Queen's Economics Department Working Paper No. 1046

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

Weili Ding
Queen's University

Steven Lehrer
Queen's University

Department of Economics
Queen's University
94 University Avenue
Kingston, Ontario, Canada
K7L 3N6

9-2005

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

Weili Ding
Queen's University

Steven F. Lehrer
Queen's University

September 19, 2005

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. To explore and truly understand the heterogeneous impacts of class size on student achievement requires more flexible estimation approaches. We consider several semi and nonparametric strategies and find strong evidence that i) higher ability students gain the most from class size reductions while many low ability students do not benefit from these reductions, ii) there are no significant benefits in reducing class size from 22 to 15 students in any subject area, iii) no additional benefits from class size reductions for minority or disadvantaged students, iv) significant heterogeneity in the effectiveness of class size reductions across schools and in parental and school behavioural responses.

* We are grateful to Alan Krueger for generously providing a subset of the data used in the study. We are grateful to Caroline Minter Hoxby and Richard Murnane for comments and suggestions which have helped to improve this paper. 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 and new federal and state/provincial legislation and appropriations will promote further shrinkage of class sizes in North America. Policymakers continuously draw from the perceived 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.[1] What has been largely ignored in the discussion of the results from Project STAR is that the prescription of smaller classes does not benefit different students in an equal manner and there remains substantial divide in debates regarding the optimal number of students per classroom as well as understanding which group of pupils receive the largest benefits.[2]

This paper takes a closer look at the heterogenous effects of CSR in kindergarten. Kindergarten was the first year of the program and the available evidence indicates that there were extremely few violations of the randomization protocol. With few violations estimates from kindergarten are statistically reliable and present the cleanest possible evidence on the impacts of class size with Project STAR data.[3] This paper provides substantial new evidence on the heterogeneous effects of

[1] 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."

[2] For example, Barnett et al. (2004) survey the literature and state that minority students and students attending inner-city school benefited most. Finn (2002) states that the benefits of small classes were two to three times greater for minority students but does not draw a distinction between inner city schools and suburban /rural schools.

[3] Violations to the randomization protocol were severe after Kindergarten and estimates of the experimental impact in Grade 1 are distorted from causal parameters such as an intent to treat. Further, assignment of refreshment

reduced class size, an active and arguably the most highly politicized area of debate in education.[4] Specifically, we consider more flexible estimation approaches and examine whether there is indeed a tipping point at which class size gains begin to accrue.[5] In addition, we examine whether there are heterogeneous impacts of small class by race, economic background and school characteristics.[6]

Finally, since randomization was done within schools we explore program heterogeneity by undertaking a closer examination of the effectiveness of class size reductions in each school. The idea that program heterogeneity across locations is likely to be an important source of differences in effectiveness of a full program is well established in labor economics.[7] In the context of CSR, it

samples in grade 1 does not appear to be random within schools. Since Kindergarten was not mandatory in Tennessee differences on numerous dimensions unobserved to the researcher further contaminate inference using samples from the later years. See Ding and Lehrer (2004) for a strategy to obtain estimates of causal parameters when there are many violations to the randomization protocol.

[4]This debate has extended to discussions of whether class size effects exist in studies that employ non-experimental data. The well known survey by Hanushek (1986) finds no evidence to support reducing class sizes in the U.S. Further, Hanushek (1999a) finds no evidence in international comparisons, where "extraordinarily large" differences in class sizes occur without commensurate differences in student performance. Krueger (2003) reanalyzes Hanushek's data and reaches a different conclusion.

[5]In their review of evidence from Project STAR, The Manitoba Teachers' Society (2001) state "the "tipping point" does seem to be between 19 and 20 students". In a broader survey, the American Education Research Association (2003) conclude that "the number of students in a class should range from 13 to 17", suggesting a tipping point following 18.

[6]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 twice the small class benefit (Finn and Achilles, 1990), 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). Since past research has reported larger gains for disadvantaged students increasing the political appeal of CSR policies. Yet, much of this research has employed statistical models that allow for limited forms of heterogeneity.

