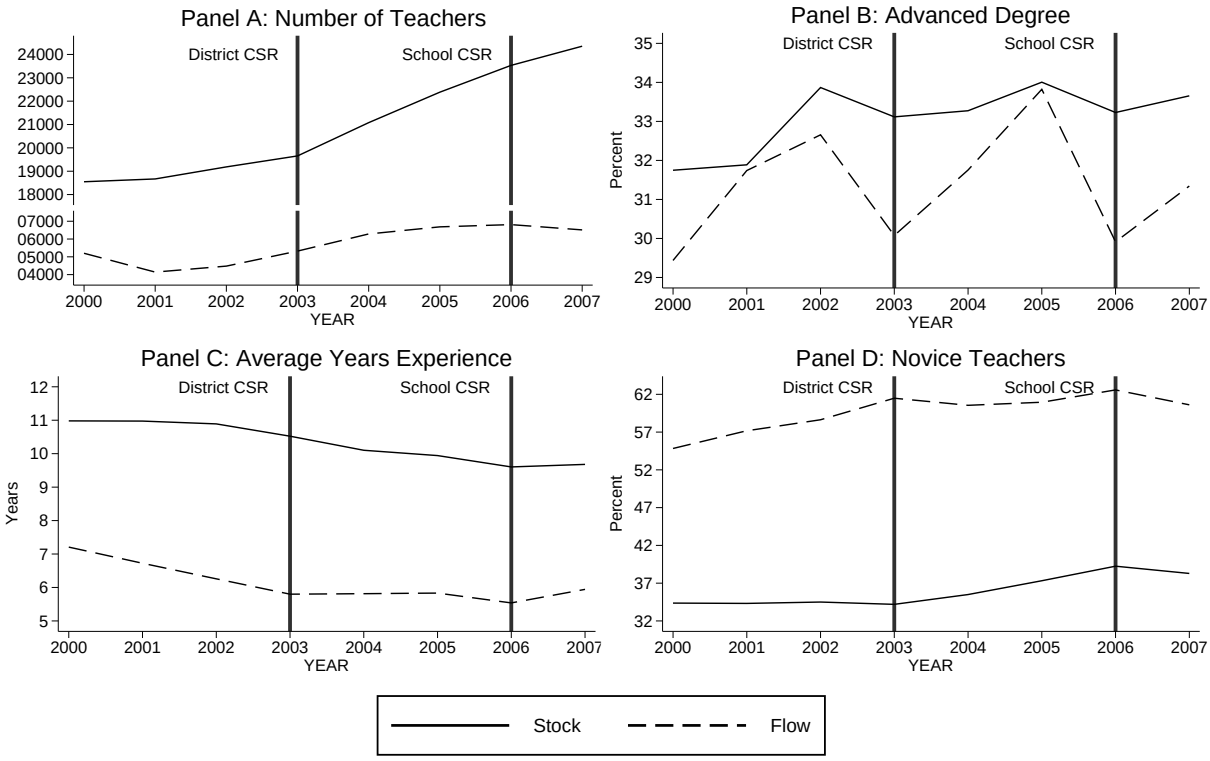
5 Market for Teachers in State X During CSR

Before analyzing the achievement outcomes associated with CSR and the subsequent teacher hiring in State X, it is important to consider the general state of the teacher labor market, as well as any factors that may have led to changes in the supply or demand for teachers over the same time period. Such an 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 the trends in both the stock and flow over time in the number of teachers, percent with an advanced degree, average experience, and percent with three or fewer years of experience. Here, the focus is 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 into teaching 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.

Figure 1: Teacher Stock and Flow Trends

Grades 4-6 in Core Courses



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 accompanied by an increase in the number of teachers entering the data each year of roughly 2,000 by the fourth year of CSR.[4] 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 increasing in the other years. 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. For a more detailed discussion of changes in the observable characteristics of teachers in State X over CSR implementation, see Dieterle (2012).

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 was facing 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

[4]Note the scale break in Panel A.

