

# Class Size and Student Achievement: Experimental Estimates of Who Benefits and Who Loses from Reductions\*

Weili Ding  
Queen's University  
dingw@queensu.ca

Steven F. Lehrer  
Queen's University and NBER  
lehrers@queensu.ca

July 2008

## Abstract

Class size proponents draw heavily on the results from Project STAR to support their initiatives. Adding to the political appeal of these initiative are reports that minority and economic disadvantaged students received the largest benefits. We extend this research in two dimensions. First we conduct a more detailed examination of the heterogeneous impacts of class size reductions on measures of cognitive and non-cognitive achievement using both conditional and unconditional quantile regression strategies. Second to address correlated outcomes from the same treatment(s) we account for over-rejection of the null hypotheses by using multiple inference procedures. We present evidence of substantial heterogeneity in the impacts of class size reductions on measures of cognitive achievement. Our evidence indicates that higher ability students gain the most from class size reductions while many low ability students do not benefit from these reductions. Further, the multiple inference procedures render the few significantly differential impacts of smaller classes by race and free lunch status when the outcomes were assumed independent to appear at a frequency that one could reasonably expect due to chance.

\*We are grateful to Alan Krueger for providing a subset of the data used in the study. We wish to thank Caroline Hoxby and Richard Murnane for suggestions which helped improve this paper. Lehrer wishes to thank SSHRC for research support. We are responsible for all errors.

# 1 Introduction

Unlike vouchers, charter schools, teacher testing, and other controversial reform strategies, class size reduction (CSR) proposals have intuitive and political appeal. Parents assume that their children will get more individualized instruction and attention, thereby improving student achievement, and teachers believe that it gives them a shot at creating true learning communities. In 2004, 33 states had laws that restricted class size at the K-3 level and new federal and state/provincial legislation and appropriations will promote further shrinkage of class sizes in North America. Policymakers continuously draw from the reported experience of Project STAR, a randomized evaluation in the late 1980s on the impacts of CSR in Tennessee to support the launch of multi-billion dollar CSR initiatives.<sup>1</sup>

Two issues have been largely ignored in the discussion of the results from Project STAR. First, if the prescription of smaller classes does not benefit students in an equal manner then an improved and more comprehensive understanding of which group of pupils received the largest benefits is needed.<sup>2</sup> To assess the distributional effects of class size reductions we consider unconditional quantile regression to determine where the treatment effects are concentrated in the test score distribution.<sup>3</sup> From a policy perspective estimating quantile impacts of inputs to an education production function (in addition to mean impacts) is likely of importance since societal costs associated with poor development of cognitive and non-cognitive skills exist primarily at the low end of the achievement distribution, with the costs increasing substantially at the very low end. Additionally, we examine whether there are heterogeneous impacts of small class by race, economic background and school characteristics accounting for a more comprehensive set of possible interactions between individual and school factors. Finally, we consider traditional quantile regression strategies that explicitly allow the impacts of smaller classes to vary across the conditional achievement distribution.<sup>4</sup>

Second, as students in Project STAR completed a battery of exams in each year, a special set of techniques are needed to evaluate whether CSR is effective with multiple outcomes. These techniques incorporate the dependence in student test scores across multiple subjects for the same student. Failing to account for multiple outcomes from the same treatment(s) may lead to finding significant impacts when there are none.<sup>5</sup> For example, if the effectiveness of CSR is assessed on six outcomes each at a significance level of 5% (two-sided tests), the chance of finding at least one false positive statistically significant test increases to 15.9%. Accounting for multiple outcomes can have a substantial influence on the rate of false positive conclusions which may affect education policy whenever there is an opportunity to select the most favorable results from an analyses; as without choice there is no influence. We adopt two multiple testing procedures that i) controls for the probability of at least one rejection of a true null hypothesis, ii) allow the number of false rejections one is willing to tolerate to vary with the total number of rejections, to present a more detailed analysis of CSR effectiveness.<sup>6</sup>

We also investigate the impacts of CSR on non-cognitive skills such as listening, motivation and self-concept using multiple inference and allowing for flexible heterogeneity. The majority of Project STAR research has focused solely on test scores in reading, mathematics and word recognition. Several researchers have criticized the focus of education policy on cognitive skills and shown the importance of non-cognitive skills on a variety of education and labor market outcomes.<sup>7</sup> With the recent public availability of measures of non-cognitive performance in Project STAR, we have a chance to examine whether CSR has appealing impacts on non-cognitive skills.<sup>8</sup>

This paper is organized as follows. In the next section we provide a brief review of the Project STAR experiment and describe the data used in this study. In order to minimize issues related to non-random violations to the experimental protocol that occurred in subsequent years of the study affecting the estimates, we only report analysis using data collected in the first year of the experiment.<sup>9</sup> The statistical approaches employed are

discussed and empirical results are reported in Section 3. We find strong evidence that i) students with higher test scores received larger impacts from CSR, ii) higher ability students gain the most from CSR while many low ability students do not benefit from these reductions, iii) there are few additional benefits from CSR for minority or disadvantaged students. Some of the differences between our findings and earlier work are related to a) treating the impacts of school factors in a more general manner and b) the different ways in which researchers use test score measures as outcome variables. Multiple inference procedures that account for general correlations in student outcomes between subject areas suggest that the impacts of CSR had positive impacts on measures of cognitive achievement but did not yield non-cognitive benefits. Moreover, these procedures render the few significantly differential impacts of CSR by race and free lunch status observed when the outcomes are treated as independent to appear at a frequency that one could reasonably expect to be due to chance. A concluding section summarizes our findings and discusses directions for future research.

## 2 Project STAR Experiment

The Student/Teacher Achievement Ratio (STAR) was a four-year longitudinal class-size study funded by the Tennessee General Assembly and conducted by the State Department of Education. Over 7,000 students in 79 schools were randomly assigned into one of the three intervention groups: small class (13 to 17 students per teacher), regular class (22 to 25 students per teacher), and regular-with-aide class (22 to 25 students with a full-time teacher's aide) as the students entered kindergarten. Teachers were also randomly assigned to the classes they would teach.

In theory, random assignment circumvents problems related to selection in treatment. However, following the completion of Kindergarten there were significant non-random movements between control and treatment groups as well as in and out of the sample

which complicates any analysis.<sup>10</sup> By grade three over 50% of the subjects who participated in kindergarten left the STAR sample and approximately 10% of the remaining subjects switch class type annually. Ding and Lehrer (2008) present evidence of selective attrition and demonstrate that the conditional random assignment of the newly entering students failed in the second year of the experiment as among this group of students those on free lunch were significantly more likely to be assigned to regular (larger) classes.<sup>11</sup> To reduce concerns regarding potential biases from non-random violations to the experimental protocol, our analysis focuses solely on data from the first year of the experiment.<sup>12</sup>

At the end of the kindergarten year the majority of the students completed six exams to measure their performance in different dimensions. The students completed the Reading, Listening Comprehension, Mathematics and Word Recognition sections of the Stanford Achievement test.<sup>13</sup> In our analysis, we employ total scaled scores by each subject area. Scaled scores are calculated from the actual number of items correct adjusting for the difficulty level of the question to a single scoring system across all grades. Scaled scores vary according to the test given, but within the same test they have the advantage that a one point change on one part of the scale is equivalent to a one point change on another part of the scale.<sup>14</sup> While the instructional objectives of the listening comprehension component are similar to that of the reading test, they focus on a different set of skills that measure ability to comprehend spoken communication.<sup>15</sup> Finally, the students completed the Self-concept and Motivation Inventory test presenting measures of two non-cognitive skills: self-concept and motivation, which are obtained by the child's response to 24 questions that are prefaced with the statement, "What face would you wear if ...". The student selects and blackens the tone of one of five different faces for each question. Even at the kindergarten level these tests have been found to have moderate internal consistency (Davis and Johnston (1987)), and is scored from 24 to 120, with higher scores indicating more positive outcomes. The motivate inventory is scored from 8 to 40 and the self concept scale ranges from 16 to 80.

The public access data on Project STAR contains information on teaching experience, the education level and race of the teacher, the gender, race and free lunch status of the student. Summary statistics on the Project STAR kindergarten sample are provided in Table 1. Between 79.7% to 92.5% of the participants completed each of the examinations as some were not offered in certain schools or some students were absent on certain test days. Nearly half of the sample is on free lunch status. There are few Hispanic or Asian students and the sample is approximately  $\frac{2}{3}$  Caucasian and  $\frac{1}{3}$  African American. There are nearly twice as many students attending schools located in rural areas than either suburban or inner city areas. There are few students in the sample (9.0%) attending schools located in urban areas. Regression analysis and specification tests found no evidence of any systematic differences between small and regular classes in any student or teacher characteristics in kindergarten, suggesting that randomization was indeed successful. However, among black students those on free lunch status were more likely to be assigned to regular classes than small classes (33.67% vs. 27.69%,  $\Pr(T > t) = 0.0091$ , one sided test).

