Canadian Labour Market and Skills Researcher Network

Working Paper No. 77

Experimental Estimates of the Impacts of Class Size on Test Scores: Robustness and Heterogeneity

Weili Ding
Queen's University

Steven F. Lehrer
Queen's University
NBER

June 2011

CLSRN is supported by Human Resources and Skills Development Canada (HRSDC) and the Social Sciences and Humanities Research Council of Canada (SSHRC).

All opinions are those of the authors and do not reflect the views of HRSDC or the SSHRC.

Experimental Estimates of the Impacts of Class Size on Test Scores: Robustness and Heterogeneity*

Weili Ding
Queen's University
dingw@queensu.ca

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

April 2011

Abstract

Proponents of class size reductions 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 economically disadvantaged students received the largest benefits from smaller classes. We extend this research in two directions. First, to address correlated outcomes from the same class size treatment, we account for the over-rejection of the Null hypotheses by using multiple inference procedures. Second, we conduct a more detailed examination of the heterogeneous impacts of class size reductions on measures of cognitive and noncognitive achievement using more flexible models. We find that students with higher test scores received greater benefits from class size reductions. Furthermore, we present evidence that the main effects of the small class treatment are robust to corrections for the multiple hypotheses being tested. However, these same corrections lead the differential impacts of smaller classes by race and free-lunch status to become statistically insignificant.

* We would like to thank Caroline Hoxby, Richard Murnane, an anonymous referee and seminar participants at the 2010 International Workshop on Applied Economics of Education, XIX Meeting of the Economics of Education Association, Simon Fraser University and the 2004 AEA annual meetings session on Education for the Disadvantaged for helpful comments and suggestions. We are grateful to Alan Krueger for generously providing a subset of the data used in the study. This paper is a revised version of the second chapter of Weili Ding's University of Pittsburgh 2002 thesis. 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 both 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 and new federal and state/provincial legislation and appropriations will promote further shrinkage of class sizes in North America. To support the launch of multi-billion dollar CSR initiatives, policymakers continue to draw from the reported experience of Project STAR, a randomized evaluation in the late 1980s on the impacts of CSR in Tennessee.

Two issues have been largely ignored in the discussion of the results from Project STAR. First, since students in Project STAR completed a battery of exams 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. 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, because without choice there is no influence. We adopt two multiple testing procedures that i) control for the probability of at least one rejection of a true Null hypothesis, and 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.[1]

Second, existing analyses of Project STAR data has focused almost exclusively on the estimation of average effects. If smaller classes do not benefit students equally, a more comprehensive understanding of which group of pupils received the largest benefits is needed. That is, the average effects reported in past studies do not shed light on the distribution of treatment effects. 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. From a policy perspective, estimating quantile impacts of inputs to an education production function (in addition to mean impacts) is likely important. This is because 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 the small class treatment heterogeneously impacts the achievement of students of different races, economic backgrounds and school characteristics, in order to account for a more comprehensive set of possible interactions between individual and school factors.

Understanding the heterogeneity in treatment across these dimensions is important as many researchers have hypothesized that the effects of CSR might vary across different types of students.[2] As such, proponents of class size reductions argue that there may be equity grounds to justify these policies, particularly if CSR initiatives are effective for students in the lower tail of the achievement distribution. Several researchers offer support for this claim. Lazear (2001) argues that smaller classes reduce opportunities for classroom disruption and if it is the case that classes with a greater proportion of lower achieving students are more disruptive, then lower achieving students might benefit the most from class size reductions. Ferguson (2003) contends that CSRs are more effective for minority students since these children may be more sensitive to teachers' perceptions and expectations. By receiving increased attention the students' work habits will improve and behavioral problems will decline. Grissmer (2002) likewise suggests that smaller classes are more efficient for economically disadvantaged students since they have had fewer prior investments into their human capital so that these investments reflect a larger contribution to their stock of human capital. Last, there is substantial evidence that teachers value their working conditions and like the idea of teaching smaller classes (e.g. Shapson (1980). Teachers may also change their teaching methods when faced with fewer students in the classroom.[3] Taken together, there are many potential pathways, in addition to influencing the kind of socializing experiences a student has in school, through which CSR may have heterogenous effects.

Using multiple inference procedures and allowing for flexible heterogeneity, we also investigate the impacts of CSR on non-cognitive skills, such as listening, motivation and self-concept. 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 have shown the importance of non-cognitive skills on a variety of education and labor market outcomes.[4] With the recent public availability of measures of non-cognitive performance from Project STAR, we have a chance to examine whether CSR has positive and statistically significant impacts on non-cognitive skills.[5]

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 that may bias the estimates, we only report analysis using data collected in the first year of the experiment. In Section 3, we discuss the statistical approaches that we employed and we report the empirical results. We find strong evidence that i) estimates of the mean impact of CSR for the full sample are robust when corrected for multiple correlated outcomes, ii) there are few additional benefits from CSR for minority or disadvantaged students, iii) students with higher test scores benefited the most from CSR. The multiple inference procedures that account for general correlations in student outcomes among 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 reveal that the few significantly (when the outcomes are treated as independent) differential impacts of CSR by race and free lunch status are likely due to chance. Some of the differences between our findings and earlier work are related to our more general treatment of the impacts of school factors and the different procedures by which researchers use test score measures as outcome variables. 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 entering kindergarten in 79 schools were randomly assigned to 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). Teachers were also randomly assigned to the classes they would teach, overcoming the student-teacher sorting bias (Rothstein (2010)) that plagues estimates of education production functions.