Table 2: Projected Hiring in State X				
	Year			
Hiring Need	*2004-05*	*2005-06*	*2006-07*	*2007-08*
CSR	4,324	2,378	11,821	974
Enrolment Growth	3,297	3,024	3,134	3,451

subjects from CSR and student enrolment growth, as shown in Table 2. Hiring needs driven by enrolment 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 enrolment growth. The projected difference for the grades studied here are likely to be even more stark, as the numbers in Table 2 include high school grades that were relatively unaffected by CSR. For the years and grades studied here, the student population never saw growth rates above 1.55% and in 2003-04 at the introduction of CSR, actually saw a decrease from the prior year. 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 enrolment growth based hiring it is also likely that the sudden increase in the number of teachers hired shown in Figure 1 was in fact largely due to CSR.[5] 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 enrolment 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 to State X.

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.

[5]Note that the trend in actual hiring may have been smoother than the projected numbers due to pre-emptive hiring.

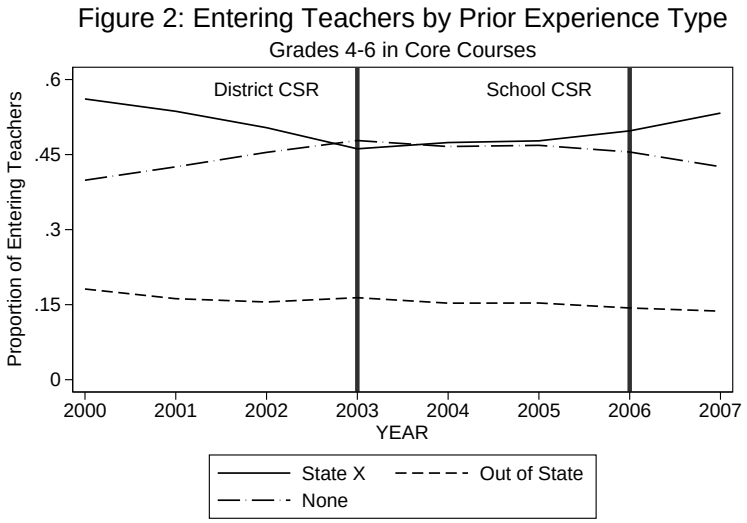


Figure 2 plots the proportion of all entering teachers in the fourth through sixth grade core course sample that have no prior experience, experience in State X, and experience outside of State X. Note that some teachers identified as entering the data will have both in- and out-of-state experience. We see that the proportion of new entrants that have prior out-of-state experience stays roughly level at about 15%. We do see a rise in the proportion with no prior experience and a fall in those with prior experience in the state. The result is an eroding of a pre-policy gap of nearly 15 percentage points in favor of hiring a larger proportion of 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 pose the biggest threat to validity. In particular, over the time period studied, 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). 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.

6 Empirical Methodology

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 in section 8 along with several sensitivity checks and then our preferred estimates that account for teacher attrition are presented in section 9. The baseline specification discussed first provides for a more tractable comparison among the several estimators considered. The intuition presented here for interpreting the results broadly applies to the other estimates as well. The main strategies used here are based on OLS estimation 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:[6]

$$\begin{aligned} A_{igst} =& \zeta_t + \lambda A_{igst-1} + X_{igst}\beta + Cohort_{igst}\gamma_1 + \gamma_2 \overline{A}_{-igst-1} + f(Exp_{igst}) \\ &+ \gamma_3 CS_{igst} + \phi_g + c_i + \delta_s + e_{igst} \end{aligned} \tag{6.1}$$

where

i, g, s, t index student, grade, school, and year

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

A_{igst} is student i's test score
ζ_t are year fixed effects
A_{igst-1} is student i's prior test score
X_{igst} are student demographics[7]
$Cohort_{igst}$ are teacher cohort indicators
$\overline{A}_{-igst-1}$ is the average prior test score of student i's classmates
$f(Exp_{igst})$ is a cubic in teacher experience
CS_{igst} is a proxy measure of class size [8]
ϕ_g are grade fixed effects
c_i is an unobserved student heterogeneity term
δ_s are school fixed effects

Note that the OLS estimation of (6.1) (our preferred strategy) treats c_i as if it were equal to zero for all students. While this assumption may not hold in practice, there is evidence that OLS estimation of the lag score specification typically performs well. Using simulated data, Guarino et al. (2011) 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. Broadly, 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 in section 8, we will argue that such assignment policies are unlikely in practice.