[7]For example, Grogger and Karoly (2005) note, the recent trend in the design of Welfare-to-Work programs at

is highly probable that not only did the impact of CSR vary across schools in the sample which contain different student populations but also that the method in which instruction was undertaken in small classes varied. Understanding the relative extent of components such as teaching strategies, teaching experience, feedback to parents, teacher quality as well as student ability and characteristics in influencing achievement gains in small classes is of crucial policy importance. After all, if policymakers outside of Tennessee seek to use Project STAR as a guideline they must account for the fact that it was conducted in a different time-period and location with substantially different populations.[8]

Our results yield new evidence on the heterogeneous impacts of CSR. We find strong evidence that i) higher ability students gain the most from CSR while many low ability students do not benefit from these reductions, ii) there are no significant benefits in reducing class size from 22 to 15 students in any subject area, iii) no additional benefits from CSR for minority or disadvantaged students, iv) significant heterogeneity in the effectiveness of CSR across schools and in parental and school behavioral responses. Finally, we find that the positive effects of CSR on achievement outcomes in the STAR kindergarten sample are being driven by slightly over 25% of the participating schools. Since there does not appear to be a systematic relationship between kindergarten class size and academic achievement understanding why it works in some schools but not in others is essential. Our analysis of the publicly available STAR data does 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.

---

the state and local levels is towards greater program heterogeneity. In different locations, policy makers implement a variety of different program components such as amount of job search assistance or training or case management strategies to reach their goals of increased employment.

[8]Methods described in Hotz, Imbens and Mortimer (2005) can be used to sort out the sources of differences in the variation of average treatment effects across schools to shed light on how effective a CSR programs will be in a new location.

2 Project STAR

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.[9]

At the end of the kindergarten year the majority of the students completed the Reading, Mathematics and Word Recognition sections of the Stanford Achievement test.[10] 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.[11] Scaled scores are arbitrary and vary according to the test given, but within the same

[9] The majority of research ignores these problems or treats non-compliance as random and ignorable. Instrumental variable procedures have also been employed to analyze data from Project STAR. However as discussed in Ding and Lehrer (2004), in the presence of selective attrition even an IV estimate is biased from a causal effect.

[10] 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. Generally, scores are reported in terms of percentiles, grade equivalents or standard scores, all of which compare or rate one student's performance in relation to a norm group.

[11] The raw score is simply the number of correct responses a student gives to test items. Total percent scores divide the raw score by the total number of items on the test. Raw scores are converted to scaled scores by use of a psychometric technique called a Rasch model process. The Rasch model, developed by George Rasch in 1960, is a one parameter logistic model that examines how performance relates to knowledge as measured by items on a test.

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. We present Kernel density estimates of the scaled scores for Reading, Mathematics and Word Recognition in kindergarten of the STAR program in Figure 1.[12] Notice that the although graphs for the scores tend to be unimodal and somewhat bell shaped, they are clearly non-normal as indicated by the deviations from the normal distributions superimposed on Figure 1.

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

Intuitively the idea is that the probability that an exam taker of a certain ability level answers a question correctly is based solely on the difficulty level of the item. The estimated coefficient is on the ability continuum where the probability of a correct response is 50%.