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.[6] By grade three, over 50% of the subjects who participated in kindergarten had left the STAR sample, and approximately 10% of the remaining subjects switch class type annually. Ding and Lehrer (2010a) 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 grade 1 students, students on free lunch status were significantly more likely to be assigned to regular (larger) classes.[7] To reduce concerns regarding these potential biases from non-random violations to the experimental protocol in subsequent years, in this paper our analysis focuses solely on data from the first year of the experiment.[8]

At the end of the kindergarten year the majority of the students completed six exams, measuring their performance in different dimensions. The students completed the Reading, Listening Comprehension, Mathematics and Word Recognition sections of the Stanford Achievement Test.[9] 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. This allows for 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.[10] 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 in that they measure the ability to comprehend spoken communication.[11] Finally, the students completed the Self-concept and Motivation Inventory Test presenting measures of two non-cognitive skills: self-concept and motivation. These measures are obtained from the child's response to 24 questions that are prefaced with the statement "What face would you wear if ...". The student selects a face by coloring in one of five different faces for each question. The overall test has moderate internal consistency (Davis and Johnston (1987)), and is scored from 24 to 120, with higher scores indicating more positive outcomes. The motivation 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 the teaching experience, education level, gender and race of the teacher, and 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 since some tests 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 analyses and specification tests have found no evidence of any systematic differences between small and regular classes in any student or teacher characteristics in kindergarten, suggesting that randomization was successful.

3 Empirical Results

3.1 Multiple Test Outcomes

Past research using Project STAR data has treated test scores in different subject areas as being independent from one another when attempting to estimate causal impacts.[12] However, the assumption of independence across dependent variables (test scores) may not appear plausible in an economics of education context, since the test scores in multiple domains (e.g. reading, writing and math) are likely highly correlated. Making adjustments for the use of multiple outcomes has a long history in psychology (Benjamini and Yekutiele (2001)) and biostatistics (Hochberg (1988)). These techniques have also been adopted in some studies within education (Williams et al. (1999)), as well as in studies in economics that examine multiple child outcomes (Kling and Liebman (2004) and Anderson (2008)). Accounting for the possibility that the multiple outcomes correlate within the study avoids the possibility of over rejecting the Null hypothesis that there are no treatment effects when using univariate statistical methods. Therefore, we need to adjust the p-value for the multiple outcomes and consider making corrections for both the familywise error rate (FWER) and false discovery rate (FDR). These p-value adjustments are based on the number of outcomes being considered and reduce the chance of making type I errors.

Formally, suppose that we want to test K hypotheses, $H_{1,}H_{2,\ldots.}H_k$ 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 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 where we also consider the three measures in the cognitive and non-cognitive domains separately. Although the FWER controls for the probability of making a Type I error,[13] we also consider accounting for the FDR, which controls the expected proportion of incorrectly rejected Null hypotheses (Type I errors) from a list of rejected hypotheses. It is a less conservative procedure with greater power than the FWER control, but at the cost of increasing the likelihood of obtaining type I errors. If all Null hypotheses are true, controlling for the FWER is equivalent to accounting for the FDR; however because increasingly more alternative hypotheses are true, controlling for the FDR can result in fewer Type II errors than controlling for the FWER.

To make corrections 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. To correct for the FDR, we use the two-step procedure developed in Benjamini, Krieger and Yekutieli (2006). This algorithm has been shown in simulation studies (e.g. Benjamini et al (2006)) to perform well and provide sharper control when p-values are positively correlated across tests, as is likely in our setting.

We begin by following earlier work and estimating the following contemporaneous achievement education production function for each component of the Stanford Achievement Test

$$A_{ij} = \beta' X_{ij} + \beta'_{CS} CS_{ij} + \varepsilon_{ij}, \tag{1}$$

where A_{ij} is the level of achievement for child i in school j, X_{ij} is a vector of school indicators and student and teacher characteristics, CS_{ij} is an indicator if student i attended a small class,[14] ε_{ij} captures both random unobserved factors as well as student invariant school specific effects. 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 β_{CS} with only kindergarten data. β_{CS} is often interpreted as an intent to treat estimate and since there were no issues with noncompliance in treatment assignment and that randomization appears to have been successful, this parameter is likely equivalent to the average treatment effect in the kindergarten year only.

Table 2 reevaluates the evidence of the mean effectiveness of CSR in kindergarten by adjusting inference of β_{CS} to account for multiple outcomes. The first three rows of Table 2 reexamines the effectiveness of CSR from OLS estimates of equation (1). The first two columns lists the Null hypotheses being tested, the specifications of the estimation equations, and the number of achievement outcomes being examined together. The third and fourth 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 that they report the number of rejections when accounting for the FDR using the Benjamini et al (2006) procedure. In general, the results indicates that the statistical significance of the mean impacts are robust to accounting for correlations between the different subject areas. However, we find that when making corrections for multiplicity with either the FWER or FDR leads to rejecting the CSR's positive impact on the motivation exam at the 10% level, which we fail to reject when we treated test scores across subject areas as independent. While this examination focused on constant treatment effects, we next examine whether the effects of CSR vary, either based on student characteristics or across the distribution of test scores.

3.2 Unconditional Quantile Regression

The estimated coefficient from an OLS school fixed-effects regression using the same empirical model and the same sample of students provides an estimates of the benefits from small class at the conditional mean, therefore, is potentially much less informative with regard to the relation between achievement and class size than the results for the various quantiles. We next allow the effects of class size reductions to vary for individuals at the different points of the unconditional test score distribution. Unlike estimates of the conditional mean from OLS estimates, the semiparametric estimator we consider will generate results that are robust to a monotone transformation of the dependent variable. Further, interpretation of estimates from an unconditional quantile regression strategy differ substantially from a (conditional) quantile regression strategy which additionally requires giving the residual a structural interpretation.[15] As such, we estimate the contribution of each explanatory variable to the unconditional quantiles of test scores, which permits us to answer questions 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.