The main coefficients of interest are the estimates of γ_1, the average quality of entry

[7]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 while attending a particular school.

[8]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 (6.1) 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.

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 substantial sorting of teachers into geographically small markets (Boyd et al. 2005; Lankford et al. 2002), each school may face a different level of average teacher 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 inclusion of school fixed effects controls for the time-invariant quality level of teacher supply that different schools face by relying on within school comparisons of teachers. In section 8 we will consider an alternative approach that relies on within school-grade-year variation.

The experience profile can be thought to capture three distinct factors: teaching-specific human capital accumulation, non-random sorting of students to teachers based on experience, and non-random attrition of teachers. Focusing on the human capital piece of the experience profile, 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 more detail later.

Care should be taken in interpreting the estimates of equation (6.1), as State X enacted many policies over the introduction of CSR. Note, however, that changes in state policy that affect all students and 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 policy changes that alter the quality of teachers hired in a particular year and are therefore captured in the estimates of γ_1. As mentioned in the previous section, the expansion of alternative certification pathways represents the most salient threat to the results. However, given the relatively low take up rate of these new options, it is unlikely that the estimates of cohort quality are being driven by this policy.

The approach adopted here captures potential CSR effects that would be difficult to identify given the available data. For example, the school-level class-size averages within the enforcement grade groupings are only available starting with the year directly before school-level enforcement.[9] This data limitation makes it difficult to identify individual schools that may have hired additional 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 teacher value-added measures may also capture changes over time in resources that complement a teacher's ability to raise achievement. If CSR led to a reduction in these resources, then part of the change in measured teacher effectiveness over time may be capturing these changes as well. There is some suggestive evidence, discussed later, that this is not a large problem in interpreting the results. Finally, while the estimation strategies employed here are more susceptible to omitted variables bias than comparable quasi-experimental designs, other approaches that estimate Local Average Treatment Effects or rely on defining treatment groups will tend to be ill-suited for identifying the sort of general equilibrium effects that underpin the teacher quality hypothesis studied here. We will specifically consider the possibility of treatment spillovers in section 10.

7 Confirming Prior CSR Achievement Effect Estimates

Before presenting the baseline results, we estimate the CSR policy effect within the framework discussed in section 6. 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 (6.1) is adapted by replacing the cohort indicators, teacher experience, and class size variables with CSR treatment-by-year indicators:

$$\begin{aligned} A_{igst} =& \zeta_t + \lambda A_{igst-1} + X_{igst}\beta + (T \times Year_{st})\gamma_1 + \gamma_2 \overline{A}_{-igst-1} + \phi_g \\ &+ c_i + \delta_s + e_{igst} \end{aligned} \tag{7.1}$$

Two separate regressions are estimated based on school- or district- level CSR enforce-

[9] 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.

ment. 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 cohort.

Table 3: 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)
Tx2004-2005	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

Table 3 presents the estimates of (7.1) 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.[10] Even the ninety-five

[10]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

percent confidence 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 3, 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 3 generally concur with those found in State X in prior paper using similar data and treatment definitions, but employing a Comparative Interrupted Time Series estimation approach. Both suggest, 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 paper 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.

8 Baseline Estimates and Sensitivity

Table 4 presents the baseline estimates of the cohort effects (γ_1) from equation (6.1) in the first column.[11] Of particular interest are the estimated coefficients on the teacher entry cohort dummy variables. These estimates reflect the conditional mean performance of students in classrooms taught by teachers entering the data in each year, relative to those students in classrooms taught by teachers already in the State X public elementary school system at the beginning of the panel. The policy-relevant comparison is between pre-CSR and post-CSR cohorts. Again 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

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

[11] See Appendix Table 2 for other estimates from these regressions.

been standardized within grade and year.[12] 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 (0.0319-0.00929=0.0226; p-value=0.000)[13] worse than students with a 2002-2003 cohort teacher.