## 3 Empirical Results

### 3.1 Quantile Treatment Effects

#### 3.1.1 Conditional Quantile Regression

In our work we first allow the effect of class size to vary for individuals at the different points of the conditional test score distribution which we view as reflecting the distribution of unobserved ability. We estimate the following contemporaneous achievement education production function by quantile regression for each component of the Stanford Achievement Test<sup>16</sup>

$$A_{ij} = \beta' X_{ij} + \beta'_{CS} CS_{ij} + \varepsilon_{ij}. \quad (1)$$

where  $A_{ij}$  is the level of achievement for child  $i$  in school  $j$ ,  $X_{ij}$  is a vector of school indicators, student and teacher characteristics,  $CS_{ij}$  is an indicator if student  $i$  attended a small class,<sup>17</sup>  $\varepsilon_{ij}$  captures random unobserved factors. Controlling for school effects is necessary since randomization was done *within* schools. By randomly assigning class type and teachers to students,  $CS_{ij}$  is uncorrelated with unobserved factors such as the impact of pre-kindergarten inputs, family and community background variables, etc., permitting unbiased estimates of  $\beta_{CS}$  with only kindergarten data. Quantile regression provides a flexible approach to characterizing the effects of observed covariates such as class type on different percentiles of the conditional achievement distribution. This allows us to investigate whether small class induces a more complex change in the test score distribution as opposed to a simple constant impact. Implicitly we are allowing class size and ability to be two separate factors in the generation of achievement to interact in unknown ways. If ability and class size are substitutes we would expect the marginal returns on class size to decrease when ability is increasing. If ability and class size are complements then marginal returns to class size would be higher for the more able.

The quantile regression results for class size coefficients from equation (1) are presented in Figure 1. In all the subject areas higher ability students benefit more from reduced class sizes, indicating that smaller class size complements unobserved ability. Students in the lowest quantiles (0.05 and 0.10) do not gain from smaller classes as the benefits are not statistically different from zero. In all subjects students in the highest quantile (0.95) gain a larger amount from a small class compared with the OLS coefficient estimates. There are substantial difference between the OLS and quantile regression class size coefficients in the extreme quantiles in all subject areas, whereas the other quantiles have impacts similar to OLS. In particular, in word recognition the quantile regression coefficients differ greatly in magnitude from the OLS estimates in the extreme quantiles.

### 3.1.2 Unconditional Quantile Regression

We next allow the effect of class size reductions to vary for individuals at the different points of the unconditional test score distribution. We estimate the contribution of each explanatory variable to the unconditional quantiles of test scores, which permits us to answer question such as what is the impact on a specific quantile of math test scores of assigning everyone to a small class, holding everything else constant. To better interpret the estimates we present information on the quantiles of each test score distribution in Appendix Table 1.

We estimate equation (1) via the Firpo, Fortin and Lemieux (2007) regression method that essentially replaces the original outcome variable ( $A_{ij}$ ) by a simple transformation known as the recentered influence function. The recentered influence function for the quantile of interest  $q_\tau$  is formally defined as

$$RIF(A; q_\tau) = q_\tau + \frac{\tau - I(A \leq q_\tau)}{f_A(q_\tau)} \quad (2)$$

where  $f_A$  is the marginal density function of  $A$ ,  $I$  is an indicator function . Since the  $RIF(A; q_\tau)$  defined in equation (2) is unobserved in practice, we use its sample analog that replaces the unknown quantities by their estimators as follows

$$RIF(A; \hat{q}_\tau) = \hat{q}_\tau + \frac{\tau - I(A \leq \hat{q}_\tau)}{\hat{f}_A(q_\tau)} \quad (3)$$

where  $\hat{q}_\tau$  is the  $\tau$ th sample quantile and  $\hat{f}_A$  is the kernel density estimator. Once the dependent variable is replaced by the transformation defined in equation (3) a simple OLS regression allows us to recover the impact of changes in the explanatory variables on the unconditional quantiles of  $A_{ij}$ . Intuitively at each quantile this procedure changes the outcome variable in equation (1) in such a way that the mean of the recentered influence function corresponds to the statistic of interest.

Figure 2 presents by subject area unconditional quantile regression estimates of the impact of attending a small class on levels of kindergarten achievement.<sup>18</sup> These esti-

mates provide more information about the impact of class size reductions than the OLS estimates. They show at different achievement levels how important attending a small class is. Notice that in all the subject areas in the top row of Figure 2 (mathematics, reading, word recognition) there is clear evidence of heterogeneity that the OLS estimate is often not captured within the 95% confidence interval. In addition, tests of treatment effect homogeneity between quantiles are firmly rejected. In these subjects students with higher test scores have received larger impacts from being assigned to a small class than students in the lower quantiles of the test score distribution. Additionally, those students in the lowest quantiles in mathematics do not receive a statistically significant impact from being assigned to a small class. In contrast, examining the test scores presented in the bottom row of Figure 2 (listening, self-concept and motivation) we do not find significant evidence of treatment effect heterogeneity. Interestingly small class only has a significantly positive impact on test scores in motivation at 7 of the 19 quantiles.

By exploring treatment effect heterogeneity we are attempting to enter the “black box” of CSRs. Our evidence indicates that there was considerable heterogeneity in the impacts of small classes on the distributions of test scores in mathematics, reading and word recognition; heterogeneity that would be left unexplored by only reporting mean impacts. In particular, we find that the impact of small classes is not significantly different from zero in the bottom 20% of the math distribution and in over 60% of the quantiles of the motivation test score. To improve the effectiveness of class size reductions one could simply target students who have larger responses to the intervention. reforms. In the next section, we take a closer look at how the relationship between class size and achievement varies across subgroups that are easy to identify.

### **3.2 Education for the Disadvantaged**

Class size reductions have played a large role in recent policy debates searching for mechanisms to reduce the achievement gap between disadvantaged children and other children.

The reported positive results that CSR is more beneficial for minority and inner city children has substantial political appeal. While the quantile regression results presented evidence that in subsamples defined by race or free-lunch status, individuals in the higher quantiles received larger gains than their counterparts in the lower quantiles of the conditional achievement distribution, in this subsection we conduct a more comprehensive examination of whether students on fee lunch and minority children gain more in small classes on average. To accomplish this goal, we interacted the individual student and teacher characteristics with class size and estimated the following equation

$$A_{ij} = \beta' X_{ij} + \beta'_{CS} CS_{ij} + \beta'_{XCS} CS_{ij} X_{ij} + \varepsilon_{ij} \quad (4)$$

The results for all six subjects are presented in the Table 2. Notice from the bottom row the interaction terms are jointly insignificant at conventional levels in all subject areas with the exception of self-concept. Further, the interaction between small class and free lunch status was individually insignificant in all subject areas. Similarly African Americans students did not perform significantly different in smaller classes with the exception of self-concept skills but this effect is only significant at the 10% level.

In order to consider the most flexible method to evaluate whether there was heterogeneity in the impact of small class treatment across groups we consider a fully saturated model that contains all possible interactions between student and teacher covariates. The results are presented in Table 3. The effect of small class is highly significant in math, reading and word recognition, but the only interaction between small class and another input that is significantly related to academic performance is that with the indicator for being female in a small class on the motivation exam. In specifications that consist of the full set of interactions there are substantially large negative impacts for being a minority or free lunch student. Further, the interaction between black and free lunch status is highly significant and positively relate to achievement on all four Stanford Achievement tests. F tests on the full set of interaction terms indicate that they are jointly significant

on the mathematics, listening and reading examination. F-tests on the joint significance of the individual demographic characteristics and small class indicators are only significant on the self-concept and motivation tests.<sup>19</sup> However, the inclusion of this large set of regressors only explains a limited amount of the variation in self-concept and motivation scores.

These discrepancies between our results and earlier work are due to two major features. First, prior work ran multiple regressions separately on samples defined by class types and then compared the magnitude of the estimated coefficient on the free lunch variable as opposed to pooling the sample and including interaction terms. By running the regressions on two separate subsamples defined by class assignment does not restrict the student invariant school effects nor the other covariates to have the same impacts across subsample. This could distort inference if there are some partial correlations between these variables. Further, many of these comparisons were exploiting variation across schools and do not account for school heterogeneity. Since randomization was done within and not between schools, these comparisons ignore the experimental variation which provides exogenous variation to identify any impacts. Further, only 34% of the African American and Hispanic students in the full kindergarten sample attend schools that also contain white or Asian students. In fact, there are 15 schools that only consist of minority students and 15 schools for which there was not a single minority student in the kindergarten sample. Thus, using raw differences from specification held on samples defined by class types could lead the results to be confounded by factors that vary across schools. In contrast, the pooling approach not only includes school indicators and exploits the experimental variation but the interpretation of the interaction terms from a regression using the pooled sample as intercept or slope shifts is straightforward. Lastly, there are gains in efficiency of the estimates by using the full sample.

