Class-size Reduction Policies and the Quality of Entering Teachers

Steven Dieterle
Michigan State University
October 31, 2011

Abstract:

Previous research has established the potential for achievement gains from attending smaller classes. However, large state-wide class-size reduction (CSR) policies have not been found to consistently realize such gains. A leading explanation for the disappointing performance of CSR policies is that schools are forced to hire additional teachers of lower quality to meet the new class-size requirements. This paper uses administrative data from an anonymous, diverse state to explore whether there were noticeable changes in the quality, measured by value-added to mathematics achievement, of newly hired teachers around the introduction of CSR. If so, were these changes large enough to account for the disappointing performance of the policy? The results suggest that while there was a modest fall in the relative average quality of newly hired teachers and those retained beyond their first year, this drop is not nearly large enough to explain the failure of CSR to produce sizeable achievement gains. Furthermore, schools facing CSR pressure saw similar falls in quality as those that did not, likely due to labor market competition forcing all schools along the effective teacher supply curve. Therefore, between-school differences in the quality of incoming teachers cannot explain the failure of previous quasi-experimental treatment-control comparisons to find achievement effects from statewide CSR. In addition to providing insight into CSR, the results are informative for assessing any potential intervention that may drastically increase the short-run demand for teachers.

The author would like to thank Gary Solon, Todd Elder, Jeff Wooldridge, Cassie Guarino, Steven Haider, Otavio Bartalotti, and Quentin Brummet, as well as seminar participants at Michigan State University for helpful comments. All errors are the responsibility of the author.

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 large scale 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 estimated effects consistently falling short of what might be expected from the experimental research. In the state studied here, a previous paper concluded there was no CSR effect. Due to the high costs of implementation, on the order of $21 billion over nine years in Florida (Florida Department of Education (FDOE)) 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 underperformance 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.

Using administrative data covering grades four through six from an anonymous, diverse state (subsequently referred to as State X) around the implementation of a state-wide CSR program, this paper addresses two separate, but related, questions. Did the CSR-induced demand increase lead to schools hiring and retaining lower quality teachers, here measured by value-added to student mathematics achievement?[1] If so, what effect did this fall in quality have on average achievement and can it explain a large portion of the unrealized CSR gains?

[1] Similar results obtained using reading test scores instead of mathematics 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

The value-added estimates of cohort performance found here indicate a small 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 0.0182 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.

Even if the average quality of cohorts had not changed, there may have been a 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, and was generally experienced 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 result implies that a differential change in the quality of newly hired teachers is not the mechanism preventing the expected achievement gains from CSR. Further, it suggests a role for competition for teacher candidates pushing all schools along the effective teacher supply curve in connected labor markets.

---

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.

The results are informative beyond providing a better understanding of CSR programs and how they relate to the pool of employed teachers. 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). 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 changes in teacher characteristics that accompanied CSR, section 6 discusses the empirical strategy used, section 7 gives and discusses the main results, section 8 tests the sensitivity of the results to the use of alternative value-added measures, and section 9 concludes.

2. Literature Review/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 the STAR data 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.

The finding of statistically and practically significant class-size effects from the Tennessee STAR experiment 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 CSR 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 (2010) found null effects for fourth through eighth grade of CSR in Florida. The effectiveness of CSR in State X will be explored in more detail in section 7.

In light of the experimental results discussed above, the fact that gains are not consistently realized with large-scale CSR programs is puzzling. One possibility is that the experimental Tennessee STAR results could be an anomaly. Indeed, not all experimental and quasi-experimental studies find significant class-size effects (Hoxby 2000). The size and scope of STAR limit our ability to assess whether the results would hold up under repeated experiments. Given this limitation, a recent paper by Rockoff (2009) that discusses the results of several class-size experiments from the beginning of the twentieth century provides some additional context for the STAR results. Rockoff 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, including the small scale of these early studies and some experimental design issues. 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. Given these concerns, these experiments serve, at most, as suggestive evidence that the results of Tennessee STAR may be an anomaly.

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 2010). 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, it is common to look at changes in observable teacher characteristics associated with the implementation of CSR. 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 student performance on exams 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.

Given the evidence from the education production function literature, 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. Only a few papers explore the teacher-quality explanation using student test scores.

Jepsen & Rivkin (2009) analyze California's 1996 CSR program and back out an estimate of the relationship between teacher cohort size and quality. Within the framework of their fixed effects CSR analysis, the authors examine whether the estimated effects of teacher experience and certification differed across years. Intuitively, this identifies the quality of new cohorts of teachers since those teachers identified as inexperienced or uncertified in a given year are more likely to be new. They find no statistically or practically significant differences in the estimated experience or certification effects across years. With cohort sizes ranging from under 4,000 to over 5,000 teachers during this period, the authors conclude that there is no evidence that cohort size is related to teacher quality. Jepsen and Rivkin's approach is limited, however, by the available data. Specifically, without access to individual-level data on students and teachers, the authors cannot identify which teachers make up a hiring cohort and cannot look specifically at the performance of students in those classrooms.

Kane & Staiger (2005) use individual-level data from Los Angeles to analyze California's CSR program. They calculate value-added for teachers hired just before and immediately after CSR and find no differences among the two cohorts of teachers.[2] They also find no evidence of differential attrition among the two cohorts. Kane & Staiger benefit from having individual-level data. However, they are limited by having data on a single, albeit very large, school district. Given the potential differences between Los Angeles and other districts in the state, it is difficult to conclude how general these results are. With data on individual students and teachers for an

[2] Unfortunately the Kane & Staiger (2005) paper was unpublished and is unavailable via the internet. This information comes from a related paper by Staiger & Rockoff (2010).

entire state, it is possible to assess the change in teacher quality associated with CSR more directly and to look across heterogeneous districts and schools.

3. Institutional Details: CSR in State X

In November of 2002, State X voters approved a constitutional amendment which created a new state wide CSR program. The program was set to begin the following school year, 2003-2004. Separate class-size maximums were set for different grade levels, as shown in Table 1. The law also established per-pupil allocations from the state government for each year a district or school was found to be in compliance with the law. There is anecdotal evidence 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.