Overall, the estimated post-CSR cohort effects range from 0.0069 (p-value=0.150) to 0.0285 (p-value=0.000)standard deviations lower than the two pre-CSR cohorts.[14] The magnitude of the differences seen in column (1) of Table 4 are small relative to the unrealized CSR achievement gains in State X. Recall that a simple extrapolation of the STAR results would place the expected achievement gain at roughly one-eighth of a standard deviation. By comparison, we estimate a fall in average student performance of, at most, one-thirty-fifth of a standard deviation only in the classes taught by newly hired teachers. That is not to say that the achievement effects for the students in these classrooms was trivial. Rather, when we consider the contribution changes in teacher quality may play in understanding the CSR achievement effect estimates (Section 10), it is unlikely that cohort differences of the size seen here will generate large changes in CSR performance.

In addition to the baseline estimates, we consider two main sensitivity checks. The first is to address the unobserved student heterogeneity term (c_i) found in equation (6.1). Recall that our baseline estimator ignores the presence of c_i, which, loosely speaking, will lead to inconsistent cohort effect estimates if the hiring cohort of a student's teacher is correlated with c_i. 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 (6.1)).[15] Moving forward, it is helpful to distinguish between the fixed effects estimator used to control for unobserved heterogeneity at the individual level, and the inclusion of group level (grade, year, or school) fixed effects. Columns (2) and (3) of Table 4 display cohort effects estimated by FE both excluding and including the school fixed effects, respectively.[16] Finally, we also consider a 2SLS version

[12]There is no agreement on the preferred choice between scale scores and grade-year standardized scale scores. Here, the main conclusions that can be drawn do not differ with this choice. See Reardon & Galindo (2009) for a brief discussion of the two approaches.

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

[14]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

[15]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.

[16]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 without the school effects.

Table 4: Baseline Cohort Effect Estimates and Sensitivity					
	(1)	**(2)**	**(3)**	**(4)**	**(5)**
Prior Score	*Lag*	*Gain*	*Gain*	*Lag*	*Lag*
Estimator	*OLS*	*FE*	*FE*	*FDIV*	*OLS*
Entry Cohort					
2001-2002	-0.0035	0.0028	-0.0011	0.0019	-0.0035
N=2824	(0.0033)	(0.0079)	(0.0076)	(0.0039)	(0.0025)
2002-2003	-0.0093***	-0.0142**	-0.0107	-0.0154***	-0.0102***
N=2856	(0.0025)	(0.0065)	(0.0071)	(0.0035)	(0.0030)
2003-2004	-0.0162***	-0.0273***	-0.0215**	-0.0248***	-0.0148***
N=3378	(0.0047)	(0.0050)	(0.0083)	(0.0026)	(0.0047)
2004-2005	-0.0221***	-0.0365***	-0.0327***	-0.0305***	-0.0198***
N=4037	(0.0046)	(0.0098)	(0.0089)	(0.0050)	(0.0046)
2005-2006	-0.0304***	-0.0442***	-0.0439***	-0.0387***	-0.0254***
N=4247	(0.0024)	(0.0066)	(0.0063)	(0.0042)	(0.0023)
2006-2007	-0.0319***	-0.0434***	-0.0414***	-0.0395***	-0.0293***
N=4492	(0.0045)	(0.0100)	(0.0083)	(0.0052)	(0.0034)
2007-2008	-0.0265***	-0.0166*	-0.0154**	-0.0245***	-0.0241***
N=3390	(0.0047)	(0.0092)	(0.0075)	(0.0049)	(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	2,752,060	2,752,060	2,752,060	1,329,658	2,752,060
R-Squared	0.653	0.399	0.412	–	0.674

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

of the Arellano & Bond (1991) dynamic GMM estimator, referred to here as the First Differenced Instrumental Variables (FDIV) estimator, in order to address the presence of c_i while not constraining $\lambda = 1$.[17]

Comparing columns (1) through (4) of Table 4 shows that each estimator leads to the general conclusion that the post-CSR cohorts have lower estimated 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 the ever-present 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

[17]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.

of the estimates if schools were using some 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. Importantly, the estimators that explicitly control for c_i require additional assumptions that may not be tenable in practice and tend to reduce identifying variation (See Appendix B).