Second the method in which student performance is measured varies substantially across samples. In our study we used scaled scores for outcomes from the Stanford

Achievement test since they are developmental. Within the same test they have the advantage that a one point change on one part of the scale is equivalent to a one point change on another part of the scale allowing one to measure growth across grades as well as within grades. Scaled scores represent the sole method of determining whether a student actually demonstrated growth in a subject area but cannot be combined across subjects. Alternative methods to estimate student performance with STAR data represent monotonic transformations of scaled scores or raw scores. These methods include percentile scores, standard scores and grade equivalent scores. Percentile scores represent ranks within a sample and simply provide the percentage of students whose scores were at or lower than a given score. While useful to compare a student's performance in relation to other students, they create a uniform distribution that places too much weight on scores near the mean when estimating equations via OLS. To construct standard scores researchers assume that any non-normality in the observed distribution of test scores is an artifact and convert each percentile point into the standard score that would correspond to that percentile in a normal distribution. Standard scores provide a measure of how much standard deviation one's score is from a mean and provide an equal unit of measurement on a single test. However, they are not developmental and cannot be used to measure growth within a subject area or combined across subjects. Further, it is much easier to interpret marginal effects and translate results to policymakers with scaled scores as these adjust for difficulties in test scoring which could occur from ceiling effects. To illustrate, consider a 10 percentile score increase on the kindergarten math exam from our sample. A move from the median to the 60th percentile is equivalent to moving 10 scaled points or  $0.018\sigma$ , whereas moving from the 80th to the 90th percentile involves 27 scaled points or  $0.294\sigma$ . The transformations from one measure to another changes the variation in outcome scores to be explained by the regressors. The relationship between standard scores, percentiles scores and scaled scores also varies from test to test. We replicated all of the analysis in Tables 2 and 3 with both standard scores and percentile scores and

there were several differences in the significance of the findings.<sup>20</sup> While the methods to specify dependent variables in labor economics and health economics studies have been an active area of study where dependent variables have i) nonnegative outcomes and ii) skewed outcome distributions, this has been an understudied area in the economics of education literature that we believe warrants further investigation.<sup>21</sup>

We replicated the analysis in Table 2 with a sample of students from inner city schools and compared the coefficients obtained to those of the sample of students from other schools. We did not find any significant differences for the class size variable or interactions, lending little support to the claim that the impacts of smaller classes are significantly larger in inner city schools.<sup>22</sup> The discrepancy between our work and earlier studies comes from the inclusion of school effects which are required since randomization was done within and not across schools.<sup>23</sup>

Finally, we attempted to investigate the significant heterogeneity in the effectiveness of CSR across schools.<sup>24</sup> Since there does not appear to be a consistent relationship between kindergarten class size and academic achievement understanding why it works in some schools but not in others is essential. We categorized schools on the basis of whether CSR was effective in each of mathematics, reading and word recognition subject tests. We then tried to determine in those schools that small classes outperformed regular classes in all 3 subject areas if there were systematic differences in any of the school or teacher factors. Unfortunately the publicly available STAR data did not yield many insights into the sources of program heterogeneity and future work requires more data collected during the process evaluations that are currently unavailable to outside researchers. While the above analysis consisted of a more comprehensive examination of the heterogeneous impacts of CSR in the first year of Project STAR, we treated each test score outcome as independent and there are many reasons to believe that there are substantial positive correlations in performance across subject areas. We next account for these factors and examine whether the above results and the effectiveness of CSR are robust to multiple inference.

### 3.3 Multiple Test Outcomes

Making adjustments for the use of multiple outcomes has a long history in psychology (Benjamini and Yekutieli (2001)) and biostatistics (Hochberg (1988)). These techniques have also been adopted in some studies within education (Williams et al. (1999)) as well as studies in economics that examine multiple child outcomes (Kling and Liebman (2004) and Andersson (2007)). The motivation for these tests is that without accounting for the fact that outcomes collected within the study are related one may over reject the Null hypothesis of no treatment effects when using univariate statistical methods. Therefore one needs to adjust the p-value for the multiple outcomes and we consider making corrections for both the Familywise error rate (FWER) and false discovery rate (FDR). These p-value adjustments reduce the chance of making type I errors and are based on the number of outcomes being considered. Formally, suppose that we wish to test  $K$  hypotheses,  $H_1, H_2, \dots, H_k$  of which only  $l < K$  are true, the FWER is simply the probability of making one or more type I errors (i.e. one of  $l$  true hypotheses in the family is rejected) among all the single hypotheses when performing multiple pairwise tests on a families of hypotheses that are similar in purpose. We consider three families in our analysis. The first family consists of all six student performance examinations and we also separately consider the three measures in the cognitive and non-cognitive domains. While, the FWER controls for the probability of making a Type I error,<sup>25</sup> we also consider accounting for the FDR rate which controls the expected proportion of incorrectly rejected null hypotheses (Type I errors) in a list of rejected hypotheses. It is a less conservative procedure with greater power than FWER control, at a cost of increasing the likelihood of obtaining type I errors.

For the FWER, we use the free step-down method (Holland and Copenhaver, 1987) that allows the different p-values (which are clustered at the classroom level) to be arbitrarily correlated. The two-step procedure developed in Benjamini, Krieger and Yekutieli (2006) is used for the FDR since Benjamini et al (2006) present evidence from simulations

that the algorithm performs well when p-values are positively correlated across tests (as in our case) therefore providing sharper control. If all Null hypotheses are true, controlling for the FWER is equivalent to accounting for the FDR; however as increasingly more alternative hypotheses are true controlling for the FDR can result in fewer Type II errors than controlling for the FWER.

Table 4 reevaluates the evidence on the mean effectiveness of CSR in kindergarten by adjusting the estimates presented in the text for multiplicity. The first three columns lists the Null hypotheses being tested, the specifications of the estimation equations and the number of subjects that are being examined together. The fourth and fifth columns reports the number of outcomes that are statistically significant when tested independently at the 5% and 10% significance levels respectively. The next two columns present the number of Null hypotheses rejected using the Holland and Copenhaver (1987) method at the 5% and 10% significance levels. The last two columns correspond to the previous two except report the number of rejections when accounting for the FDR using the Benjamini et al (2006) procedure.

The first three rows of Table 4 reexamines the effectiveness of CSR from OLS estimates of equation (1). We find that accounting for multiplicity leads to rejecting the positive impact of CSR on the motivation exam at the 10% level, which assuming independence failed to reject. However, the effects of accounting for multiple tests on the conclusions one can draw from Project STAR are more dramatic when we consider estimates from the fully saturated model in Table 3. When examining all six tests together, the effects of CSR on three cognitive achievements are significantly different from zero at the 5% level when the tests are treated as independent, while the effects of CSR on non-cognitive achievements are insignificant in the saturated model. Once we account for multiplicity only the impact of CSR on reading remains significant at the 5% level. This result holds whether we correct for the FDR or FWER. At the 10% level, the impact of CSR on mathematics becomes significant when we account for the FWER whereas accounting for

FDR yields equivalent results to assuming independence.

Evaluating the effects of small classes interacted with student demographics with multiple inference procedures suggests that the few differential impacts of CSR by race and free lunch status when the outcomes are independently assessed now appear at a frequency that one could reasonably expect due to chance. None of the interactions between small class and student demographic characteristics are significant at the 5% level even when we assume the outcomes are independent, with the exception of being a female student in a small class on motivation assessment. With the criterion of 10% significance level, there is positive impact of being an African American student in a small class on the self-concept evaluation when assuming independence. This positive result becomes insignificant once we account for multiple inference by FWER or FDR. Taken together, this casts doubt that there are truly heterogeneous mean impacts from CSR across groups defined by race or free lunch status in kindergarten.<sup>26</sup>

## 4 Conclusion

This paper provides new evidence in one of the most active and highly politicized subject areas in the education reform debate: the effects of reduced class size. Our empirical analysis of the STAR project complement existing studies using this data by first demonstrating that higher ability students gain the most from CSR while many low ability students do not benefit from these reductions in Kindergarten. These differences in performance across the conditional achievement distribution also holds in subsamples defined by race and free lunch status. Second, we also explored heterogeneity by estimating the impact of attending a small class on the distribution of test scores. We find substantial heterogeneity in these impacts in the cognitive subject areas as students with higher test scores receive larger impacts from small classes. Third, using statistical corrections for multiple inference, we do not find any evidence in Kindergarten for additional benefits

from CSR for minority or disadvantaged students. It may well be that CSR are more effective for some groups of students defined by alternative criteria in which case policy might be more effective targeting specific populations rather than mandating across the board reductions. Finally, we find mixed evidence on the effectiveness of CSR in kindergarten as it leads to significant improvement in cognitive achievement measures but does not appear to provide substantial benefits to the development of non-cognitive skills.

Understanding why CSRs were only effective in some subjects but not others is clearly a direction for future research. Moreover, since there was substantial heterogeneity in the effectiveness of CSR across schools, increased attention should be paid to determine why some schools were able but other schools were not to translate smaller classes into gains in student cognitive achievement.<sup>27</sup> As teaching practices varied across and within schools uncovering whether certain practices are partially responsible for the extent of heterogeneity in treatment effectiveness is important for education policy. In conclusion, we suggest that the substantial heterogeneity in the impacts from class size reduction witnessed in Kindergarten should promote further investigation by integrating qualitative data collected in the process evaluations into quantitative empirical analyses to improve our understanding of how class size affects the production of education outcomes.

## Notes

<sup>1</sup>For example, The US Department of Education in a 1998 report titled “Reducing Class Size: What Do We Know?” states “In sum, due to the magnitude of the Project STAR longitudinal experiment, the design, and the care with which it was executed, the results are clear: This research leaves no doubt that small classes have an advantage over larger classes in student performance in the early primary grades.”