[12]The selection of scores is of critical importance in interpreting the results and previous work employed transformations of the scaled scores as outcome variables, which has drastic effect upon their results. Krueger (1999) uses percentile scores that provide the percentage of students in the regular class sample (this is his norming group) whose scores were at or lower than a given score. Percentile scores are useful to compare a student's performance in relation to other students. However, differences in percentile units cannot be used to compute gains since scores are not constant across the entire scale. After all, the long tails shown in Figure 1 clarify that increasing from 50 to 51 percent on the percentile score is not equivalent to increasing performance from 95 to 96 percent. Further, standard estimation techniques will place a disproportionate amount of weight on scores near the mean where observations are clustered since this transformation reduces the weight place on observations in the tails of Figure 1. Note Krueger (1999) also averages scores across subject areas and between the Stanford Achievement test and the Basic Skills First test which criterion referenced (as opposed to norm referenced) test. Similarly, Finn and Achilles (1999) use grade equivalent scores which measure performance in terms of the grade level at which the typical pupil makes this raw score. Grade equivalents are known to have low accuracy for students with very high or low scores and are inappropriate for computing group statistics or in determining individual gains (Woolfolk (1990)). Standard scores employed by Mosteller (1995) are appropriate only if the distribution of the scaled scores across subject areas comes from distributions that only differ in the first moment. But, as demonstrated in Figure 1, there is non zero skewness and kurtosis and the variance differs across subject areas, making the use of scores measured in deviations from a mean a raw proxy of the real distributions.

statistics on the Project STAR kindergarten sample are provided in Table 1. Nearly half of the sample is on free lunch status. There are very 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 very 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 Methodology

As a benchmark, consider estimation of the following contemporaneous achievement education production function by subject area

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

where A_{ij} is the level of achievement for child i in school j, X_{ij} is a vector of student and teacher characteristics, CS_{ij} is the actual number of students in the class, v_j is a school fixed effect and ε_{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, class size at kindergarten is uncorrelated with unobserved factors such as the impact of pre-kindergarten inputs, family and community background variables, etc., permitting estimation of treatment effects with only one period of data. This formulation treats class size as a linear regressor which restricts the effect of a reduction from 26 to 25 students to be equal to a reduction of 18 to 17 students. Our analysis begins by relaxing this assumption.

3.1 Results I: Nonparametric Form on Class Size

Actual class size in Project STAR varied from 12 to 28. We begin by considering two strategies that relax the assumption that class size enters equation (1) linearly. First we estimate the following partial linear model for each subject area separately

$$A_{ij} = \beta' X_{ij} + h(CS_{ij}) + v_j + \varepsilon_{ij} \tag{2}$$

where $h(*)$ is a nonparametric function to be estimated.[13] Since there does not exist any theory within education or economics literature that specifies how class size impacts student achievement we estimate the shape of a unspecified function rather than impose assumptions on the shape of the relationship. Our estimates will indicate the shape of the relationship in STAR data and can be used to detect the presence of any tipping points. Control variables included in X_{ij} correspond to those employed by Krueger (1999) and include indicators for each student's race, gender and free lunch status, as well as each teacher's race, experience and education.

Second, since class size is a discrete variable we flexibly model the relationship between class size and achievement through the use of class size dummy variables

$$A_{ij} = \beta' X_{ij} + \sum_{k=12}^{28} \beta_{CS}^{\prime k} I(CS_{ij} = k) + v_j + \varepsilon_{ij} \tag{3}$$

where $\beta_{CS}^{\prime k}$ is a vector of dummy variable coefficients. Due to collinearity we omit the indicator for 22 students. This class size is selected as the reference point since the majority of cost benefit analyses are based on reducing class size from 22 students to 15 students.

Estimates of equation (3) are presented in Table 2. Notice that there is no systematic evidence that alternative class sizes lead to either gains or reductions in achievements in all subject areas relative to 22 students. The coefficient estimates fluctuate from positive to negative, and the majority are statistically insignificant. Class size of 26 students perform significantly lower than 22

[13] See Robinson (1988) for a discussion of root N consistent estimation of this equation.

students and class sizes of 14 or 16 students perform significantly better. However, there are no differences between 15 students and 22 students in any subject area.

Nonparametric estimates of the effects of class size on student achievement from equation (2) are presented for each subject area in Figure 2. The figures reinforces that there is no systematic relationship between class size and achievement in any of the subject areas. Further, there is no evidence of a tipping point at which CSR are effective. The ranges over which class size changes lead to gains in achievement are either within regular classes (28 to 26 students) or within small classes (14 to 15 and 17 to 18 students).All of these reductions are a move within the same class size type and do not span the class size region between regular and small classes within schools.