The Firpo, Fortin and Lemieux (2009) regression method is applied to equation (1) and this essentially replaces the original outcome variable (A_{ij}) with a simple transformation known as the recentered influence function. The recentered influence function for the quantile of interest q_τ is formally defined as

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

where f_A is the marginal density function of A, and 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:

$$RIF(A;\widehat{q_\tau}) = \widehat{q_\tau} + \frac{\tau - I(A \leq \widehat{q_\tau})}{\widehat{f_A}(q_\tau)} \tag{3}$$

where $\widehat{q_\tau}$ is the τth sample quantile and $\widehat{f_A}$ 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 1 presents both OLS and unconditional quantile regression estimates of the impact of attending a small class on levels of kindergarten achievement by subject area.[16] The unconditional quantile regression estimates may provide more information about the impact of class- size reductions than the OLS estimates, since they can document how important attending a small class is at different achievement levels. The top row of Figure 1 presents estimates for mathematics, reading and word recognition. We observe that in these subjects that measure cognitive skills, students with higher test scores receive the most benefits from being assigned to a small class. For students in the lower quantiles of the test score distribution, the benefit is very small from an economic point of view. Students in the lowest quantiles in mathematics do not receive a statistically significant impact from attending a small class. Notice that the benefits from small class on both the reading and word recognition exams are statistically significantly different from zero at a conventional level throughout the achievement distribution. However, there is clear evidence of treatment effect heterogeneity, since the mean estimate obtained from an OLS regression is often not captured within the 95% confidence interval in these three subject areas. Simple tests of treatment effect homogeneity between quantiles are firmly rejected for these three subject areas.

Examining the conditional mean in isolation, therefore could lead to the wrong conclusion that the relation between class size and test scores does not differ sharply across these subject areas – a statement that is clearly refuted by unconditional quantile regression analysis; where we observe substantially more heterogeneity in mathematics than in the other subject areas. This increased heterogeneity may result from there being more heterogeneity in the knowledge and skills that the children bring with them to the classroom (potentially generated at the home) in mathematics relative to other domains. To summarize, we generally observe that the benefits from CSR increase over the achievement distribution on the three cognitive skills examinations.

Contrary to what is found for the cognitive skill tests. there is no evidence for treatment effect heterogeneity in the non-cognitive domains. This is illustrated in the bottom row of Figure 1 for the listening skills, self-concept and motivation tests. In each of these subject areas, test statistics fail to reject the Null of treatment effect homogeneity. The lack of treatment effect heterogeneity may exist due to the limited variation in the underlying scores or even the nature of the tests. In particular, there is substantial mass at few test scores for both the self-concept and motivation exam. The lack of heterogeneity in performance on the listening exam is not unique to the achievement distribution, but is also found to be the only subject in which the treatment effect did not vary with the proportion within school receiving treatment in Ding and Lehrer (2010b). Interestingly, being assigned to a small class yields positive benefits only for 7 of the 19 quantiles on the motivation exam.

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. However, it is difficult to ex-ante identify students who may score poorly at the end of the year and evidence presented in Ding and Lehrer (2010a) shows that those who are among the lowest scoring in the mathematics exam on entry in their classroom out gain their classmates in the subsequent grades performance. Thus, we take a closer look at how the relationship between class size and achievement varies across subgroups that are easy to identify ex-ante in the next section.

3.3 Small Classes for the Disadvantaged

Class size reductions have played a large role in recent policy debates in the search for mechanisms to reduce the achievement gap between disadvantaged children and other children. It is often reported that CSRs offer greater benefits for both minority and inner city children. For example, past research using 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 experienced 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 and those who were not (Word et al (1990)). By reporting larger gains for disadvantaged students, the political appeal of CSR policies increased. However, much of this research has employed statistical models that allow for limited forms of heterogeneity and are based on specifications of the education production function that either ignore school specific unobserved heterogeneity or treat this term as a random effect. To create a set of benchmark results, we first re-examine whether students on free lunch status and minority children gain more in small classes on average. We begin by interacting the individual student and teacher characteristics with class size and estimated the following equation

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

where ε_{ij} continues to include components that capture both random unobserved factors as well as student invariant school specific effects.

Estimates of equation (4) for all six subjects are presented in Table 3. We observe that the interaction between attending a small class and being eligible for free lunch is statistically insignificant in all six subject areas. Similarly, African Americans students did not perform significantly differently in smaller classes compared to regular sized classes. The bottom row of Table 3 contains the results from an F test of whether $\beta'_{XCS} = 0$. The test statistics reject the Null hypothesis, indicating that the interaction terms are jointly insignificant in all subject areas with the exception of self-concept.

Equation (4) allows for a limited amount of interactions. Past work with Project STAR (e. g. Dee (2004)) has presented evidence of significant complementarities between student and teacher inputs. As such, we next consider the most flexible method to evaluate whether there was heterogeneity in the impact of small class treatment across groups, by estimating a fully saturated model that contains all possible interactions between student, class type and teacher covariates. This specification imposes the fewest restrictions on an underlying model of education production function and only assumes that the unobserved inputs are additively separable from the observed inputs to the production process.

A subset of the estimated coefficients from the fully saturated model are presented in Table 4. We find that although the effects of small class attendance remain highly significant in math, reading and word recognition, but the only interaction term between small class and another input that has a significantly positive impact on academic performance is small class interacted with female student for the motivation exam. In specifications that consist of the full set of interactions there are substantially large negative impacts for being a minority student or student on free lunch. Further, the interaction between African American student and free lunch status is highly significant and positively related to achievement on all four Stanford Achievement tests.

In table 4, 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.[17] Interestingly, the inclusion of this large set of regressors appears to only explain a limited amount of the variations in self-concept and motivation scores. This reinforces why we were unable to find evidence of treatment effect heterogeneity in these subjects in Figure 1.