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 baseline estimates. Again, to generate problems, such factors must be related to the student-teacher assignment 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. 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. Our baseline estimates, on the other hand, will compare all teachers hired in different cohorts in the same school ensuring that more classes are contributing to the estimation.[18]

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. Once more, however, this slight change does not alter the conclusion that post-CSR cohorts tend to have lower value-added than pre-CSR cohorts.[19] Motivated by these results and the prior literature

[18]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 and in section 6. Here, this approach leads to a similar conclusion that students in post-CSR cohort classes perform worse.

[19]The results are also invariant to the many potential combinations of year, grade, school, school-year,

discussed above, throughout the remainder of the paper we will estimate variants of (6.1) by OLS controlling for separate grade, year, and school effects.

9 Teacher Attrition and Cohort-by-Year Effects

The estimates discussed above will combine the initial average 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 relatively productive (or unproductive) teachers that contribute more observations to the estimation by staying 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-specific indicators in (6.1) are replaced with cohort-by-year indicators:

$$\begin{aligned} A_{igst} =& \zeta_t + \lambda A_{igst-1} + X_{igst}\beta + Cohort \times Year_{igst}\gamma_1 + \gamma_2 \overline{A}_{-igst-1} + \\ & f(Exp_{igst}) + \gamma_3 CS_{igst} + \phi_g + c_i + \delta_s + e_{igst} \end{aligned} \tag{9.1}$$

Table 5 displays the estimates of equation (9.1). The baseline results from column (1) of Table 4 are also presented for reference. 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 are essentially unchanged from the previous estimates with post-CSR cohorts having average achievement 0.0033 (p-value=0.525) to 0.0277 (p-value=0.001) standard deviations below the pre-CSR cohorts.[20] The point estimates suggest the relative performance gap between pre-CSR and post-CSR cohorts drops to between 0.0078 (p-value=0.280) and 0.0192 (p-value=0.004) standard deviations in each cohort's second year. Importantly some, but not all, second year cohort effects are statistically different at conventional levels.[21]

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

school-grade, and grade-year effects that could be included in the model.

[20] First year cohort differences that are not statistically significant at the 5% level include 2001-2002 to 2003-2004 (p-value=0.310), 2002-2003 to 2003-2004 (p-value=0.525), and 2002-2003 to 2007-2008 (p-value=0.079)

[21] Second year cohort differences that are not statistically significant at the 5% level include 2001-2002 to 2003-2004 (p-value=0.057), 2001-2002 to 2004-2005 (p-value=0.280), 2001-2002 to 2005-2006 (p-value=0.063), 2001-2002 to 2006-2007 (p-value=0.196), and 2002-2003 to 2004-2005 (p-value=0.054)

Table 5: Pooled OLS Cohort and Cohort-by-Year Estimates

Specification	*Cohort*	*Cohort-by-Year*						
Equation	*(6.1)*	*(9.1)*						
Year		*2001-2002*	*2002-2003*	*2003-2004*	*2004-2005*	*2005-2006*	*2006-2007*	*2007-2008*
Entry Cohort								
2001-2002	-0.0035	-0.0411***	-0.0171***	0.00243	0.00208	0.00932	0.00929	0.0106*
	(0.00331)	(0.00512)	(0.00509)	(0.00617)	(0.00630)	(0.00564)	(0.00644)	(0.00581)
N	2824	2824	2028	1649	1547	1374	1258	1115
2002-2003	-0.0093***		-0.0453***	-0.0116***	-0.0120*	0.000914	0.0108	-0.00496
	(0.00247)		(0.00486)	(0.00403)	(0.00634)	(0.00636)	(0.00807)	(0.00553)
N	2856		2856	1990	1705	1534	1349	1225
2003-2004	-0.0162***			-0.0486***	-0.0308***	-0.00150	-0.00268	-0.00816
	(0.00465)			(0.00629)	(0.00557)	(0.00664)	(0.00980)	(0.00642)
N	3378			3378	2422	2076	1902	1706
2004-2005	-0.0221***				-0.0688***	-0.0249***	-0.00860	0.000120
	(0.00455)				(0.00726)	(0.00561)	(0.00583)	(0.00540)
N	4037				4037	2904	2457	2091
2005-2006	-0.0304***					-0.0636***	-0.0295***	-0.0198***
	(0.00237)					(0.00485)	(0.00381)	(0.00453)
N	4247					4247	2995	2489
2006-2007	-0.0319***						-0.0674***	-0.0261***
	(0.00450)						(0.00404)	(0.00563)
N	4492						4492	3080
2007-2008	-0.0265***							-0.0581***
	(0.00469)							(0.00521)
N	3390							3390
Observations	2,752,060				2,752,060			
R-squared	0.653				0.653			