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. Table 1 also shows the change

in the average class size in the state by grade grouping over eight years. With average class size dropping from 23 to 16 for the earliest grades and 24 to 19 in the middle grades, it is clear that the program did achieve the stated goal of reducing class size.

Table 1: New Class-size Maximums: State X

		District Level		
Grades	*Maximum*	*Percent Below Max Yr 1*	*Average CS Yr 1*	*Average CS Yr 8*
KG-G3	18	11.94%	23.07	16.39
G4-G8	22	41.79%	24.16	18.91
G9-G12	25	91.04%	24.10	21.94

Source: State X Department of Education

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 administrative data link students in grades one through six to teachers and schools from the 2000-2001 to the 2007-2008 school year. In addition to basic student demographics, the data include test scores 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.

Importantly, 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, certification, and degree level.

Finally, State X has made each district/school's average class size publically 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. Generally in this paper, schools are divided into those with district (school) level average class size below the maximums for grades four through eight in the year prior to district (school) level CSR

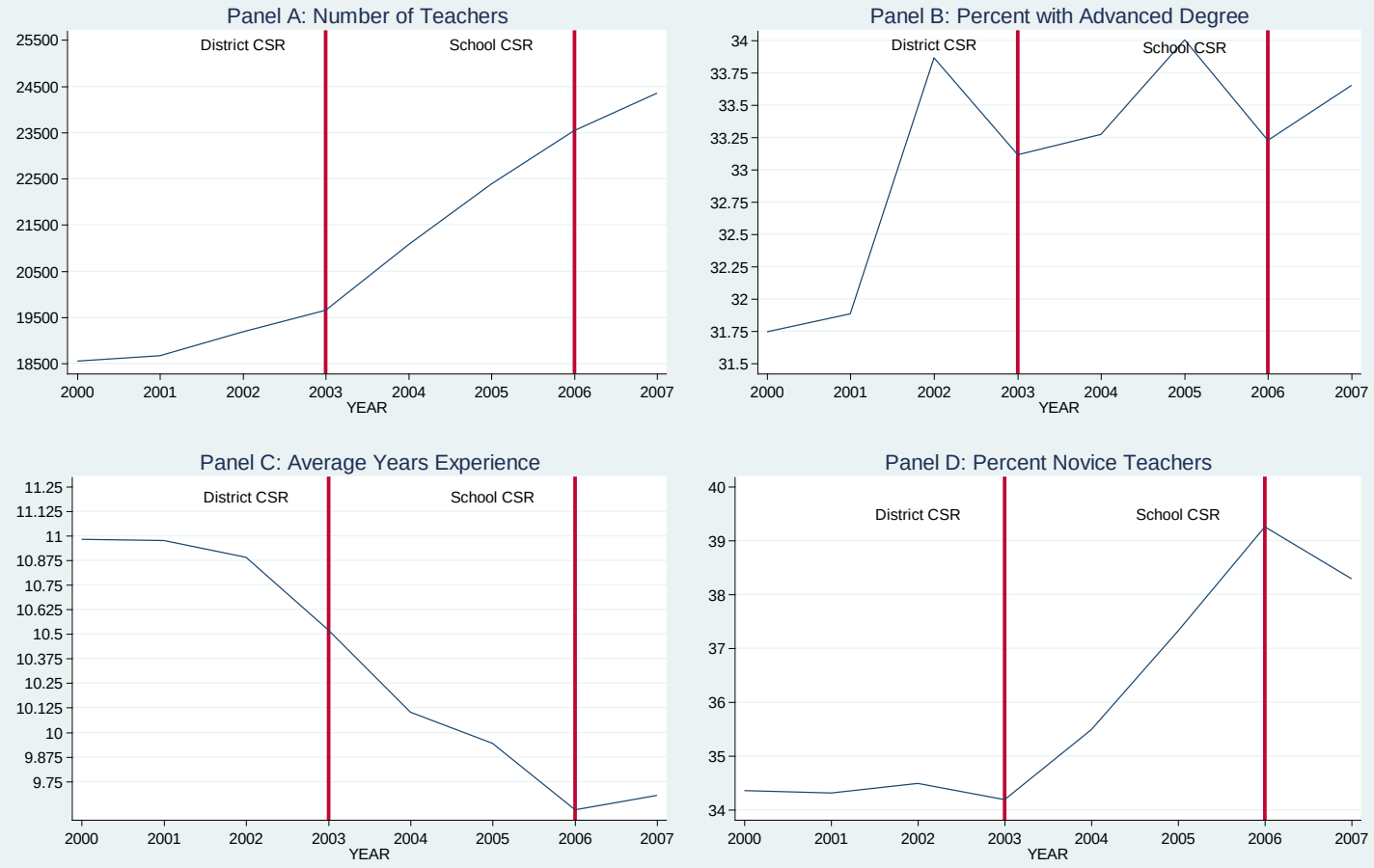
enforcement and into quartiles of average class size for those above the maximums. Descriptive statistics for the key variables used in this study are presented in Appendix Table 1. Notably, nearly 70% of the student-year observations in the data are linked to a teacher observed entering at some point in the sample period.

5. CSR and Teacher Characteristics in State X

To provide background for the subsequent analysis and to tie the current work to the previous CSR literature on changes in the teacher workforce, it is helpful to consider how teacher characteristics changed with the introduction of CSR in State X. Bohrnstedt & Stecher (1999) found fairly large swings in teacher characteristics around California's introduction of CSR with the percent with less than three years experience rising from 17% to 28%, the percent without an advanced degree rising from 17% to 23%, and the percent not fully certified going from 1% to 12% in just three years. On the other hand, Dieterle (2011) found much smaller changes in State X with only average teacher experience found to consistently drop with the introduction of CSR. Dieterle uses publicly available data aggregated at the school level for his analysis. The administrative data used in this study allow for a closer look at the characteristics of both the stock of teachers and the flow of new teachers into the system. Here, the focus is on teachers teaching a core course (those that fall under CSR requirements) in grades four through six (those VAM estimation is possible for). Figure 1 displays the trends over time in the number of teachers, percent with an advanced degree, average experience, and percent with three or fewer years of experience for all fourth through sixth grade core course teachers.