<sup>2</sup>For example, the American Education Research Association (2003) concludes their research summary by stating "There is no doubt that small classes can deliver lasting benefits, especially for minority and low-income students." Past research with Project STAR data has reported that i) minority students receive at least twice the small class benefit (Finn and Achilles (1990) and Finn (2002)), ii) larger gains are received in inner-city schools relative to urban, suburban and rural schools (Pate-Bain et al. (1992)), and iii) small classes reduced the gap between students who were economically eligible for the free lunch program versus those students who were not eligible (Word et al (1990)). These studies have reported larger gains for disadvantaged students, thus increasing the political appeal of CSR policies. Yet, much of this research has employed statistical models that allow for limited forms of heterogeneity.

<sup>3</sup>To the best of our knowledge this is the first paper in the economics of education literature to employ unconditional quantile regression.

<sup>4</sup>Quantile regression and unconditional quantile regression strategies differ substantially. Quantile regression estimates are harder to interpret for a policy audience but have a structural interpretation as they are conditional quantile functions that are multidimensional whereas estimates from unconditional quantile regression present estimates of the effects of a treatment for the entire population as the unconditional quantile function is a one dimensional function.

<sup>5</sup>This problem is well known in the theoretical statistics literature (Romano and Wolf (2005)) and the importance of making adjustments to the inference procedure when evaluating the impact of an intervention on multiple outcomes is also stressed by practitioners such as the U.S. Department of Education’s Institute of Education Sciences What Works

Clearinghouse to determine whether an education intervention is truly effective. See [http://www.whatworks.ed.gov/reviewprocess/rating\\_scheme.pdf](http://www.whatworks.ed.gov/reviewprocess/rating_scheme.pdf) for details.

<sup>6</sup>In contrast, earlier research has either combined a subset of the outcome measures collected into a single index using arbitrary weights or examined each of these outcomes independent of the others.

<sup>7</sup>This is of policy relevance since research has shown that noncognitive skills influence individual performance on cognitive tests (Borghans et al. (2006)), likelihood of school dropout (Heckman and Rubinstein (2001)) and amount of schooling obtained (Heckman, Stixrud and Urzua (2006)).

<sup>8</sup>Word et al. (1990) report not finding a significant impact of CSR with a couple of the non-cognitive measures.

<sup>9</sup>We only examine relationships in the first year of the study to present evidence that imposes minimal assumptions on the analysis. There are several complications that arose in the experiment including i) students switching between class types, ii) selective attrition and refreshment samples being assigned to class type in a non-random manner. Dealing with each of these violations to the experimental protocol results in separate series of assumptions being made as in Ding and Lehrer (2008); which presents evidence on each of the above violations. We have replicated the results and they are also robust to using the full sample of kindergarten students in higher grades where the samples are reweighted by either series logit estimates of the probability of remaining in the sample or the probability of writing the exam the previous academic year.

<sup>10</sup>The STAR experiment not only witnessed attrition in students but also in schools. Six schools left the study prior to the end of grade 3 and five of these schools left immediately after kindergarten.

<sup>11</sup>It should also be noted that attendance of kindergarten was not mandatory in Tennessee and students who entered school in grade 1 may differ in unobservables to those started in kindergarten.

<sup>12</sup>As discussed in footnote 9 the general pattern of our results holds in subsequent

years where we corrected for subsequent selection on observables. These analyses impose additional behavioral assumptions and are available upon request.

<sup>13</sup>The Stanford Achievement Test is a norm-referenced multiple-choice test designed to measure how well a student performs in relation to a particular group, such as a representative sample of students from across the nation. Norm-referenced tests are commercially published and are based on skills specified in a variety of curriculum materials used throughout the country. They are not specifically referenced to the Tennessee curriculum.

<sup>14</sup>As we discuss in Section 3.2, the selection of scores is of critical importance in interpreting the results and much of the previous work has employed transformations of the scaled scores as outcome variables, which has major effects upon their results.

<sup>15</sup>To a large extent each of these test scores may reflect a combination of cognitive and non-cognitive skills. This breakdown between cognitive and non-cognitive skills is based in part on behavior of college admission committees who consider listening a non-cognitive skill whereas reading a cognitive skill (Streyffeler et al. (2005)).

<sup>16</sup>There is limited variability in both the motivation and self-concept scores leading to substantial difficulties with the algorithm for quantile regression estimates in these subject areas.

<sup>17</sup>Following Finn et al. (2001) and Krueger (1999) our control group consists of regular class with and without teacher aides, as these studies (among others) report that the presence of a teacher aide did not significantly impact student test scores.

<sup>18</sup>The estimates of the impacts of the other explanatory variables on the quantiles of the achievement distribution are available from the authors upon request.

<sup>19</sup>We also considered simpler specifications that only considered interactions between the inputs and either the race or free lunch variable. The results are presented in Appendix Table 2. With the exception of self-concept, the effects of class size interacted with either being black or being economically disadvantaged are statistically insignificant.

<sup>20</sup>In particular, with both standard and percentile scores one would conclude that small classes affected performance on the self-concept exam at the 5% level but not the listening test at the 10% level. In total there are 7 and 9 differences in the significance of various

interactions for these alternative rescalings. We could not convert the scores to grade equivalent which tells the students standing in relation to the norm group at the time of testing. Yet, interpretation of these scores is confusing and these scores are known to have low accuracy for students with very high or low scores. Further, they are inappropriate to use for computing group statistics or in determining individual gains.

<sup>21</sup>Appendix Figure 1 presents Kernel density estimates of the distribution of test scores in all six-subject areas and demonstrates how several of these tests have a skewed or bimodal distribution, which differs from a Normal distribution. Blackburn (2007) discusses issues how the selection of dependent variables in wage regressions can lead to difficulties in interpreting coefficients. Similarly, Manning and Mullahy (2001) provide guidance on appropriate estimators for health economists when the dependent variable has or has not taken the log transformation.

<sup>22</sup>The results are available from the authors by request.

<sup>23</sup>We additionally replicated the analysis presented in Figures 1 and 2 on the subsample of students who were eligible for free lunch in kindergarten as well as on the group of African American and Hispanic children. These graphs are available upon request and we continue to find significant heterogeneity in the impacts of small classes on measures of achievement for both the subsample of students on free lunch and African American students. Notice that the patterns are nearly identical as both students on free lunch and African American students in higher quantiles of the conditional achievement distribution receive larger impacts from smaller classes than students in the lower quantiles. For example, African American students in the highest quantile receive over 5 times the benefits on mathematics from small class relative to students in the smallest quantile. Those on free lunch student in the highest quantile receive over 4 times the benefit relative to those in the lowest quantile on mathematics. Interestingly, the number of quantiles in which small class is not statistically different from zero on achievement is higher for these subsamples.

<sup>24</sup>This result is also reported in Krueger (1999).

<sup>25</sup>The FWER maintains the overall probability of making a Type I error at a fixed  $\alpha$  (i.e. 5%) but with an ever increasing number of tests this comes at the cost of making more Type II errors. The sequential procedure we use performs tests in order of increasing

p-values with smaller p-values tested at a tougher threshold to maintain the FWER at a desired level.

<sup>26</sup>Further, if we expand the analysis to evaluate the estimates in Table 2 and Appendix Table 2, we find that none of the interactions between small class and student demographics remains significant once we account for the FDR or FWER with the sole exception of the interaction between small class and being a female student on the motivation.

<sup>27</sup>There is limited examination in the economics of education literature on how class size may affect student achievement. It has been hypothesized that the teacher will have more time to transmit knowledge and exert less effort to discipline (Lazear (1999)). Among other claimed benefits are better assessment techniques, more small group instruction and students becoming less passive. The available evidence suggests that teaching practices do not vary with class size as hypothesized. For example, Betts and Shkolnik (1999) find no association between class size and text coverage and correspondingly no more time devoted to material in one class over another even after controlling for teacher fixed effects. Yet they do find teachers in large classes spent more time on discipline and less time on individualized attention. Finally, Shapson et al., (1980) present experimental evidence on teacher behavior across 4 class sizes (16, 23, 30 or 37 students). The authors conducted a two-year study of 62 Toronto area classes of grade four and five students from eleven schools. They found that class size makes a large difference to teachers in terms of their attitudes and expectations, but little or no difference to students or to instructional methods used. Teachers in class sizes of 16 and 23 were pleased because they had less work to do in terms of evaluating students' work, than did the teachers with larger class sizes. They conclude that teachers need to be trained in instructional strategies for various size classes.