While the mean effects of small class obtained from equation (1) were robust to accounting for multiple testing, the above discussion of Tables 3 and 4 treated each test score outcome as independent to be comparable to previous STAR studies. In Table 5, we summarize the results of applying either the correction for the FWER or FDR used in Table 2 to assess the significance of the interactions of the treatment with various student demographics in either Table 4. Notice we do not find that there remains a statistically significant differential impacts of CSR after we apply the corrections to the statistical inference procedure. To a large degree this should not come as a surprise since none of the interactions between small class and student demographic characteristics were significant at the 5% level when we assumed the outcomes are independent, with the exception of the interaction between female student and small class on the motivation assessment. These findings cast doubt that there are truly heterogeneous mean impacts from CSR across groups defined by race or free lunch status in kindergarten. In terms of mean effects in Table 4, once we account for multiple correlated outcomes, 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 remains significant when we account for the FWER, whereas accounting for FDR yields equivalent results to assuming independence.

Even ignoring issues related to multiple inference, there are clear discrepancies between our results and those from other papers on Project STAR data in regards to whether smaller classes benefit the disadvantaged more. These differences arise from two major features in the analysis. First, prior work conducted analyses separately on samples defined by class types and then compared the magnitude of the estimated coefficients on the free lunch variable rather than pooling the sample and including interaction terms. In some papers (e.g. Finn and Achilles (1990) researchers were exploiting variation across schools and did not account for student invariant school heterogeneity. Since randomization was done within and not between schools, these comparisons ignore the experimental design that provides exogenous variation to identify causal impacts and are necessary to achieve an unbiased estimate of β_{CS}. Further F tests indicate that school effects should be accounted for in all the specifications of the education production function that we consider.

Moreover, it is important to note that only 34% of the African American and Hispanic students in the full kindergarten sample attend schools that also have white or Asian students. In fact, there are 15 schools that consist only of minority students and 15 schools for which there was not a single minority student in the kindergarten sample. These schools do not have any within school variation that can be used to identify racial gaps or heterogeneity by race. Thus, results using raw differences from specifications estimated using samples defined by class types could lead the results to be confounded by factors that vary across schools and may end up having a different set of schools contribute to the treatment effect. The performance differences of students in small and regular classes in these schools is clearly different as documented in Appendix Table 2. This table first documents how the mean performance of minority and Caucasian students varies in both small and regular classroom across schools based on the racial distribution of the kindergarten cohort. For example, in mathematics for minority students there is a large 11 point difference in student performance when only examining the schools that did not exhibit racial heterogeneity. In contrast, in the schools that had both minority and Caucasian students, minority students did not perform in a significantly different manner between small and regular classes. Differences between schools in columns 2 and 3 with columns 6 and 7 need to be accounted for as there are likely substantial differences in neighborhood and community inputs to the production process. In addition, school differences are needed to be accounted for since the randomization is done within schools. Further, there are gains in efficiency of the estimates by using the full sample of students and including interactions with school fixed effects in the specification of the education production function.

Second, the method in which student performance is measured varies substantially across samples. In our study, we use scaled scores for outcomes from the Stanford Achievement test since they are developmental and are considered by the test publisher to be the natural unit of measurement for a norm referenced tests.25 Alternative measures to estimate student performance with STAR data represent monotonic transformations of the scaled scores or raw scores. These measures include percentile scores (e.g. Krueger (1999)), standard scores (e.g. Schanzenbach (2007)) and grade equivalent scores (e.g. Finn et al (1999)). 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 they 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. In addition, it is much easier to interpret marginal effects and translate results to policymakers with scaled scores because 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. For completeness, the empirical distribution of kindergarten test scores in all six subject areas is presented in Appendix Figure 1. A move from the median to the 60th percentile is equivalent to moving 10 scaled points or 0.018σ whereas moving from the 80th to the 90th percentile involves 27 scaled points or 0.294σ. The transformation from one measure to another changes the variation in outcome scores to be explained by the regressors. The relationship between standard scores, percentile scores and scaled scores also varies from test to test. We replicated all of the analysis in Tables 3 and 4 with both standard scores and percentile scores, and there were several differences in the significance of the findings.[18] While the methods to specify dependent variables in labor economics and health economics have been an active area of study (e.g. Blackburn (2007) and Manning and Mullahy (2001)) where dependent variables have i) nonnegative outcomes and ii) skewed outcome distributions, such issue has been understudied in the economics of education literature, which we believe warrants further investigation since the empirical distribution of test scores are often skewed and differ from a Normal distribution.

We also replicated the analysis that generated Table 3 with a subsample of students from inner-city schools, and compared the estimated coefficients obtained to those obtained running the same specification with students from other school districts. We did not find any significant differences in the estimated magnitude 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.[19] The discrepancy between our work and earlier studies comes largely from the inclusion of school fixed effects in equation (4). We believe these are necessary to achieve an unbiased estimate of β'_{CS} since randomization was done within and not across schools.[20]

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 complements existing studies of this large and influential experiment, and provides three new findings. First, we find that estimates of the mean impact of CSR for the full sample are robust to statistical corrections for multiple inference. Second, these same corrections reject any evidence for additional benefits from CSR for minority or disadvantaged students in Kindergarten. Third, we find substantial heterogeneity in the impact of attending a small class on the distribution of test scores in all cognitive subject areas. The results indicate that students with higher test scores benefit the most from small classes in these subject areas. Taken together, we find mixed evidence on the effectiveness of CSR in kindergarten. This is because CSR leads to significant improvement in cognitive achievement measures, it appears to provide few benefits in the development of non-cognitive skills.