Figure 1: Teacher Level Trends

All Teachers Grades 4-6 in Core Courses



In panel A, we see the number of teachers in this group rising steadily over the introduction of CSR from under 19,500 before CSR to nearly 24,500 after five years. This rise in the stock of teachers was associated with a less clear pattern in the percentage of teachers with an advanced degree, shown in panel B. While the percentage with an advanced degree drops by three-quarters of a percentage point with the introduction of CSR and the change to school-level enforcement, the percentage increases in the other years. Finally, panels C and D show trends in teacher experience, the one characteristic previously found to be related to CSR implementation

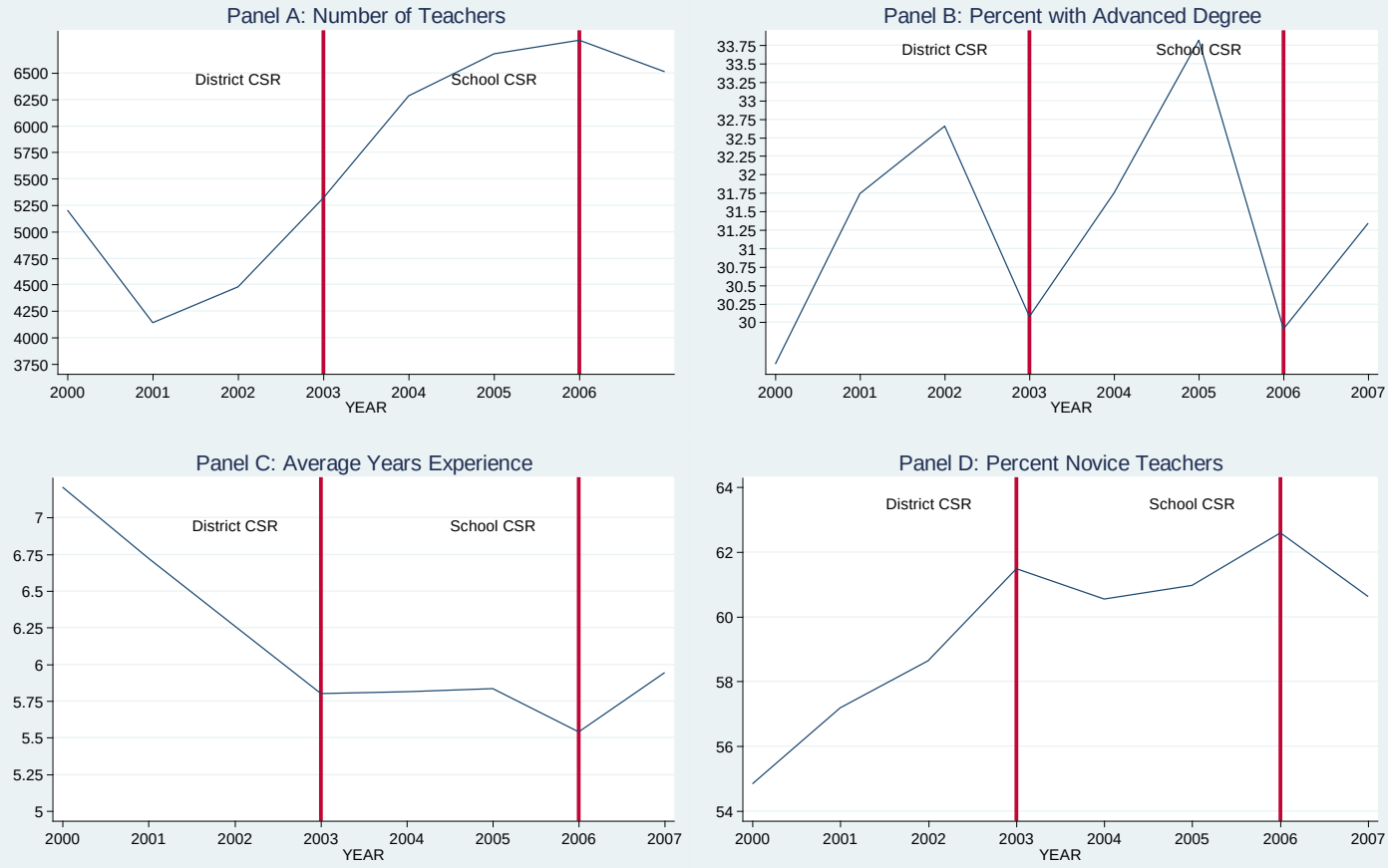
in State X (Dieterle 2011). Average experience drops from a pre-CSR level of roughly eleven years to nearly 9.5 years by the introduction of school-level enforcement four years later. While this drop in the average experience of the stock of teachers is in no way trivial, given prior research that finds that early experience matters more for student achievement, it is unclear what impact this may have had on CSR performance. Panel D addresses this issue of early experience by showing the trend in the percentage of teachers considered novices, those with three or fewer years of experience. The percentage novice increases by five percentage points, from 34% to 39%, over the implementation of CSR. Such a large increase in novice teachers may well have contributed to the perceived performance of CSR in State X. However, to the extent these new teachers stayed in teaching, the drop in achievement associated with having more novice teachers may only be temporary as the teachers gain experience.