## References

- [1] American Education Research Association (2003), “Class Size: Counting Students Can Count Research Points,” *Research Points*, 1(2), 1 - 4.
- [2] Anderson, M. (2007) “Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects,” forthcoming in the *Journal of the American Statistical Association*.
- [3] Benjamini, Y., and D. Yekutieli (2001) “The Control of the False Discovery Rate in Multiple Testing Under Dependency,” *The Annals of Statistics*, 29(4), 1165 – 1188
- [4] Benjamini, Y., A. Krieger, and D. Yekutieli (2006) “Adaptive Linear Step-up Procedures that Control the False Discovery Rate,” *Biometrika*, 93(3), 491 – 507
- [5] Betts, J. R. and J. L. Shkolnik (1999), “The Behavioral Effects of Variations in Class Size: The Case of Math Teachers,” *Educational Evaluation and Policy Analysis*, 21(2), 193 - 213.
- [6] Blackburn, M. L. (2007), “Estimating Wage Differentials Without Logarithms,” *Labour Economics*, 14(1), 73 - 98.
- [7] Borghans, L., B. ter Weel, and B. Weinberg (2006), “Interpersonal Styles and Labor Market Outcomes,” forthcoming in *Journal of Human Resources*.
- [8] Ding W. and S. F. Lehrer (2008) “Estimating Treatment Effects from Contaminated Multi-Period Education Experiments: The Dynamic Impacts of Class Size Reductions,” forthcoming in the *Review of Economics and Statistics*.
- [9] Finn, J. D, S. B. Gerber, C. M. Achilles, M. Charles and J. Boyd-Zaharias, Jayne (2001), “The Enduring Effects of Small Classes.” *Teachers College Record*, 103(2), 145-83.
- [10] Finn, J. D., and C. M. Achilles (1990), “Answers about Questions about Class Size: A Statewide Experiment,” *American Educational Research Journal*, 27, 557 - 577.
- [11] Firpo, S., N. M. Fortin and T. Lemieux (2007), “Unconditional Quantile Regressions,” *NBER Working Paper T339*.
- [12] Hamushek, E. A. (1999b), “Some Findings from an Independent Investigation of the Tennessee STAR Experiment and from Other Investigations of Class Size Effects,” *Educational Evaluation and Policy Analysis*, 21, 143 - 163.

- [13] Heckman, J. J. and Y. Rubinstein (2001), "The Importance of Noncognitive Skills: Lessons from the GED Test Program," *American Economic Review*, 91(2), 145 - 149.
- [14] Heckman, J. J., J. Stixrud and S. Urzua (2006), "The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior," *Journal of Labor Economics*, 24(3) 411 - 482.
- [15] Hochberg, Y. (1988), "A Sharper Bonferroni Procedure For Multiple Tests of Significance," *Biometrika*, 75(4), 800 - 802.
- [16] Holland, B. and M. D. Copenhaver (1987), "An Improved Sequentially Rejective Bonferroni Test Procedure," *Biometrics*, 43, 417 - 442.
- [17] Kling, J. R., and J. B. Liebman (2004): "Experimental Analysis of Neighborhood Effects on Youth," *Princeton IRS Working Paper 483*.
- [18] Krueger, A. B. (1999), "Experimental Estimates of Education Production Functions," *Quarterly Journal of Economics*, 114(2), 497 - 532.
- [19] Lazear, E. P. (2001), "Educational Production," *Quarterly Journal of Economics*, 116(3), 777 - 803.
- [20] Manning, W. G. and J. Mullahy (2001), "Estimating Log Models: To Transform Or Not To Transform?," *Journal of Health Economics*, 20(4), 461-494.
- [21] Mosteller, F. (1995), "The Tennessee Study of Class Size in the Early School Grades," *The Future of Children: Critical Issues for Children and Youths*, 5, 113 - 127.
- [22] Nye, B., L. V. Hedges and S. Konstantopoulos (1999), "The Long-Term Effects of Small Classes: A Five-Year Follow-Up of the Tennessee Class Size Experiment," *Educational Evaluation and Policy Analysis*, 21(2), 127 - 142.
- [23] Pete-Bain, H., C. M. Achilles, J. Boyd-Zaharas and B. McKenna (1992), "Class Size Does Make a Difference," *Phi Delta Kappan*, 253 - 256.
- [24] Romano, J., and M. Wolf (2005), "Exact and Approximate Stepdown Methods for Multiple Hypothesis Testing," *Journal of the American Statistical Association*, 100(469), 94 - 108
- [25] Shapson, S. M., E. N. Wright, G. Eason and J. Fitzgerald (1980), "An Experimental Study of the Effects of Class Size," *American Educational Research Journal*, 17, 141 - 152.

- [26] Streyffeler, L., .E. M. Altmaier, S. Kuperman and L. E. Patrick (2005), “Development of a Medical School Admissions Interview Phase 2: Predictive Validity of Cognitive and Non-Cognitive Attributes,” *Medical Education* 10(14) 1 - 5. *Online*.
- [27] Williams, V., L. Jones, and J. Tukey (1999), “Controlling Error in Multiple Comparisons, with Examples from State-to-State Differences in Educational Achievement,” *Journal of Educational and Behavioral Statistics*, 24(1), 42 – 69.
- [28] Woolfolk, Anita E. (1990), “*Educational Psychology*,” 4th ed. Boston, Allyn and Bacon.
- [29] Word, E., J. Johnston, H. Bain, D. B. Fulton, J. Boyd-Zaharias, N. M. Lintz, C. M. Achilles, J. Folger and C. Breda (1990), *Student/Teacher Achievement Ratio (STAR): Tennessee’s K–3 Class-Size Study*, Nashville, TN: Tennessee State Department of Education

Table 1: Summary Statistics of the Project STAR Kindergarten Sample

Variable	Number of Observations	Mean	Standard Deviation
Mathematics Test Score	5871	485.377	47.698
Reading Test Score	5849	434.179	36.762
Word Recognition Test Score	5789	436.725	31.706
Listening Skills Test Score	5837	537.4746	33.140
Motivation Skills Test Score	5038	25.64887	2.513
Self-Concept Skills Test Scores	5038	55.950	5.170
Teacher is African American	6282	0.165	0.371
Teacher has Master's Degree	6304	0.347	0.476
Years of Teaching Experience	6304	9.258	5.808
Student on Free Lunch Status	6301	0.484	0.500
Student is White	6322	0.669	0.470
Student is African American	6322	0.326	0.469
Student is Hispanic	6322	7.909*10E-4	0.028
Student is Asian	6322	2.201*10E-3	0.470
Student is Female	6326	0.486	0.500
Assigned to Small Class Treatment	6325	0.300	0.458
Class Size	6325	20.338	3.981
Inner City School	6325	0.226	0.418
Suburban School	6325	0.223	0.416
Rural School	6325	0.461	0.491
Urban School	6325	0.090	0.286

Table 2: Does the Impact of Class Size Vary by Student or Teacher Characteristics?  
 Estimation of Education Production Function with the Small Class Interactions

	Mathematics	Reading	Word Recognition	Listening Comprehension	Self Concept	Motivation
Kindergarten Small Class	12.095 (4.741)*	9.779 (3.090)**	9.749 (3.922)*	5.351 (3.045)	0.756 (0.584)	0.037 (0.227)
Female Student	7.816 (1.342)**	5.681 (0.958)**	5.640 (1.162)**	2.482 (0.893)**	-0.054 (0.172)	-0.072 (0.083)
Black Student	-16.258 (2.881)**	-7.544 (1.816)**	-7.066 (2.079)**	-17.221 (1.939)**	0.587 (0.378)	0.185 (0.197)
Student on Free Lunch	-20.123 (1.570)**	-14.918 (1.034)**	-16.022 (1.228)**	-15.994 (1.120)**	-0.745 (0.241)**	-0.082 (0.116)
Black Teacher	-3.122 (4.468)	-1.464 (3.215)	-1.711 (3.729)	1.282 (3.252)	0.601 (0.509)	0.048 (0.198)
Teacher has Masters Degree	-4.301 (2.631)	-0.483 (1.675)	0.066 (1.930)	-0.299 (1.555)	0.434 (0.311)	0.103 (0.136)
Years of Teaching Experience	0.584 (0.242)*	0.430 (0.145)**	0.414 (0.165)*	0.410 (0.191)*	0.046 (0.026)	0.006 (0.010)
Small Class *Female Student	-4.644 (2.391)	-1.057 (1.644)	-2.160 (1.951)	0.487 (1.598)	0.518 (0.336)	0.412 (0.149)**
Small Class *Black Student	-1.518 (3.905)	-0.416 (2.890)	0.458 (3.374)	-1.020 (2.860)	0.818 (0.496)	0.249 (0.236)
Small Class *Free Lunch Stu.	-0.000 (2.917)	0.616 (1.968)	0.062 (2.197)	2.270 (2.028)	0.480 (0.364)	0.071 (0.174)
Small Class *Black Teacher	11.999 (7.725)	5.216 (4.736)	2.941 (4.835)	6.850 (5.234)	-0.935 (0.664)	-0.258 (0.302)
Small Class *Master Teacher	5.139 (4.927)	-1.445 (3.000)	-0.057 (3.546)	1.744 (2.879)	0.202 (0.578)	-0.000 (0.244)
Small Class *Tch Experience	-0.467 (0.403)	-0.412 (0.257)	-0.326 (0.301)	-0.487 (0.259)	-0.077 (0.044)	-0.022 (0.021)
Constant	490.733 (2.906)**	438.305 (1.722)**	436.061 (2.073)**	544.697 (2.074)**	55.264 (0.306)**	25.526 (0.125)**
Observations	5809	5728	5790	5776	5000	5000
R-squared	0.27	0.27	0.23	0.26	0.05	0.03
Test of joint significance of all interactions	1.46 [0.191]	0.77 [0.597]	0.49 [0.816]	1.50 [0.177]	2.31 [0.034]*	1.89 [0.082]

Note: Standard errors corrected at the classroom level in parentheses. Regression equation includes information on school identifiers as well as interactions between the school indicators and student race being black. \* Significant at 5%; \*\* Significant at 1%.