It may well be that CSR is more effective for some groups of students defined by alternative criteria on specific subjects, in which case policy might be more effective if it targets specific student sub-population rather than mandating across-the-board reductions. Understanding why CSRs were only effective in some subjects but not others is clearly a direction for future research.[21] Since 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. Further, there is a growing body of research that documents substantial within school variation in teacher quality. Several researchers, including Word et al. (1990) and Hanushek (1999), have suggested that the pattern of findings in the Project STAR study is also consistent with the existence of substantial within school differences in teacher quality. While teacher quality is common to all students in a classroom, evidence presented in Dee (2004) suggests that teachers in Project STAR were more effective with students whose race matches their own.[22] While these explanations could explain differences across schools in the mean effects of CSR, extremely large (and arguably implausible) effects of teacher quality on student achievement would be required to explain the heterogeneity in effects exhibited in Figure 1, particularly if both teachers and students were independently randomly assigned to classrooms.

We postulate that the larger effects from CSR in the higher quantiles of student achievement presented in Figure 1 may suggest that family background is very important and that interventions within schools may only reinforce at home preparation for a small fraction of the population. However, a limitation of the STAR data is the limited number of home inputs that were collected. In particular, we do not have any direct knowledge of how parents change their investments in their children as a response to their child being assigned to a small class or the extent and pattern of heterogeneity in the parental input decisions. In conclusion, we suggest that the substantial heterogeneity in the impacts from class size reduction witnessed in kindergarten should promote further investigation, using both qualitative and quantitative data to improve our understanding of the pathways through which class size contributes to the production of education outcomes.

Notes

[1] In contrast, earlier research has either examined each of these outcomes independently or combined a subset of the outcome measures collected into a single index using arbitrary weights.

[2] More generally, CSR policies are expensive. In times of shrinking government budgets, it is worth knowing whether CSRs should be implemented universally or in a targeted fashion.

[3] Rice (1999) and Word et al. (1990) among others, report that teachers do change the methods they use when assigned to a class with fewer students. In contrast, Shapson (1980) and Bandiera et al. (2010), among others, do not find evidence of changing teacher behavior.

[4] This is of policy relevance since research has shown that non-cognitive skills influence individual performance on cognitive tests (Borghans et al. (2008)), the likelihood of school dropout (Heckman and Rubinstein (2001)) and the amount of schooling obtained (Heckman, Stixrud and Urzua (2006)).

[5] The Word et al. (1990) report does not find a significant impact of CSR on some non-cognitive measures.

[6] 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 schools left immediately after kindergarten.

[7] It should also be noted that attendance of kindergarten was not mandatory in Tennessee. Students who entered school in grade 1 may differ in unobservables from those who started in kindergarten.

[8] The general pattern of our results holds in subsequent years where we corrected for subsequent selection on observables using inverse probability weighting. Specifically, the samples are reweighted by either series logit estimates of the probability of remaining in the sample, or the probability of having written the exam in the previous academic year. These analyses impose additional behavioral assumptions and are available upon request.

[9] 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.

[10] As we discuss in Section 3.3, the selection of scores is of critical importance in interpreting the results. Much of the previous work has employed transformations of the scaled scores as outcome variables, which has major effects upon their results.

[11] 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 the behavior of college admission committees who consider listening to be a non-cognitive skill and reading to be a cognitive skill. (Streyffeler et al. (2005)).

[12] One alternative is to collapse outcome variables into a single measure or score. However, the correct way of combing and weighting different outcome variables is not obvious and measurement error in the dependent variable may increase. That being said, Cunha et al. (2010) have made progress on this issue of combining outcomes by demonstrating that one can define a scale for output is not invariant to monotonic transformation. Specifically, they anchor test scores to the adult earnings of the child, which has a well-defined cardinal scale. However, this information is not available for the full Project STAR sample. More generally, often one would like to evaluate an early childhood intervention after it has been completed, and not until years later when the participants enter the labor market.

[13] The FWER maintains the overall probability of making a Type I error at a fixed α (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.

[14] Following Finn et al. (2001) and Krueger (1999), our control group consists of regular classes with and without teacher aides, because these studies (among others) report that the presence of a teacher aide did not significantly impact student test scores. Our independent analyses confirm these results.

[15] That being said, in many papers results from a quantile regression analysis are often incorrectly interpreted as being from an unconditional quantile regression analysis. See Ding and Lehrer (2005) for an analysis of this dataset with conditional quantile regression estimators.

[16] The estimates of the impacts of the other explanatory variables on the quantiles of the achievement distribution are available from the authors upon request.

[17] We also considered less flexible specifications that only include interactions between the inputs and either the race or free lunch variable. The results are presented in Appendix Table 1. With the exception of self-concept, the effects of class size interacted with either being black or being economically disadvantaged are statistically insignificant.

[18] In particular, with both standard and percentile scores one would conclude that small classes benefited performance on the self-concept exam at the 5% level and the listening test at the 10% level. In total there are 7 to 9 differences in the significance of various interactions for these alternative rescalings. As the SAT-9 codebooks are no longer published, we could not convert the scores to grade equivalents, which yields the students' standing in relation to the norm group at the time of testing. However, the interpretation of these scores is confusing and they are known to have low accuracy for students with very high or low scores. Furthermore, these scores are inappropriate to use for computing group statistics or in determining individual gains.

[19] The results are available from the authors by request.

[20] Additionally, we replicated the analysis presented in Figure 1 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. 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. The patterns are nearly identical as both students on free lunch and African American students in higher quantiles benefit more from smaller classes than students in the lower quantiles. For example, African American students in the highest test score quantile receive over 5 times the benefits on mathematics from small class relative to students in the smallest quantile. Students on free lunch in the highest quantile receive over 4 times the benefit relative to those in the lowest quantile in mathematics. The number of quantiles in which small class is not statistically different from zero on achievement is greater for these subsamples.

[21] There is limited examination in the economics of education literature on the mechanism

of 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 disciplining students (Lazear (2001)). 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. They do find teachers in large classes spent more time on discipline and less time on individualized attention. 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, relative to the teachers with larger class sizes. They conclude that teachers need to be trained in instructional strategies for various sized classes.