Figure 2 explores trends in the same characteristics as in Figure 1, but for the flow, rather than the stock, of teachers into the state public elementary school system. Recall that the data follow all first through sixth grade teachers in public schools in State X. Therefore, a teacher entering the data may be 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. Panel A shows that the number of fourth through sixth grade core course teachers entering the data each year increases by roughly 2,000 by the fourth year of CSR. The percent with an advanced degree, in Panel B, shows a similar pattern to the stock of teachers, dropping only in years of a change in CSR policy. The drop is more pronounced, with the percentage of entering teachers with advanced degrees falling by 2.5 and 3.75 percentage points with the introduction of district- and school- level CSR enforcement, respectively. The

average experience of entering cohorts, shown in Panel C, actually falls more before CSR than after. Similarly, the percent novice increases more before CSR.

Figure 2: Teacher Level Trends

Entry Cohorts in Grades 4-6 in Core Courses



6. Empirical Methodology

The empirical methodology used here follows from the standard value-added approach to education production function estimation and consists of two broad steps. The first set of estimates uses several variants of the value-added framework to answer the general question of whether schools hire teachers of lower quality with the CSR induced demand increase. Selected results from the first step are then used to create estimates of the change in overall average

achievement that can be attributed to the change in the stock of teachers. Additionally, I explore the extent to which these changes can explain the disappointing CSR achievement effects from estimators that compare treated schools (those for which the new CSR maximums were binding at introduction) to untreated schools.

The main estimation 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:[3]

$$A_{ist} = \zeta_t + \lambda A_{ist-1} + X_{ist}\beta + Cohort_{ist}\gamma_1 + \gamma_2 \bar{A}_{-ist-1} + f(Exp_{ist}) + \gamma_3 CS_{ist} + g_{ist} + c_i + \delta_s + e_{it} \quad (6.1)$$

where

A_{ist} is student i's test score in school s in year t

ζ_t are year fixed effects

A_{ist-1} is student i's prior test score

X_{ist} are student demographics

$Cohort_{ist}$ are teacher cohort indicators

$\bar{A}_{-ist-1}$ is the average prior test score of student i's classmates

$f(Exp_{ist})$ is a cubic in teacher experience

CS_{ist} is a proxy measure of class size

g_{ist} are grade fixed effects

c_i is an unobserved student heterogeneity term

δ_s are school fixed effects

Note that the preferred estimation strategy effectively ignores the presence of the unobserved student heterogeneity. While the assumptions underlying the preferred estimation are unlikely to hold, there is evidence that OLS estimation of the lag score specification may perform well in practice. 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. Using simulated data, Guarino et al. (2011)

[3] See Appendix B: Measuring Teacher Quality, for a detailed discussion of value-added models and estimation.

find that the lag score specification estimated by OLS is fairly robust, compared to other common VAM estimators, to different teacher and student sorting mechanisms. 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 may be introducing additional assumptions and greatly reducing the identifying variation, while not capturing much more of the assignment mechanism which threatens the validity of VAM estimates. The sensitivity of the results to the choice of value-added approach is explored in section 8.

In practice, the class size is measured by the number of students linked to a teacher in a given year in the test score 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) I 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.

The main coefficients of interest are the estimates of γ_1, the average quality of entry cohorts of teachers. Specifically, interest lies in comparing the average quality of entry cohorts 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 one of the biggest threats to the validity of estimates in the value-added framework,

the fact that certain schools may tend to service higher- or lower- ability students on average. In effect, the school fixed effects control for sorting into schools, and the lagged achievement and average classmates' lagged achievement control for within-school sorting and matching of teachers to students.

Secondly, the school fixed effects help in identifying whether schools were hiring teachers of lower quality in CSR years. Consider a case in which schools face teacher supply curves consisting of candidates of homogeneous quality. However, given evidence that there is substantial sorting of teachers into geographically small markets (Boyd et al. 2005; Lankford et al. 2002), allow each school to face a different level of teacher quality. If CSR disproportionately induced hiring in schools that faced supplies of lower quality teachers, without controlling for these school level differences in supply, we would see an overall negative relationship between CSR years and the average quality of new entrants. However, in this scenario there is no actual relationship between CSR and the quality of teachers hired as schools were able to hire additional teachers of the same quality. The inclusion of school fixed effects controls for the time-invariant quality level of teacher supply that different schools face.

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 not to interpret the estimates of (6.1) as necessarily a pure CSR hiring effect. In particular, the effects of any changes that may be related to the quality of teachers hired in a given year are included in the estimates of γ_1. State X enacted many policies over the same time period, including 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 Educator 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 the estimated cohort quality may be capturing these changes.

While issues such as this threaten the causal validity, the estimates still provide an informative look at how teacher quality may have changed with CSR. In fact, this simple approach captures potential CSR effects that would be difficult to parameterize given the available data. For example, the school-level class-size averages are only available starting with the year directly before school-level enforcement. 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.

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 are replaced with cohort-by-year indicators:

$$A_{ist} = \zeta_t + \lambda A_{ist-1} + X_{ist}\beta + Cohort \times Year_{ist}\gamma_1 + \gamma_2 \bar{A}_{-ist-1} + f(Exp_{ist}) + \gamma_3 CS_{ist} + g_{ist} + c_i + \delta_s + e_{it} \quad (6.2)$$

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. In this way, the cohort-by-year estimates are likely preferred to those discussed above; however the previous estimators without cohort-by-year effects are maintained as they provide a much more tractable comparison among the many other modeling and estimation choices.

The estimates of equations (6.1) and (6.2) can be thought of as identifying the statewide general equilibrium relationship between cohorts and teacher quality. 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. For the next set of estimates, the entry cohorts are further divided into teachers entering schools based on their pre-district and pre-school level enforcement class-size averages. For the pre-district level enforcement, schools are grouped based on the district-level averages in the year before CSR, with those districts already under the new maximums for fourth through eighth grade in one group and the remaining districts divided into quartiles by average class size. The schools are grouped similarly using the pre-school level enforcement class-size averages.

The previous estimates all identify changes in average performance of cohorts. To assess the change in the distribution of teacher quality among hiring cohorts, individual teacher value-added estimates are obtained by replacing the various cohort dummies with individual teacher indicators Due to computational constraints, this estimation is done separately by district.