Table 3: Estimation of Education Production Function with the Full Set of Interactions

	Mathematics	Reading	Word Recognition	Listening Comprehension	Self Concept	Motivation
Kindergarten Small Class	10.652 (4.543)*	8.984 (3.000)**	8.446 (3.858)*	4.161 (2.847)	0.704 (0.590)	0.025 (0.232)
Female Student	8.052 (2.936)**	6.893 (2.120)**	7.045 (2.632)**	1.825 (1.935)	-0.181 (0.304)	-0.249 (0.138)
Black Student	-31.248 (5.163)**	-14.478 (3.470)**	-15.355 (3.928)**	-27.034 (3.509)**	0.172 (0.635)	-0.069 (0.299)
Student on Free Lunch	-19.684 (3.187)**	-15.359 (2.003)**	-16.912 (2.284)**	-15.544 (2.282)**	-0.877 (0.425)*	-0.090 (0.202)
Black Teacher	-7.588 (7.689)	-4.739 (4.342)	-6.400 (4.680)	-5.358 (5.003)	0.254 (0.639)	-0.062 (0.344)
Teacher has Masters Degree	-13.429 (5.182)**	-4.844 (3.345)	-6.115 (4.010)	-1.924 (3.069)	0.432 (0.675)	0.103 (0.293)
Years of Teaching Experience	0.498 (0.324)	0.350 (0.183)	0.202 (0.219)	0.156 (0.192)	0.023 (0.038)	0.009 (0.016)
Small Class *Female Student	-4.113 (2.338)	-0.874 (1.640)	-1.985 (1.951)	0.756 (1.579)	0.537 (0.332)	0.430 (0.147)**
Small Class *Black Student	-0.245 (3.856)	0.334 (2.848)	1.785 (3.380)	0.183 (2.914)	0.905 (0.510)	0.333 (0.244)
Small Class *Free Lunch Stu.	0.299 (2.854)	0.815 (1.905)	0.337 (2.149)	2.456 (1.994)	0.498 (0.367)	0.054 (0.174)
Small Class *Black Teacher	10.988 (7.240)	4.764 (4.688)	2.043 (4.789)	6.573 (5.242)	-0.996 (0.662)	-0.310 (0.317)
Small Class *Master Teacher	5.431 (4.810)	-1.305 (2.921)	0.076 (3.485)	1.721 (2.791)	0.152 (0.582)	0.008 (0.244)
Small Class *Tch Experience	-0.333 (0.393)	-0.338 (0.246)	-0.200 (0.295)	-0.396 (0.238)	-0.071 (0.045)	-0.022 (0.021)
Female Black Student	3.671 (2.972)	0.220 (1.753)	0.037 (2.197)	3.620 (1.956)	-0.198 (0.374)	-0.051 (0.186)
Female student on Free Lunch	-4.725 (2.515)	-2.813 (1.596)	-3.109 (1.933)	-4.470 (1.731)*	0.135 (0.327)	0.023 (0.158)
Female Student *Black Teacher	5.912 (3.680)	3.071 (2.237)	2.862 (2.502)	2.204 (2.082)	0.575 (0.453)	0.514 (0.224)*
Female Student *Master Teacher	5.625 (2.371)*	1.200 (1.660)	1.315 (2.068)	1.060 (1.596)	0.013 (0.333)	-0.029 (0.155)
Female Student *Teach exp.	-0.239 (0.217)	-0.095 (0.134)	-0.094 (0.163)	0.087 (0.133)	0.004 (0.026)	0.012 (0.011)
Black Student on Free Lunch	9.415 (3.286)**	5.681 (2.196)*	6.382 (2.574)*	6.831 (2.170)**	-0.178 (0.428)	0.227 (0.206)

Black Student *Black Teacher	10.107 (5.704)	3.422 (3.808)	4.276 (4.384)	3.002 (4.247)	0.260 (0.562)	0.152 (0.293)
Black Student *Master Teacher	2.500 (4.268)	-0.032 (2.975)	1.841 (3.391)	3.208 (2.834)	0.375 (0.515)	0.377 (0.280)
Black Student * Teach Exp.	0.498 (0.381)	0.287 (0.275)	0.297 (0.307)	0.232 (0.272)	0.040 (0.042)	-0.007 (0.019)
Free lunch Stu. *Black Teacher	1.483 (4.634)	2.003 (2.762)	2.432 (3.151)	1.954 (3.310)	0.333 (0.529)	-0.048 (0.248)
Free Lunch Stu.*Master Tch	2.099 (2.771)	3.341 (1.824)	2.732 (2.146)	0.410 (1.796)	-0.009 (0.381)	0.061 (0.181)
Free lunch Stu.* Teach exp.	-0.151 (0.264)	-0.113 (0.169)	-0.054 (0.181)	-0.038 (0.179)	0.006 (0.033)	-0.007 (0.015)
Black Teacher *Master teacher	-6.351 (5.712)	-2.694 (3.482)	-3.626 (4.175)	-2.048 (4.077)	-1.129 (0.756)	-0.395 (0.333)
Black Teacher *Teach Exp.	-0.597 (0.592)	-0.143 (0.368)	0.012 (0.410)	0.362 (0.606)	-0.019 (0.064)	-0.012 (0.026)
Master Teacher *Teach Exp.	0.516 (0.431)	0.242 (0.287)	0.406 (0.326)	0.043 (0.247)	0.003 (0.051)	-0.005 (0.021)
Constant	493.858 (3.387)**	439.883 (2.018)**	438.831 (2.548)**	547.804 (2.286)**	55.575 (0.407)**	25.587 (0.174)**
Observations	5809	5728	5790	5776	5000	5000
R-squared	0.28	0.27	0.24	0.26	0.06	0.03
Test of joint significance of all interactions	2.10 [0.036]*	1.61 [0.045]*	1.36 [0.138]	1.94 [0.009]**	1.15 [0.293]	1.38 [0.128]
Test of joint significance of interaction between small class & student variables	1.06 [0.365]	0.17 [0.919]	0.48 [0.698]	0.59 [0.624]	3.20 [0.024]*	3.77 [0.011]*

Note: Standard errors corrected at the classroom level in parentheses. Regression equation includes information on school identifiers.

\* Significant at 5%; \*\* Significant at 1%

Table 4: Evaluating the Impacts of Small Classes Adjusting for Multiple Outcomes

Null Hypothesis being tested	Control variables include the full set of interactions	Number of Subjects being tested	Number of rejected P-value@.05 Independent	Number of rejected P-value @.10 Independent	Number of rejected P-value@.05 Account for FWER	Number of rejected P-value @.10 Account for FWER	Number of rejected P-value@.05 Account for FDR	Number of rejected P-value @.10 Account for FDR
Small Class =0	No	All 6	5	6	5	5	6	5
Small Class =0	No	3 Cognitive	3	3	3	3	3	3
Small Class =0	No	3 Non-Cognitive	2	2	2	2	2	2
Small Class =0	Yes	All 6	3	3	1	2	1	3
Small Class *Black Stu.=0	Yes	All 6	0	1	0	0	0	0
Small Class *Free L. Stu.=0	Yes	All 6	0	0	0	0	0	0
Small Class =0	Yes	3 Cognitive	3	3	3	3	3	3
Small Class *Black Stu.=0	Yes	3 Cognitive	0	0	0	0	0	0
Small Class *Free L. Stu.=0	Yes	3 Cognitive	0	0	0	0	0	0
Small Class =0	Yes	3 Non-Cognitive	0	0	0	0	0	0
Small Class *Black Stu.=0	Yes	3 Non-Cognitive	0	1	0	0	0	0
Small Class *Free L. Stu.=0	Yes	3 Non-Cognitive	0	0	0	0	0	0

Note: Each cell entry lists the number of hypotheses that reject the hypothesis in the first column at a specific level with a given procedure.

Appendix Table 1: Summary Statistics on the Percentiles of the Kindergarten Test Scores

Subject -> Percentile↓	Mathematics	Reading	Word Comprehension	Listening	Self- Concept	Motivation
5%	418	395	386	486	48	22
10%	429	402	396	498	49	23
20%	444	410	405	509	52	24
25%	454	414	405	516	53	24
30%	459	418	410	516	54	25
40%	468	425	418	528	55	25
50%	484	433	427	536	56	26
60%	494	439	435	545	58	26
70%	506	447	449	554	59	26
75%	513	453	458	560	59	27
80%	520	456	463	565	60	27
90%	547	474	480	578	62	28
95%	576	492	494	595	63	30
Minimum	288	315	315	397	24	12
Maximum	626	627	593	671	72	36

Appendix Table 2 : Does The Impact of Education Production Function Inputs Vary by Student Race or with Student Free Lunch Status?