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

Note: Models include teacher experience cubic, a class size proxy, student demographic variables, and school, grade and year dummies

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 here that these estimates come from a specification that includes a cubic term in teacher experience. This implies that much of this observed improvement for cohorts is being driven by compositional changes of the cohort, rather than human capital accumulation that is common to all cohorts.

These results suggest 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 (0.0688-0.0120=0.0568; 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.[22]

[22]The above estimates identify changes in mean cohort performance. Appendix C presents results for individual teacher value-added that provide a similar conclusion

10 Implications for Prior Quasi-experimental CSR Effect Estimates

The estimates of equations (6.1) and (9.1) can be thought of as identifying the state-wide 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 6 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 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 enrolment 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 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.[23]

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

[23]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.

Table 6: Estimates of New Cohort Effects by CSR Intensity

CSR Intensity	*None*	*Q1*	*Q2*	*Q3*	*Q4*
Entry Cohort			*District Enforcement*		
2001-2002	-0.0050	-0.0039	0.0004	-0.0165*	0.0053***
	(0.0076)	(0.0050)	(0.0091)	(0.0097)	(0.0012)
2002-2003	-0.0151***	0.0038	-0.0014	-0.0188***	-0.0197***
	(0.0055)	(0.0078)	(0.0053)	(0.0042)	(0.0029)
2003-2004	-0.0251***	-0.0199	-0.0164***	-0.0171*	-0.0044**
	(0.0062)	(0.0121)	(0.0045)	(0.0087)	(0.0019)
2004-2005	-0.0227***	-0.0292***	-0.0167***	-0.0375***	-0.0059***
	(0.0052)	(0.0061)	(0.0065)	(0.0102)	(0.0016)
2005-2006	-0.0320***	-0.0240***	-0.0338***	-0.0276***	-0.0336***
	(0.0050)	(0.0073)	(0.0067)	(0.0040)	(0.0028)
2006-2007	-0.0388***	-0.0176**	-0.0222***	-0.0668***	-0.0229***
	(0.0069)	(0.0082)	(0.0078)	(0.0078)	(0.0042)
2007-2008	-0.0357***	-0.0391***	-0.0251***	-0.0163	-0.0078***
	(0.0084)	(0.0060)	(0.0069)	(0.0129)	(0.0024)
Observations			2,752,060		
R-Squared			0.0653		
Entry Cohort			*School Enforcement*		
2001-2002	-0.0088*	-0.0117	-0.0159	0.0055	0.0488***
	(0.0045)	(0.0182)	(0.0101)	(0.0115)	(0.0056)
2002-2003	-0.0074**	-0.0197*	-0.0201*	-0.0073	-0.0126
	(0.0034)	(0.0113)	(0.0113)	(0.0147)	(0.0078)
2003-2004	-0.0226***	-0.0147	-0.0052	0.0042	0.0163*
	(0.0043)	(0.0104)	(0.0117)	(0.0162)	(0.0088)
2004-2005	-0.0225***	-0.0097	-0.0378*	-0.0178	-0.0206*
	(0.0045)	(0.0114)	(0.0223)	(0.0132)	(0.0118)
2005-2006	-0.0278***	-0.0326**	-0.0263**	-0.0219**	-0.0617***
	(0.0036)	(0.0135)	(0.0117)	(0.0085)	(0.0039)
2006-2007	-0.0306***	-0.0329**	-0.0504***	-0.0195*	-0.0376***
	(0.0051)	(0.0066)	(0.0113)	(0.0100)	(0.0080)
2007-2008	-0.0308***	-0.0314*	-0.0204	0.0040	-0.0160**
	(0.0049)	(0.0182)	(0.0164)	(0.0146)	(0.0077)
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