All of the above estimates are aimed at uncovering the change in teacher quality associated with increased CSR hiring. By using a select set of these estimates it is possible to more directly test the impact that the change in the stock of teachers had on average achievement in State X and whether this change can explain the fact that quasi-experimental estimates show disappointing CSR achievement effects. As shown in section 5, the number of new teachers increased while the average experience of teachers in State X went down over the implementation of CSR. 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 will be presented along with the change since 2001. A second set of estimates, based on interacting a CSR district treatment dummy with all included covariates in (6.1),[4] is used to show a similar breakdown of the total contribution of the stock of teachers to achievement differences among these schools. This directly addresses the question of how much of the lack of achievement gains found in quasi-experimental estimates of the CSR policy in State X can be explained by differential changes in teacher quality.

7. Results

Before presenting the main results, I estimate the CSR policy effect within the value-added framework discussed in section 6. These results will help to establish the extent to which CSR in State X 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:

$$A_{ist} = \zeta_t + \lambda A_{ist-1} + X_{ist}\beta + (T \times YEAR_{st})\gamma_1 + \gamma_2 \bar{A}_{-ist-1} + g_{ist} + c_i + \delta_s + e_{it} \quad (7.1)$$

Two separate regressions are estimated based on school- or district- level CSR enforcement. For the district-level enforcement, treatment is defined by being in a district that was above the new class-size maximum in the year before CSR. The school-level treatment status is similarly determined by the school average class size the year prior to school-level enforcement. It is

[4] This is effectively the same as running separate regressions using the treated and untreated samples.

important to note that the regressions include year and school dummy variables and the omitted treatment category is for the 2001-2002 cohort.

Table 2 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. A previous paper by another author found small and statistically insignificant CSR achievement effects in State X using similar administrative data. Beginning with the district-CSR results, we find that 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 20004-2005 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 estimates of Krueger (1999) would be on the order of one-eighth of a standard deviation.[5] Even the ninety-five percent confidence intervals for these estimates fall short of half of the rough Tennessee STAR benchmark.

Moving on to the school-level enforcement results in the last column of Table 2, we see negative estimates of the treatment-by-year effects after the switch to school level enforcement during the 2006-2007 school year. 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

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

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.

Table 2: Estimated CSR Mathematics Achievement Effects for State X

Sample	*G4-G6*	*G4-G6*
CSR Level	*District*	*School*
T x 2002-2003	-0.0170	-0.0323
	(0.0180)	(0.0244)
T x 2003-2004	0.0163	-0.0284*
	(0.0152)	(0.0143)
T x 2004-2005	0.0264**	-0.00604
	(0.0125)	(0.0102)
T x 2005-2006	0.00902	-0.0459***
	(0.0183)	(0.0164)
T x 2006-2007	-0.00522	-0.0410*
	(0.0186)	(0.0231)
T x 2007-2008	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; clustered at district (school) level for district (school) CSR

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

The results found in Table 2 generally concur with those found by a previous study. 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.[6] Importantly the difference in the estimated achievement effects and the Tennessee STAR benchmark allows for the possibility that the average quality of the newly hired teachers did change with CSR and that this change may have affected the performance of the policy.

[6] Importantly, direct replication and extension of the previous paper's approach finds that the estimated CSR effects can be sensitive to the included covariates and the particular model used. In some cases, point estimates are found on the order of 0.06 standard deviations, and in many cases the standard errors are too large to rule out effects close to the one-eighth of a standard deviation benchmark from Tennessee STAR. The value-added modeling approach used here is much less sensitive to the choice of included covariates, but does limit the data used in estimation by requiring a lagged test score for each student.

Tables 3 and 4 display the estimates based on the general specifications found in equations (6.1) and (6.2). 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.

Table 3 presents the baseline estimates of the cohort effects in the first column.[7] Again I use the convention of shading district CSR enforcement years in light gray and school CSR enforcement years in dark gray. For reference, the initial cohort size is also presented. All specifications are estimated using developmental scale test scores that have been standardized within grade and year.[8] We see 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) worse than students with a 2002-2003 cohort teacher. Importantly, the estimated cohort effects for these two cohorts are strongly statistically different with a p-value of 0.0000.

Overall, the estimated post-CSR cohort effects range from 0.0069 to 0.0285 standard deviations lower than the two pre-CSR cohorts. The magnitude of the differences seen in Table 3 fall short of what would be needed to explain why CSR policies do not produce the gains expected based on experimental results. Recall that a simple extrapolation of the STAR results would place the expected achievement gain at roughly one-eighth of a standard deviation.

[7] See Appendix Table 2 for other estimates from these regressions.
[8] 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.

Differences in cohort performance of, at most, one-thirty-fifth of a standard deviation do not compare to the difference in estimated CSR effects and the Tennessee STAR benchmark.

Table 3: Pooled OLS Cohort Estimates with Cohort Indicators or Cohort-by-Year Indicators with Cohort Size

Specification	*Cohort*	*Cohort-by-Year*						
Equation	*(6.1)*	*(6.2)*						
Year		*2001-2002*	*2002-2003*	*2003-2004*	*2004-2005*	*2005-2006*	*2006-2007*	*2007-2008*
Entry Cohort								
2001-2002	-0.00345	-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	-0.00929***		-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***	-0.0210**		-0.0486***	-0.0308***	-0.00150	-0.00268	-0.00816
	(0.00465)	(0.00920)		(0.00629)	(0.00557)	(0.00664)	(0.00980)	(0.00642)
N	3378			3378	2422	2076	1902	1706
2004-2005	-0.0221***	-0.00368	-0.00705		-0.0688***	-0.0249***	-0.00860	0.000120
	(0.00455)	(0.0128)	(0.00750)		(0.00726)	(0.00561)	(0.00583)	(0.00540)
N	4037				4037	2904	2457	2091
2005-2006	-0.0304***	-0.00655	-0.0113	-0.0100		-0.0636***	-0.0295***	-0.0198***
	(0.00237)	(0.00822)	(0.00711)	(0.0129)		(0.00485)	(0.00381)	(0.00453)
N	4247					4247	2995	2489
2006-2007	-0.0319***	-0.00422	-0.00696	-0.00655	0.0177*		-0.0674***	-0.0261***
	(0.00450)	(0.0121)	(0.00718)	(0.00894)	(0.00954)		(0.00404)	(0.00563)
N	4492						4492	3080
2007-2008	-0.0265***	0.00170	-0.00551	0.00498	-0.00110	0.00283		-0.0581***
	(0.00469)	(0.0110)	(0.00872)	(0.00569)	(0.00803)	(0.0118)		(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