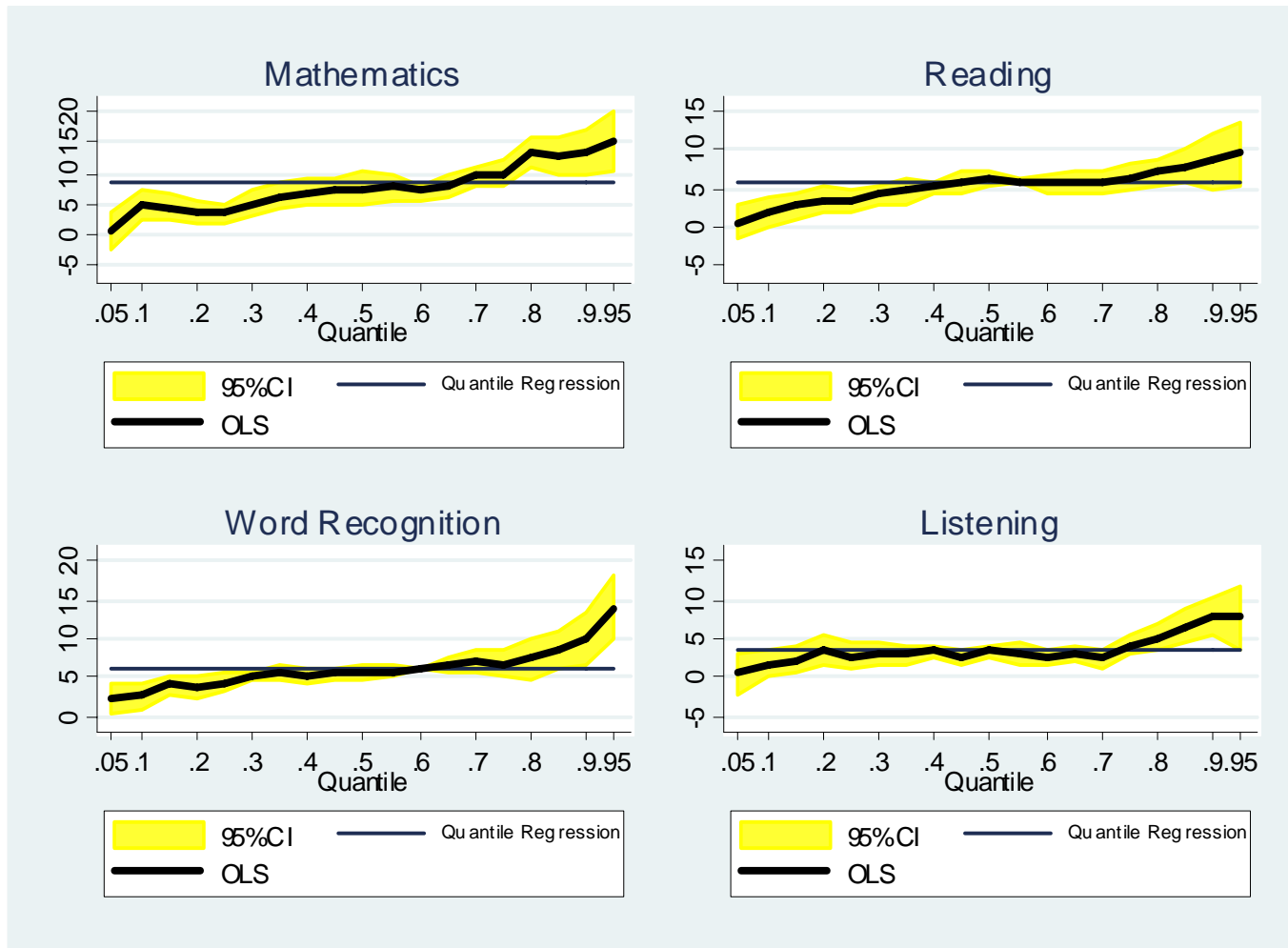
Estimation of Education Production Function with the Black Student Interactions						
	Mathematics	Reading	Word Recognition	Listening Comprehension	Self Concept	Motivation
Kindergarten Small Class	7.920 (2.247)**	5.170 (1.345)**	5.681 (1.603)**	2.634 (1.364)	0.447 (0.252)	0.030 (0.106)
Female Student	5.457 (1.396)**	5.354 (1.009)**	5.117 (1.247)**	1.978 (0.976)*	0.078 (0.179)	0.007 (0.077)
Black Student	-20.362 (6.424)**	1.176 (4.152)	-5.692 (4.536)	-7.112 (3.010)*	-6.387 (1.163)**	-2.151 (0.610)**
Student on Free Lunch	-21.744 (1.612)**	-15.754 (1.075)**	-17.284 (1.255)**	-16.648 (1.035)**	-0.560 (0.230)*	-0.101 (0.107)
Black Teacher	-8.633 (6.000)	-3.435 (3.696)	-5.424 (4.546)	0.636 (4.745)	-0.007 (0.419)	-0.320 (0.253)
Teacher has Masters Degree	-3.111 (2.272)	-1.147 (1.335)	-0.438 (1.602)	-0.441 (1.417)	0.528 (0.252)*	0.046 (0.110)
Years of Teaching Experience	0.291 (0.231)	0.208 (0.136)	0.165 (0.160)	0.061 (0.140)	-0.001 (0.025)	0.003 (0.012)
Small Class *Black Student	3.050 (4.170)	2.609 (3.011)	2.253 (3.065)	3.405 (2.565)	0.918 (0.443)*	0.290 (0.209)
Black Female Student	3.192 (2.536)	0.061 (1.553)	-0.254 (1.831)	1.913 (1.558)	0.066 (0.311)	0.138 (0.150)
Black Student on Free Lunch	8.868 (2.981)**	5.707 (1.985)**	6.267 (2.409)**	6.905 (2.103)**	-0.089 (0.427)	0.172 (0.203)
Black Student *Black Teacher	13.554 (6.973)	5.531 (4.838)	6.961 (5.577)	4.744 (5.564)	0.440 (0.647)	0.485 (0.314)
Black Student *Master Teacher	4.091 (4.500)	1.536 (3.152)	2.463 (3.379)	3.565 (2.888)	0.106 (0.489)	0.388 (0.253)
Black Student * Teach Exp.	0.266 (0.399)	0.217 (0.261)	0.308 (0.275)	0.379 (0.324)	0.045 (0.041)	-0.015 (0.018)
Constant	492.778 (2.963)**	437.504 (1.828)**	437.115 (2.148)**	543.367 (1.730)**	57.096 (0.361)**	26.063 (0.177)**
Observations	5809	5728	5790	5776	5000	5000
R-squared	0.28	0.28	0.24	0.27	0.06	0.04
Estimation of Education Production Function with Interactions on Free Lunch Status						
	Mathematics	Reading	Word Recognition	Listening Comprehension	Self Concept	Motivation
Kindergarten Small Class	8.192 (2.401)**	5.206 (1.553)**	6.009 (1.792)**	1.888 (1.471)	0.345 (0.252)	0.031 (0.109)
Female Student	7.031 (2.880)*	6.549 (2.000)**	6.296 (2.481)*	2.330 (1.896)	-0.009 (0.294)	-0.110 (0.131)

Black Student	-23.361 (3.832)**	-10.030 (2.906)**	-9.494 (3.268)**	-22.929 (2.639)**	1.041 (0.496)*	0.195 (0.221)
Student on Free Lunch	17.767 (8.301)*	-16.798 (4.895)**	-1.133 (4.673)	9.452 (3.810)*	0.329 (0.980)	0.584 (0.462)
Black Teacher	-3.002 (3.897)	-1.156 (2.882)	-2.124 (3.319)	1.801 (2.697)	0.107 (0.425)	-0.241 (0.177)
Teacher has Masters Degree	-4.780 (2.355)*	-1.273 (1.463)	-0.232 (1.764)	0.319 (1.458)	0.506 (0.313)	0.111 (0.134)
Years of Teaching Experience	0.563 (0.210)**	0.363 (0.126)**	0.360 (0.145)*	0.230 (0.150)	0.017 (0.026)	-0.006 (0.011)
Female Student on Free Lunch	1.115 (2.992)	1.542 (2.110)	0.651 (2.258)	3.394 (1.904)	0.785 (0.343)*	0.147 (0.160)
Black Student on Free Lunch	-3.233 (2.419)	-2.603 (1.550)	-3.086 (1.887)	-3.161 (1.610)	0.062 (0.308)	0.008 (0.144)
Small Class *Free Lunch Stu.	10.315 (4.634)*	4.333 (3.275)	4.728 (3.717)	7.221 (3.156)*	-0.569 (0.676)	-0.019 (0.306)
Free lunch Stu. *Black Teacher	8.402 (3.465)*	3.486 (2.165)	3.283 (2.387)	4.545 (2.060)*	0.404 (0.428)	0.434 (0.204)*
Free Lunch Stu*Master Tch	5.865 (2.434)*	1.349 (1.683)	1.441 (2.102)	1.039 (1.647)	0.011 (0.339)	-0.029 (0.161)
Free lunch Stu.* Teach exp.	-0.290 (0.221)	-0.110 (0.133)	-0.095 (0.162)	0.064 (0.136)	0.002 (0.025)	0.010 (0.011)
Constant	490.630 (2.832)**	438.321 (1.873)**	435.708 (2.296)**	546.095 (1.937)**	55.269 (0.339)**	25.493 (0.145)**
Observations	5809	5728	5790	5776	5000	5000
R-squared	0.28	0.28	0.25	0.27	0.06	0.04

Note: Standard errors corrected at the classroom level in parentheses. Regression equation includes information on school identifiers as well as interactions between the school indicators and student race (top panel) and between the school indicators and student being on free lunch (bottom panel).