Table 7: 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	-0.0068***	0.0272***	0.0204***	-	-	-
	(0.0010)	(0.0030)	(0.0029)	-	-	-
2002-2003	-0.0074***	0.0270***	0.0195***	-0.0006	0.0002***	-0.0009**
	(0.0010)	(0.0030)	(0.0030)	(0.0004)	(0.0000)	(0.0004)
2003-2004	-0.0090***	0.0275***	0.0185***	-0.0022***	-0.0002***	-0.0019***
	(0.0015)	(0.0031)	(0.0029)	(0.0008)	(0.0001)	(0.0007)
2004-2005	-0.0115***	0.0271***	0.0156***	-0.0047***	-0.0001***	-0.0048***
	(0.0020)	(0.0030)	(0.0027)	(0.0012)	(0.0000)	(0.0012)
2005-2006	-0.0149***	0.0267***	0.0118***	-0.0081***	-0.0005***	-0.0086***
	(0.0017)	(0.0030)	(0.0028)	(0.0010)	(0.0001)	(0.0010)
2006-2007	-0.0182***	0.0261***	0.0079***	-0.0114***	-0.0011***	-0.0125***
	(0.0021)	(0.0029)	(0.0028)	(0.0013)	(0.0002)	(0.0013)
2007-2008	-0.0233***	0.0265***	0.0032	-0.0165***	-0.0007***	-0.0172***
	(0.0028)	(0.0030)	(0.0032)	(0.0020)	(0.0002)	(0.0019)

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

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. To assess the effect on average achievement of the change in average quality of new cohorts and the drop in average experience, the contribution of each of these components is calculated using the estimates of equation (6.1). The estimated contribution to average achievement in the state of the cohort composition and teacher experience are calculated in each year as $\overline{COHORT}_t\hat{\gamma}_1$ and $\overline{\hat{f}(EXP_t)} = \overline{EXP_t}\hat{\beta}_1 + \overline{EXP_t^2}\hat{\beta}_2 + \overline{EXP_t^3}\hat{\beta}_3$, respectively.

Both the total contribution and the separate contribution of each component are presented in Table 7, 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 0.0172 standard deviations. This difference is driven more by the relative performance of the cohorts than by the drop in teacher experience.[24]

To directly assess the role of these same changes in explaining the lack of estimated CSR achievement gains, estimates are used from a modified version of (6.1) in which a CSR treatment dummy is interacted with all included regressors. Table 8 displays the evolution of the total contribution (cohort composition plus experience) of teachers to average performance separately for schools considered "treated" and "untreated" based on the districts pre-CSR class-sizes. Table 8 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 previously used to estimate CSR policy effects. Specifically, prior CSR effect studies rely on treatment-control comparisons (Difference-in-difference (DinD), Comparative Interrupted Time Series,

[24]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 8: 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	0.0204***	0.0214***	-0.0010	-	-	-
	(0.0034)	(0.0057)	(0.0066)	-	-	-
2002-2003	0.0195***	0.0204***	-0.0009	-0.0008	-0.0010	0.0001
	(0.0033)	(0.0061)	(0.0069)	(0.0005)	(0.0008)	(0.0010)
2003-2004	0.0193***	0.0169**	0.0024	-0.0010	-0.0045***	0.0035**
	(0.0031)	(0.0065)	(0.0072)	(0.0009)	(0.0010)	(0.0013)
2004-2005	0.0161***	0.0150**	0.0011	-0.0043***	-0.0064***	-0.0021
	(0.0031)	(0.0060)	(0.0067)	(0.0015)	(0.0017)	(0.0023)
2005-2006	0.0133***	0.0082	0.0051	-0.0071***	-0.0132***	0.0061***
	(0.0033)	(0.0058)	(0.0066)	(0.0012)	(0.0011)	(0.0016)
2006-2007	0.0097***	0.0035	0.0062	-0.0107***	-0.0179***	0.0072***
	(0.0030)	(0.0064)	(0.0071)	(0.0015)	(0.0014)	(0.0021)
2007-2008	0.0084***	0.0018	0.0066	-0.0120***	-0.0196***	0.0076***
	(0.0031)	(0.0076)	(0.0082)	(0.0017)	(0.0022)	(0.0028)