3.2 Results II: Differential Treatment Effects by Unobserved Ability

In our work we allow the effect of class size to vary according to student ability and run quantile regression estimates of the equation (1). Quantile regression provides a more flexible approach to characterizing the effects of observed covariates such as class size on different percentiles of the conditional achievement distribution. 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 are presented in Figure 3. 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 at least twice as much from a one person reduction in class size 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. (coeff.=-6.129, s.e. 2.714)

3.3 Results III: School Differential Treatment Effects

To address program heterogeneity across schools we conducted a simple comparison using one sided t-tests for each school between small classes and regular classes. We find that the gains from reduced class size are driven by 20 (out of 79) schools where in all three tests (reading, math and word) the small class average scores are statistically greater than the regular class averages (at the 10% level), which are the schools without doubt find small classes effective. We label these schools "effective schools". 39 schools have either zero or negative small class effects on all three tests, which without doubt have not experienced benefit from small classes, if not lending proof for effective regular classes. We label these schools "ineffective schools". The remaining 19 schools see small classes outperform regular classes in some tests and no difference in other tests (only in one school do small classes perform better in one test, worse in another test and no difference in a third test), which suggest that small classes are sometimes effective but the evidence is not conclusive. OLS regressions of equation (1) excluding the 20 effective schools never find positive or significant small class effects in any subject area. Similarly, estimating (1) excluding the 20 effective schools where class size is replaced by class type finds that small classes are not significantly related to achievement in any subject area.

When we compare the 20 effective schools with the 39 ineffective schools, we find from one sided t-tests that in each subject area small class students score significantly higher and regular class students significantly lower in effective schools than their counterparts in ineffective schools. The appearance of effectiveness in these schools is not only because they generated stronger student

performance in their small classes than the ineffective schools but also because they failed to generate as strong a student performance in their regular classes as in ineffective schools. Moreover, there is little evidence that the effective schools are doing better than the ineffective schools: the average test scores of the effective schools are not statistically different from those of the ineffective schools in each subject area. These findings lend some support to the concern that some effective schools might have allocated different resources to their small and regular classes as there is no obvious rationale why their regular class students should perform worse than those in the ineffective schools when their small class students could do better than those in the ineffective schools. Note that the effectiveness of small class is exaggerated if better resources are assigned to small classes.

Further, experimental treatments also differ in their attractiveness so that the number and characteristics of subjects who remain in subsequent periods may differ following kindergarten. We examine whether individuals who subsequently leave the STAR experiment are systematically different from those who remain in terms of initial behavioral relationships.[14] We estimate the following contemporaneous specification of an education production function in kindergarten by subject area

$$A_{ij} = \beta' X_{ij} + \beta'_{CS} CS_{ij} + \beta'_L L_{ij} Z_{ij} + v_j + \varepsilon_{ij} \tag{4}$$

where L_{ij} is an indicator for attrition and $Z_{ij} = [X_{ij}, CS_{ij}]$. The vector β'_L allows for both a simple intercept shift and differences in slope coefficients for future attritors.

The results are presented in Table 3. Wald tests indicate that the β'_L coefficient vector is significantly different for attritors and non-attritors in all subject areas. The attrition indicator is significantly negatively related to test score performance in all three subject areas, indicating that the levels of performance for subsequent attritors is significantly lower in kindergarten. In all subject areas, the joint effect of attrition on all student characteristics and class type is significantly different

[14]Note that the STAR study not only witnessed attrition in students but also in schools. Six schools left the experiment prior to the end of grade 3. Each of these schools was ineffective and five of them left immediately after kindergarten.

from zero. The interaction between subsequent attrition and free lunch status is significantly and negatively related to mathematics achievement indicating that students on free lunch status that left scored significantly lower than free lunch students who remained in the sample in that subject only. In all subject areas, the interaction between female and subsequent attrition is statistically significant indicating that female attritors out performed female non-attritors in kindergarten, but the magnitude is small. Finally, in both mathematics and word recognition attritors received half the gain of reduced class sizes, suggesting that non-attritors obtained the largest gains in kindergarten which if unaccounted for may bias future estimates of the class size effect upwards.