Table 3 also displays the results from equation (6.2) that allows for separate cohort-by-year effects. Recall from the previous section that the motivation for these estimates was to explicitly allow the average performance of a cohort to change as its composition changes. Note that the

definition of entering teacher used allows for reentering teachers to be part of a new cohort. For example, if a teacher leaves the sample after the 2002-2003 school but reenters in 2005-2006, they will be considered part of the 2005-2006 cohort. This definition implies that while a given cohort has no observations in the year directly before entry, there will be some observations for that cohort in earlier years.

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 relatively unchanged from the previous estimates with post-CSR cohorts having average achievement 0.0033 to 0.0277 standard deviations below the pre-CSR cohorts. The point estimates suggest the relative performance gap between pre-CSR and post-CSR cohorts drops to between 0.0078 and 0.0182 standard deviations in each cohort's second year. Not all second year post-CSR cohort effects are statistically different from the pre-CSR estimates at traditional significance levels. For instance, the p-value is 0.3030 for the t-test of the null that the second-year effects for the 2001-2002 and 2004-2005 cohorts are the same. On the other hand we do reject the null that the 2002-2003 cohort and the 2003-2004 cohorts were equally effective in their second years (p-value=0.0039).

Also note that pre-CSR cohorts become comparable to the baseline teachers after three or four years with year-specific cohort effects statistically indistinguishable from zero. The two post-CSR cohorts observed for at least four years, 2003-2004 and 2004-2005, also appear to level off to be roughly comparable to the baseline after four years. This result suggests that the potential long-run CSR hiring effects may be even smaller than those initially observed. However, we do not observe the largest post-CSR hiring cohorts 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 and Staiger 2010). However, it appears that the long-run achievement effect of these changes may be relatively small.

Finally, 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 earlier cohorts likely face similar resources within schools as later cohorts

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) 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 to differences in other classroom level inputs.

Table 4 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. Looking first at the estimates based on the district groupings, we see 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 enrollment growth were hired by nearby schools facing CSR pressure.

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

level enforcement in the following year, had a hiring cohort estimated to be 0.0617 test score standard deviations worse than the baseline teachers, while the other schools saw cohorts between 0.0219 and 0.0326 standard deviations worse.

Table 4: Estimates of New Cohort Effects by CSR Intensity

	None	Q1	Q2	Q3	Q4
Entry Cohort			*District Enforcement*		
2001-2002	-0.00500	-0.00394	0.000417	-0.0165*	0.00531***
	(0.00760)	(0.00503)	(0.00912)	(0.00969)	(0.00115)
2002-2003	-0.0151***	0.00384	-0.00144	-0.0188***	-0.0197***
	(0.00545)	(0.00784)	(0.00529)	(0.00417)	(0.00286)
2003-2004	-0.0251***	-0.0199	-0.0164***	-0.0171*	0.00444**
	(0.00622)	(0.0121)	(0.00445)	(0.00869)	(0.00194)
2004-2005	-0.0227***	-0.0292***	-0.0167**	-0.0375***	-0.00591***
	(0.00519)	(0.00610)	(0.00648)	(0.0102)	(0.00156)
2005-2006	-0.0320***	-0.0240***	-0.0338***	-0.0276***	-0.0336***
	(0.00501)	(0.00726)	(0.00673)	(0.00404)	(0.00281)
2006-2007	-0.0388***	-0.0176**	-0.0222***	-0.0668***	-0.0229***
	(0.00689)	(0.00816)	(0.00778)	(0.00781)	(0.00416)
2007-2008	-0.0357***	-0.0391***	-0.0251***	-0.0163	-0.00780***
	(0.00834)	(0.00598)	(0.00690)	(0.0129)	(0.00244)
Observations			2,754,022		
R-squared			0.653		
			School Enforcement		
2001-2002	-0.00879*	-0.0117	-0.0159	0.00550	0.0488***
	(0.00447)	(0.0182)	(0.0101)	(0.0115)	(0.00555)
2002-2003	-0.00744**	-0.0197*	-0.0201*	-0.00728	-0.0126
	(0.00342)	(0.0113)	(0.0113)	(0.0147)	(0.00778)
2003-2004	-0.0226***	-0.0147	-0.00517	0.00423	0.0163*
	(0.00430)	(0.0104)	(0.0117)	(0.0162)	(0.00876)
2004-2005	-0.0225***	-0.00972	-0.0378*	-0.0178	-0.0206*
	(0.00449)	(0.0114)	(0.0223)	(0.0132)	(0.0118)
2005-2006	-0.0278***	-0.0326**	-0.0263**	-0.0219**	-0.0617***
	(0.00355)	(0.0135)	(0.0117)	(0.00849)	(0.00387)
2006-2007	-0.0306***	-0.0329***	-0.0504***	-0.0195*	-0.0376***
	(0.00506)	(0.00659)	(0.0113)	(0.0100)	(0.00803)
2007-2008	-0.0308***	-0.0314*	-0.0204	0.00395	-0.0160**
	(0.00487)	(0.0182)	(0.0164)	(0.0146)	(0.00773)
Observations			2,752,060		
R-squared			0.653		

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

The above estimates 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. To explore the relative performance of CSR cohorts, 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 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. To the extent that the included experience profile does a poor job of capturing the human capital accumulation for these later cohorts, the perceived pattern may better represent a short-run effect.