[22] This result is sensitive to the specification of the education production function. Interactions between student and teacher are not statistically significant at conventional levels in the fully saturated models presented in Table 4. In the top panel of Appendix Table 1, where a subset of interactions are included, the teacher and student race interaction is significant at the 10% level if the test score outcomes are treated as independent.

References

[1] American Education Research Association (2003), “Class Size: Counting Students Can Count Research Points,” *Research Points*, 1(2), 1 - 4.

[2] Anderson, M. (2008) “Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects,” *Journal of the American Statistical Association,* 484, 1481 - 1495.

[3] Bandiera, O., V. Larcinese, I. Rasul (2010), “Heterogeneous Class Size Effects: New Evidence from a Panel of University Students,” *The Economic Journal*, 120(549), 1365–1398.

[4] 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.

[5] Benjamini, Y., A. Krieger, and D. Yekutieli (2006) “Adaptive Linear Step-up Procedures that Control the False Discovery Rate,” *Biometrika,* 93(3), 491 – 507.

[6] Betts, J. R. and 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.

[7] Blackburn, M. L. (2007), “Estimating Wage Differentials Without Logarithms,” *Labour Economics*, 14(1), 73 - 98.

[8] Borghans, L., B. ter Weel, and B. Weinberg (2008), “Interpersonal Styles and Labor Market Outcomes,” *Journal of Human Resources,* 43(4), 815 - 858.

[9] Cunha, F., J. J. Heckman and S. M. Schennach, (2010), “Estimating the Technology of Cognitive and Noncognitive Skill Formation,” *Econometrica*, 78(3), 883 - 931.

[10] Davis T. M. and J. M. Johnston (1987), “On The Stability and Internal Consistency of the Self-concept and Motivation Inventory: Preschool/Kindergarten Form,” *Psychological Reports,* 61, 871–874.

[11] Dee, T. S. (2004), “Teachers, Race, and Student Achievement in a Randomized Experiment,” *Review of Economics and Statistics,* 86(1), 195–210.

[12] Ding W. and S. F. Lehrer (2005), “Class Size and Student Achievement: Experimental Estimates of Who Benefits and Who Loses from Reductions,” *Queen’s University Department of Economics Working Papers number 1046.*

[13] Ding W. and S. F. Lehrer (2010a), "Estimating Treatment Effects from Contaminated Multi-Period Education Experiments: The Dynamic Impacts of Class Size Reductions," *Review of Economics and Statistics,* 92(1), 31-42.

[14] Ding W. and S. F. Lehrer (2010b), "Estimating Context-Independent Treatment Effects in Education Experiments," *mimeo,* Queen's University.

[15] Ferguson, R. (2003), "Teachers' Perceptions and Expectations and The Black-White Score Gap" *Urban Education*, 38(4), 460-507.

[16] Finn, J. D. and C. M. Achilles (1990), "Answers about Questions about Class Size: A Statewide Experiment," *American Educational Research Journal*, 27, 557 - 577.

[17] Finn, J. D. and C. M. Achilles (1999), "Tennessee's Class Size Study: Findings, Implications, Misconceptions," *Educational Evaluation and Policy Analysis*, 21(2), 97-109.

[18] Finn, J. D., S. B. Gerber, C. M. Achilles and J. Boyd-Zaharias (2001), "The Enduring Effects of Small Classes," *Teachers College Record*, 103(2), 145-183.

[19] Finn J. D. (2002), "Class Size Reduction in Grades K-3," in A. Molnar (ed.), *School Reform Proposals: The Research Evidence*, Greenwich, CT: Information Age Publishing.

[20] Firpo, S., N. M. Fortin and T. Lemieux (2009), "Unconditional Quantile Regressions," *Econometrica,* 3(5), 953 - 973.

[21] Grissmer, D. (2002), "Cost-Effectiveness and Cost-Benefit Analysis: The Effect of Targeting Interventions," in H. M. Levin and P. J. McEwan (eds.), *Cost-Effectiveness and Educational Policy,* AEFA Yearbook: Eye on Education, Larchmont, NJ.

[22] Hanushek, E. A. (1999), "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.

[23] 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.

[24] 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.

[25] Hochberg, Y. (1988), "A Sharper Bonferroni Procedure For Multiple Tests of Significance," *Biometrika,* 75(4), 800 – 802.

[26] Holland, B. and M. D. Copenhaver (1987), "An Improved Sequentially Rejective Bonferroni Test Procedure," *Biometrics*, 43, 417 - 442.

[27] Kling, J. R., and J. B. Liebman (2004): "Experimental Analysis of Neighborhood Effects on Youth," *Princeton IRS Working Paper 483.*

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

[29] Lazear, E. P. (2001), "Educational Production," *Quarterly Journal of Economics,* 116(3), 777 - 803.

[30] Manning, W. G. and J. Mullahy (2001), "Estimating Log Models: To Transform Or Not To Transform?," *Journal of Health Economics*, 20(4), 461-494.

[31] 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.

[32] Pate-Bain, H., C. M. Achilles, J. Boyd-Zaharas and B. McKenna (1992), "Class Size Does Make a Difference," *Phi Delta Kappan,* 253 - 256.

[33] Rice, J. (1999), "The Impact of Class Size on Instructional Strategies and the Use of Time in High School Mathematics and Science Courses." *Educational Evaluation and Policy Analysis*, 21(2), 215 - 229.

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

[35] Schanzenbach D. W., (2007), "What Have Researchers Learned from Project STAR?" *Brookings Papers on Education Policy*, 2006/07, 205-228.

[36] 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.

[37] 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.*

[38] 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.

[39] 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 is Female	6325	1.000	0.000
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: Evaluating the Impacts of Small Classes Adjusting for Multiple Outcomes

Null Hypothesis being tested	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	All 6	5	6	5	5	5	6
Small Class =0	3 Cognitive	3	3	3	3	3	3
Small Class =0	3 Non-Cognitive	2	2	2	2	2	2

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. FWER and FDR respectively denotes correcting the statistical inference for the familywise error rare and false discovery rate.