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

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 teachers. Both treated and untreated schools experience a drop in the teachers' contribution to average achievement. Interestingly, the CSR schools saw a slightly smaller drop, 0.0076 test score standard deviations smaller by 2007-2008, than those schools for which CSR was not binding at introduction. This estimate is of the opposite sign needed to explain the finding of no achievement gain from CSR. Clearly, the change in average achievement attributable to the make-up of the teaching stock falls well short of explaining the lack of achievement gains.

11 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 noticeable achievement gains from CSR in State X. Despite large increases in the number of teachers, the evidence suggests that newly-hired teachers account for only slight decreases in achievement during the implementation of CSR. The overall drop in achievement from the 2001-2002 to the 2007-2008 school year attributable to changes in the average quality, experience, and cohort composition of fourth through sixth grade teachers is estimated to be only 0.0172 test score standard deviations. Furthermore, the results suggest that this decrease in quality was experienced by both treated and untreated schools alike. These treatment spillovers imply that the disappointing CSR effects found in quasi-experimental research cannot be explained by differential changes in

new teacher quality.

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. An important next step is to consider how these changes in demand may have led to a movement of teachers across schools. Understanding the mechanisms at play will help to determine whether popular CSR policies can be designed to promote achievement gains.

More generally, the results of this paper suggest that while large short-run increases in teacher demand may lead to modest declines in the value-added of newly hired teachers, these declines may not substantially affect long-run achievement. This conclusion should be interpreted with caution, as our findings reflect the experience of a single state based on teachers in grades four through six. In other states or grades, the quality of incoming teachers may fall more dramatically in response to the introduction of CSR policies.

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.

Dieterle, S. (2012). Class-size Reduction Policies and the Composition of the Teacher Workforce. Unpublished draft.

Dieterle, S., Guarino, C., Reckase, M., & Wooldridge, J. (2012). How do Principals Assign Students to Teachers? Finding Evidence in Administrative Data and the Implications for Value-added. IZA Discussion Paper 7112.

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

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. (2011). Evaluating Value-added Methods for Estimating Teacher Effects. Working paper.

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.

Reardon, S. & Galindo, C. (2009). The Hispanic-White Achievement Gap in Math and Reading in the Elementary Grades. *American Educational Research Journal, 46(3),* 853-891.

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	*(6.1)*	*(9.1)*
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

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 2011). 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:

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

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 pri-

marily 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),[25] equation (B.3), referred to as the lag score equation from here on, could be reasonably estimated by OLS.[26] 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 estimators. 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).[27] 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 (Dieterle et al. 2012). 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.[28] Practically speaking, the strict exogeneity assumption

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

[26]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.

[27]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.

[28]Note that the strict exogeneity assumption is what precludes the use of fixed effects on the lag score

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 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.[29] 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 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 by replacing the cohort indicators in (6.1) with indicators for each teacher.[30] Teachers are given a percentile rank based on their estimated value-added relative to all the teachers in the sample. Figure 3 displays histograms of the distribution of teacher percentile ranks for each entry cohort. The solid line on each graph

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*.

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

[30]Due to computational constraints, this estimation is done separately by district.

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). Prior to CSR, the percentile rank distribution of the entry cohorts is roughly uniform. Over the implementation of CSR, starting with the 2003-2004 entry cohort, there is a noticeable increase in the probability a given teacher will be ranked below the twentieth percentile.

It is important to note that the value-added estimates for later cohorts will tend to be noisier. However, if differences in the percentile rank distributions across cohorts were simply an artifact of increased noise, more outliers would be expected at both ends of the distribution resulting in a U-shaped distribution. That we only see more teachers at the low end of the percentile rank distribution for the later cohorts suggest that it is not due purely to noise. Regardless, Figure 3 provides additional suggestive evidence that teachers hired post-CSR were more likely to be low value-added teachers.

Figure 3: Percentile Rank Distributions