It is important to note that the value-added estimates for later cohorts will tend to be noisier as well. However, if differences in the percentile rank distributions across cohorts were simply an artifact of increased noise, we would expect more outliers 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. Ultimately, this analysis would be aided by having access to additional years of data to better address the issue of differences in the precision of the estimates for different cohorts. 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

Entry Cohorts

The comparison among the estimated entry cohort effects does not fully capture the contribution of these teachers to average statewide achievement. In particular, this comparison misses the fact that not all students in CSR years are taught by teachers hired in post-CSR cohorts and that the average experience in the state dropped in post-CSR years. Table 5 shows the estimated contribution to average achievement in the state of the cohort composition ($\overline{COHORT_t}\hat{\gamma}_1$) and teacher experience ($\overline{\hat{f}(EXP_t)} = \overline{EXP_t}\hat{\beta}_1 + \overline{EXP_t^2}\hat{\beta}_2 + \overline{EXP_t^3}\hat{\beta}_3$) using the estimates from column (1) of Table 3. Both the total contribution, and the separate contribution of each component are presented. Finally, the change in the contribution to statewide average achievement since 2001 is also shown. While we see the contribution attributable to these teacher characteristics fall 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.

Table 5: Estimated Contribution of Teacher Cohort Composition and Experience to Average Achievement

	Achievement Contribution			Change from 2001		
Year	$\overline{COHORT_t}\hat{\gamma}_1$	$\overline{\hat{f}(EXP_t)}$	*Total*	$\overline{COHORT_t}\hat{\gamma}_1$	$\overline{\hat{f}(EXP_t)}$	*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

To directly assess the role of these same changes in the estimated CSR policy effects, estimates are used from a modified version of (6.1) in which a CSR treatment dummy is interacted with all included regressors. Table 6 displays the evolution of the total contribution, cohort composition plus experience, of teachers to average performance separately for CSR and non-CSR schools based on pre-district CSR class-sizes. Table 6 also shows the difference in these changes between CSR and non-CSR schools. Column six is of particular interest as it relates to the type of comparison previously used to estimate CSR policy effects. We see that both CSR and non-CSR 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. Recall that a simple prediction from Krueger's analysis of STAR would suggest a CSR effect of roughly one-eighth of a standard deviation. Clearly, the change in average achievement attributable to the makeup of the teaching stock falls well short of explaining the lack of achievement gains.

Table 6: Estimated Contribution of Teacher Composition to Average Achievement: CSR vs. Non-CSR Schools

	Total Achievement Contribution			Change from 2001		
Year	*CSR*	*No CSR*	*Difference*	*CSR*	*No CSR*	*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)	(0.00223)	(0.0028)

Standard errors clustered at the district level in parentheses

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

8. Sensitivity Analysis

In addition to the OLS estimates of the lag score equation, the cohort effects were also estimated by the alternative VAMs discussed in Appendix B. Specifically, effects were estimated by a 2SLS version of the Arellano & Bond (1991) dynamic GMM estimator (FDIV) on the lag score equation used above or by OLS, random effects (RE), or fixed effects (FE) on a gain score equation obtained by constraining λ=1 in equation (6.1) and subtracting prior achievement from both sides of the equation. Table 7 displays the results from each of the estimation methods. 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.

Comparing columns (1) and (2) of Table 8 shows that the choice of estimating cohort effects by OLS on the lag score versus gain score equations makes little difference for the estimated

cohort effects. Comparing the preferred estimates in column (1) to the others reveals more variation; however, the main conclusion that can be drawn is invariant to the estimator chosen. Recall from above that students with teachers in the 2006-2007 cohort are estimated to score, on average, one-fiftieth of a standard deviation worse than those with teachers from the 2002-2003 cohort. Looking across the estimates in Table 8, all estimators suggest similar magnitudes of this effect with the largest being closer to 0.03 when estimated by FE.

Table 7: Cohort Effect Estimates from Alternative VAM Estimators

	(1)	(2)	(3)	(4)	(5)
Model	*Lag*	*Gain*	*Gain*	*Gain*	*Lag*
Estimator	*OLS*	*OLS*	*FE*	*RE*	*FDIV*
Entry Cohort					
2001-2002	-0.00345	-0.00269	0.00275	0.000374	0.00188
	(0.00331)	(0.00333)	(0.00788)	(0.00127)	(0.00388)
2002-2003	-0.00929***	-0.00846***	-0.0142**	-0.0114***	-0.0154***
	(0.00247)	(0.00260)	(0.00650)	(0.00136)	(0.00348)
2003-2004	-0.0162***	-0.0164***	-0.0273***	-0.0179***	-0.0248***
	(0.00465)	(0.00438)	(0.00501)	(0.00130)	(0.00257)
2004-2005	-0.0221***	-0.0220***	-0.0365***	-0.0231***	-0.0305***
	(0.00455)	(0.00448)	(0.00981)	(0.00129)	(0.00496)
2005-2006	-0.0304***	-0.0304***	-0.0442***	-0.0305***	-0.0387***
	(0.00237)	(0.00253)	(0.00661)	(0.00136)	(0.00415)
2006-2007	-0.0319***	-0.0313***	-0.0434***	-0.0300***	-0.0395***
	(0.00450)	(0.00464)	(0.0100)	(0.00146)	(0.00515)
2007-2008	-0.0265***	-0.0254***	-0.0166*	-0.0254***	-0.0245***
	(0.00472)	(0.00460)	(0.00918)	(0.00159)	(0.00489)
Observations	2,752,060	2,752,060	2,752,060	2,752,060	1,329,658
R-squared	0.653	0.034	0.399	0.016	--

Standard errors clustered at the district level (Cols (1), (2), (3), and (5)) or the student level (Col (4)) in parentheses; Grade-year standardized test scores are the dependent variable and the cubic in experience is included in all regressions

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

The robustness of the main result to alternative value-added approaches provides assurance that the relationship between cohort quality and CSR hiring requirements is not being driven purely by biases in the chosen estimator. Across all estimators, the effect of having a post-CSR

cohort teacher is quite small, suggesting that teacher quality, as measured by value-added, did not play a major role in the observed performance of CSR in State X.