Table 3: 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 4: 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)
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)
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)
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 and interactions with individual inputs and teacher characteristics as well as between teacher characteristics. * Significant at 5%; ** Significant at 1%.

Table 5: Evaluating the Impacts of Small Classes Estimation From Specifications That Control for the Full Set of Interactions Reported in Table 4 where the Statistical Inference Procedure Now Adjusts for Multiple Correlated Outcomes

Null Hypothesis being tested	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	All 6	3	3	1	2	1	3
Small Class *Black Stu.=0	All 6	0	1	0	0	0	0
Small Class *Free L. Stu.=0	All 6	0	0	0	0	0	0
Small Class =0	3 Cognitive	3	3	3	3	3	3
Small Class *Black Stu.=0	3 Cognitive	0	1	0	0	0	0
Small Class *Free L. Stu.=0	3 Cognitive	0	0	0	0	0	0
Small Class =0	3 Non-Cognitive	0	0	0	0	0	0
Small Class *Black Stu.=0	3 Non-Cognitive	0	1	0	0	0	0
Small Class *Free L. Stu.=0	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. FWER and FDR respectively denotes correcting the statistical inference for the familywise error rare and false discovery rate.

Appendix Table 1 : 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

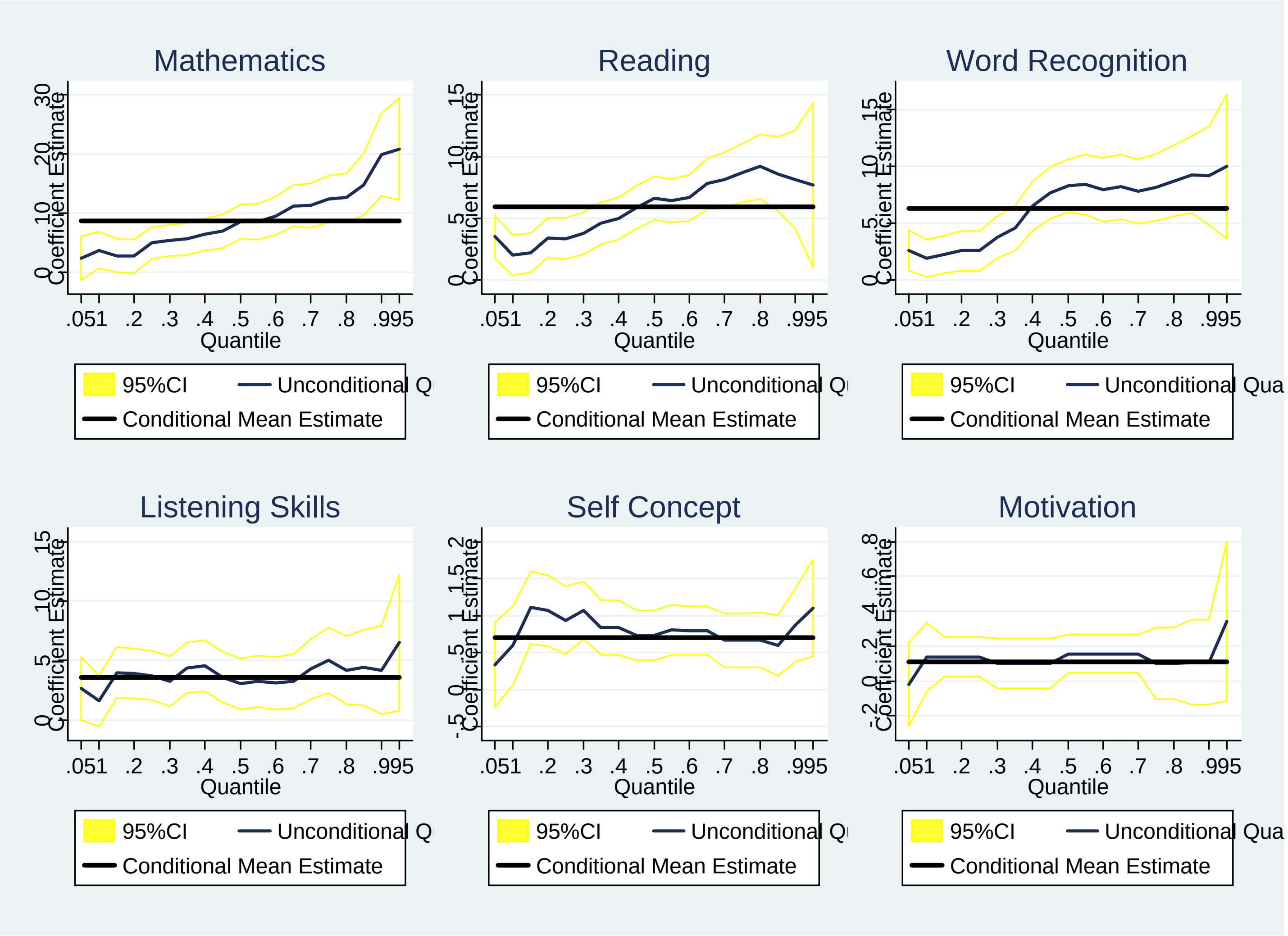
Note: Standard errors corrected at the classroom level in parentheses. Regression equation includes information on school identifiers as well as interactions between the school indicators and student race (top panel) and between the school indicators and student being on free lunch (bottom panel).
* Significant at 5%; ** Significant at 1%

Appendix Table 2: Summary Information on Student Performance by Class Type and Student Race in Schools with and without Student Heterogeneity in Race.