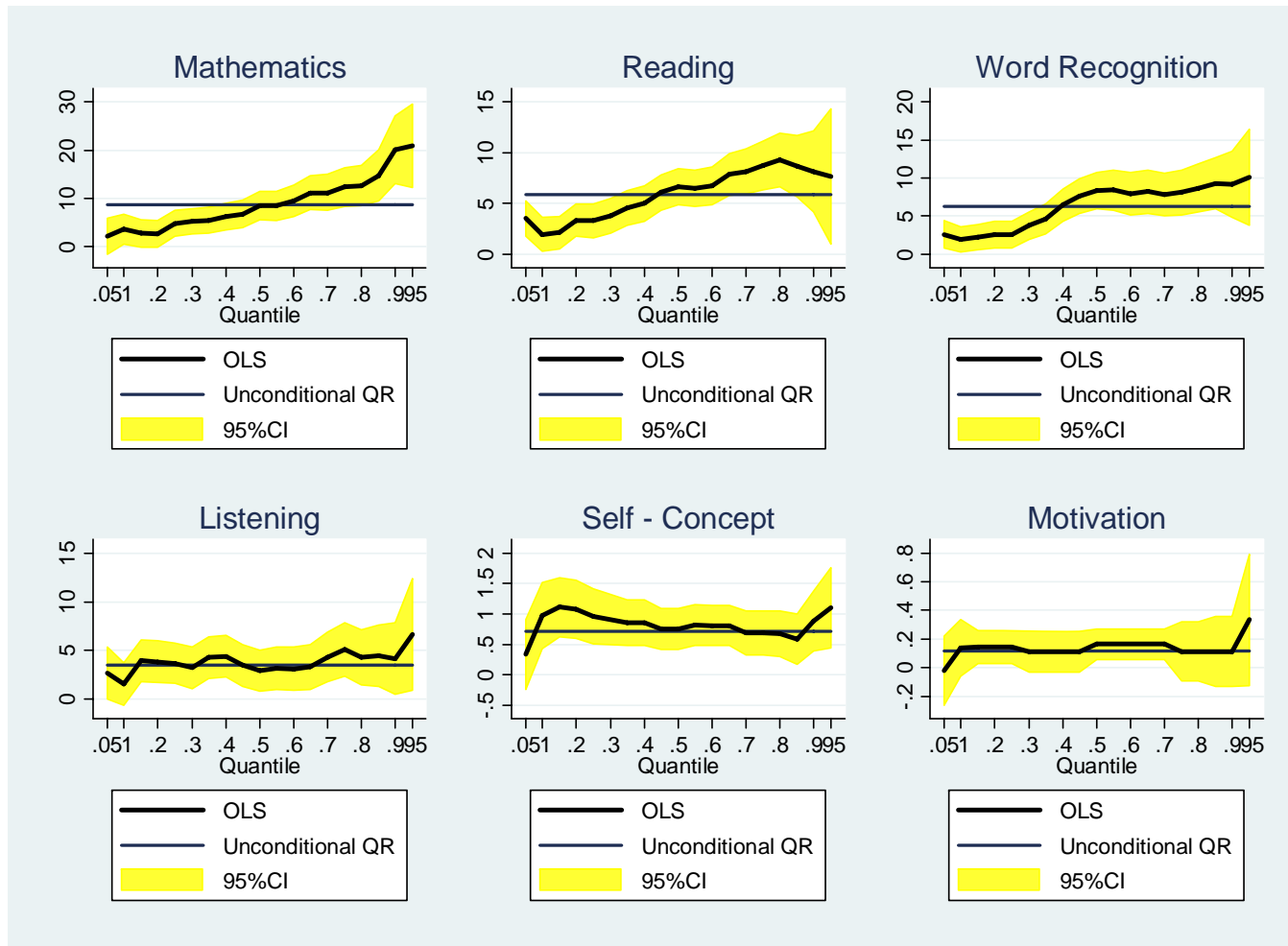
\* Significant at 5%; \*\* Significant at 1%

Figure 1: Quantile Regression and OLS Estimates of the Impact of Class Size on Kindergarten Achievement



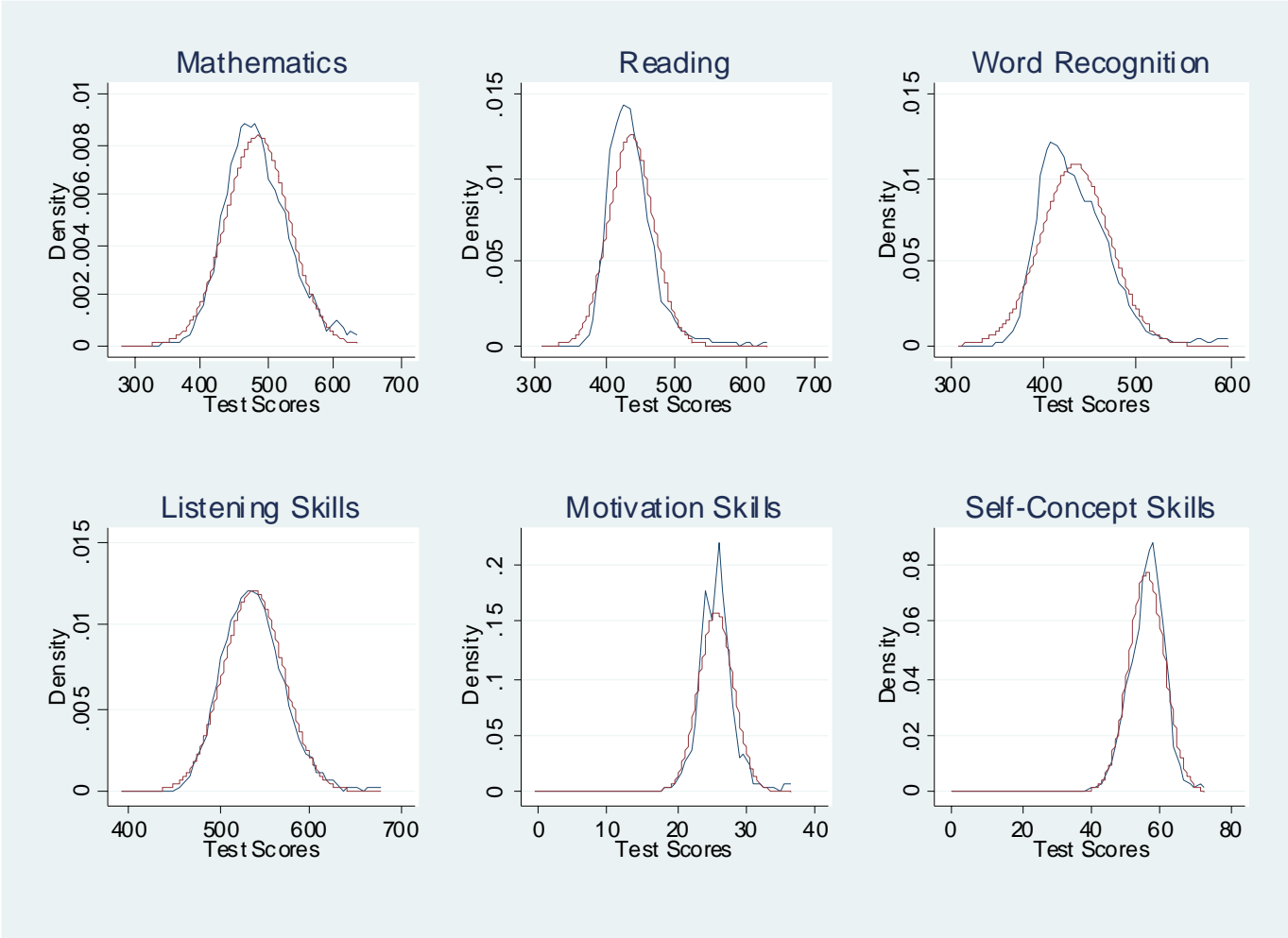
Note: The y-axis presents the estimated coefficient of the impact of small class on achievement. Specifications include student demographics (race, gender), free lunch status, class size, and teacher characteristics (race, gender, years of experience and highest education level completed).

Figure 2: Unconditional Quantile Regression and OLS Estimates of the Impact of Class Size on Kindergarten Achievement



Note: The y-axis presents the estimated coefficient of the impact of small class on achievement. Specifications include student demographics (race, gender), free lunch status, class size, and teacher characteristics (race, gender, years of experience and highest education level completed).

Appendix Figure 1: Kernel Density Estimates of Kindergarten Test Scores by Subject Area



Note: In each figure, the density function of the scaled test score data is presented with the blue line connected by dots. The red line represents the Normal density curve.