9. 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 only slight decreases in achievement attributable to newly hired teachers 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, results of this paper 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.

This leaves the actual mechanism for why state-wide CSR programs do not achieve the expected gains an open question. Exploring the possibility that other inputs may have changed is an important next step. This is especially true in cases in which CSR was implemented without full funding, as has been suggested 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. 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 drops in the value-added of newly hired teachers, this drop may not lead to significantly lower long-run achievement on average. The conclusions of this

paper come with a few caveats. Specifically, these findings reflect the experience of a single state based on teachers in grades four through six. It is possible that in other states and in other grades the quality of incoming teachers is more responsive to the quantity hired.

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.

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

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

Chingos, M. M. (2010).The Impact of a Universal Class-size Reduction Policy: Evidence from Florida's Statewide Mandate. Program on Education Policy and Governance Working Paper 10-03.

Dieterle, S. (2011). Class-size Reduction Policies and the Composition of the Teacher Workforce. Unpublished draft.

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

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.

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.

Murnane, R. J. (1975). *The Impact of School Resources on the Learning of Inner City Children*. Cambridge, MA: Ballinger Publishing Company.

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.

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.

Wooldridge, J. (2002). *Econometric Analysis of Cross Section and Panel Data*. MIT Press: Cambridge, MA.

Appendix A: Additional Tables

Appendix Table 1: Descriptive Statistics

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

Appendix Table 2: Estimates from Pooled OLS Regressions

Specification	*Cohort*	*Cohort-by-Year*
Equation	*(6.1)*	*(6.2)*
Prior Math Score	0.706***	0.706***
	(0.00564)	(0.00564)
Asian	0.0947***	0.0947***
	(0.00515)	(0.00511)
Black	-0.137***	-0.137***
	(0.00347)	(0.00347)
Hispanic	-0.0273***	-0.0273***
	(0.00242)	(0.00244)
Other Race	-0.0239***	-0.0240***
	(0.00229)	(0.00231)
Female	-0.0160***	-0.0160***
	(0.00148)	(0.00148)
Disabled	-0.185***	-0.185***
	(0.0124)	(0.0125)
Free or Reduced Lunch	-0.0585***	-0.0584***
	(0.00141)	(0.00140)
Limited English	-0.0738***	-0.0742***
	(0.01000)	(0.0100)
Age	-0.0555***	-0.0554***
	(0.00322)	(0.00322)
Foreign Born	0.0706***	0.0706***
	(0.00354)	(0.00356)
Days Present	0.00109***	0.00108***
	(3.58e-05)	(3.56e-05)
Days Absent	-0.00500***	-0.00500***
	(0.000293)	(0.000293)
Experience	0.00731***	0.00502***
	(0.000890)	(0.000699)
Experience Sq	-0.000341***	-0.000231***
	(4.72e-05)	(3.40e-05)
Experience Cu	4.23e-06***	2.76e-06***
	(6.92e-07)	(4.39e-07)
Lagged Peer Score	0.0799***	0.0789***
	(0.0131)	(0.0131)

Class Size	8.97e-05	5.00e-06
	(0.000252)	(0.000258)
*Class Size*G5*	-7.95e-05	-2.58e-05
	(0.000412)	(0.000429)
*Class Size*G6*	-0.000535	-0.000540*
	(0.000328)	(0.000320)
Observations	2,752,060	2,752,060
R-squared	0.653	0.653

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

Appendix B: Measuring Teacher Quality

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 go up more with CSR.

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:

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

where
A_{it} is the achievement of student i in year t
X_{is} is a vector of family and student characteristics for student i in year s
E_{is} is a vector of education inputs for student i in year s
c_i is unobserved student ability
u_{is} is an idiosyncratic shock to student i's achievement in year s

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:

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

Typically, we 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:

(B.3) $$A_{it} = \zeta_t + \lambda A_{it-1} + X_{it}\beta_0 + E_{it}\gamma_0 + c_i + e_{it}$$
$$e_{it} = u_{it} - \lambda u_{it-1}$$

Up to now, the assumptions made on the original model in equation (B.1) have been primarily data-driven. At this point, there is some choice over further assumptions imposed on

the model. Under the assumptions that e_{it} is serially uncorrelated and that c_i is uncorrelated with the included inputs (or equal to zero)[9], equation (B.3), referred to as the lag score equation from here on, could be reasonably estimated by OLS.[10] 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, I will also consider other value-added models and estimators. Briefly, it is also common to assume that λ=1, and to subtract A_{it-1} from both sides of equation (B.3), yielding a gain score model of student achievement:

$$\Delta A_{it} = \zeta_t + X_{it}\beta_0 + E_{it}\gamma_0 + c_i + \nu_{it} \qquad \text{(B.4)}$$
$$\nu_{it} = u_{it} - u_{it-1}$$

Equation (B.4) could then be estimated by OLS, random effects (RE), or fixed effects (FE). 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, $(\lambda\text{-}1)A_{it-1}$, on the right hand side of equation (B.4), which will likely lead to an omitted variables bias. Importantly, OLS on (B.4) does not control for the unobserved student heterogeneity in any way. RE estimation recognizes the presence of the unobserved heterogeneity term, c_i, so much as it introduces serial correlation. However, RE does not allow for c_i to be correlated with teacher assignments. Moving from OLS

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

[10] 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 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.

to RE requires an additional assumption that the included regressors are strictly exogenous in (B.4). The strict exogeneity assumption is discussed in more detail below.

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.[11] Practically speaking, the strict exogeneity assumption precludes any feedback from realized achievement shocks to future inputs. For instance, if a principal reacts to a randomly good or bad test score in one year when determining a future teacher assignment, this would violate strict exogeneity. As noted by Rothstein (2009, 2010), the fixed effects approach is useful when assignment to teachers is made based on a static characteristic of 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 and Bond (1991) (see Koedel and 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.[12] 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-

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

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

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.