	Project STAR schools that consist of only African American and Hispanic Students in Kindergarten 15 Schools		Project STAR schools that do not contain any African American or Hispanic Student in Kindergarten 10 Schools		Schools with Mixed Student Body in Kindergarten 54 Schools	
African American and Hispanic Students						
Subject Area	Small Classes	Regular Classes	Small Classes	Regular Classes	Small Classes	Regular Classes
Mathematics	479.978 (52.155)	468.756 (48.864)			478.900 (47.600)	475.009 (45.504)
Reading	433.003 (27.917)	427.769 (29.074)			435.112 (31.055)	425.387 (27.208)
Word Recognition	428.555 (30.492)	424.115 (34.858)			432.545 (38.444)	422.053 (31.130)
Listening Comprehension	522.239 (30.853)	519.439 (31.371)			527.515 (27.820)	520.996 (28.232)
Self Concept	57.148 (5.323)	55.897 (5.640)			57.058 (4.510)	56.449 (5.372)
Motivation	25.963 (2.882)	25.721 (2.812)			26.006 (2.093)	25.788 (2.791)
Caucasian and Asian Students						
Subject Area	Small Classes	Regular Classes	Small Classes	Regular Classes	Small Classes	Regular Classes
Mathematics			496.693 (48.541)	488.157 (44.993)	493.395 (46.280)	492.922 (44.808)
Reading			445.210 (33.785)	439.238 (31.892)	436.065 (31.109)	437.928 (30.965)
Word Recognition			443.764 (38.935)	437.765 (36.850)	432.636 (35.046)	432.911 (36.227)
Listening Comprehension			548.025 (32.642)	544.140 (30.998)	542.788 (28.653)	546.664 (32.844)
Self Concept			55.993 (5.345)	55.536 (4.888)	55.965 (5.659)	56.072 (4.508)
Motivation			25.578 (2.462)	25.589 (2.361)	25.529 (2.660)	25.454 (2.122)

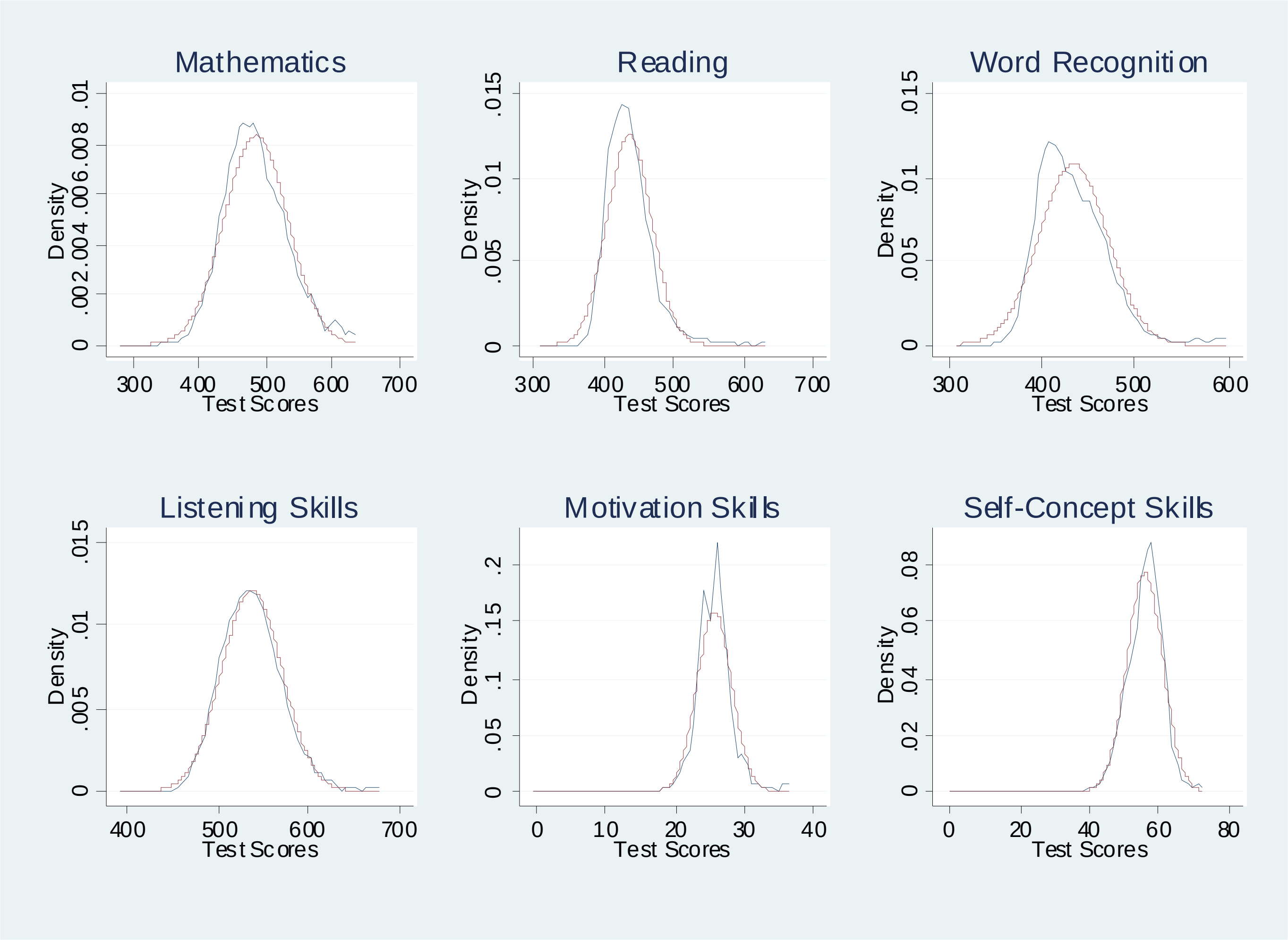
Note: Each cell contains the unconditional mean and standard deviation.

Figure 1: 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 the same covariates as used in Table 2.

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.