Not only were there interesting attrition patterns in the full sample but substantial heterogeneity across schools in the type of non-random transition by Kindergarten class assignment. Students in effective schools were less likely to move from small classes to regular classes after kindergarten (coeff.= -0.071, s. e. 0.036) and more likely to transit from regular classes to small classes than in the ineffective schools (coeff.= 0.322, s.e. 0.030).[15] This indicates that parents in effective schools quickly learnt of these performance differences and responded by requesting their children be moved to smaller classes. The students who moved out of the small classes were termed incompatible children by the schools, were taken out of the small class despite their intentions and on average scored between 8.5 and 13.2 scaled score points less than their small class counterparts. Surprisingly attrition was significantly more common in the effective schools versus the ineffective schools (coeff.= 0.190, s. e. 0.023). Yet while the ineffective schools witnessed no significant differences in attrition rates based on initial class type assignment (coeff.= -0.006, s. e. 0.019), students in effective schools initially assigned to small classes were significantly less likely to leave the sample (coeff.= -0.078, s. e. 0.016). To summarize, it appears that the heterogenous impacts of CSR were apparent to parents who either pulled their children out of school if they were performing poorly or asked for

[15] There results and the remainder reported in the paragraph are from OLS regressions that include all the regressors in equation (1) replacing class size with class type and also include school effects. The standard errors rae corrected at the school level.

their child to be moved to a smaller class predominantly in schools where CSR was statistically effective in Kindergarten.

3.4 Results IV: 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. These reported positive results have substantial political appeal particularly the claims that CSR is more beneficial for minority and inner city children. These claims are not consistent with our findings from quantile regression results which present evidence that ability and class size are complements; unless disadvantaged children possess high unobserved abilities. To further investigate whether disadvantaged and minority children gain more in small classes we interacted the individual student and teacher characteristics with class size and estimated the following equation

$$A_{ij} = \beta' X_{ij} + \beta'_{CS} CS_{ij} + \beta'_{XCS} CS_{ij} X_{ij} + v_j + \varepsilon_{ij} \tag{5}$$

The results are presented in the Table 4. The first three columns contain results using actual class size and its interactions and the last three columns use an indicator for being in a small class versus regular or regular with aide class in place of class size. Notice with either measure of class size the interaction terms are jointly insignificant at conventional levels in all subject areas. Further, the interaction between small class and free lunch status was individually insignificant in all subject areas. Similarly white and Asian students did not perform significantly different in smaller classes. Interacting inputs and characteristics with either race or free lunch status one will find that only white or Asian students on free lunch perform significantly worse than black and Hispanic students on free lunch. If free lunch captures family background, it seems that likely some characteristics of a family that initiate its white or Asian child to free lunch status have a more perverse impact on its child's academic achievement than the family background of black and Hispanic students. An alternative explanation could be that families of white and Asian students on free lunch are

on average in wore shape (more segregated from other families if the same race, more stigmatized) than the families with black and Hispanic children on free lunch in the STAR data. Note that approximately only 34% of the African American and Hispanic students in the kindergarten sample attend schools that also contain white or Asian students.

We also replicated the analysis in Tables 4 interacting the individuals regressors with an indicator variable for inner city schools.[16] This does not yield any significant shifts for the class size variable, thus there is no evidence that the impacts of smaller classes are larger in inner city schools. These discrepancies between our results and earlier work is first due to the fact that prior work ran multiple regressions separately on small classes and regular classes, comparing the magnitude of the estimated coefficients as opposed to pooling the sample and including interaction terms. Pooling is preferred since by using the full sample gains in efficiency are obtained. Further, while the interpretation of the interaction terms from a regression using the pooled sample as intercept or slope shifts is straightforward, the subsample approach generally does not restrict unobserved school factors to be fixed across subsamples which distorts inference.

Second, past analysis has focused heavily on comparing the magnitude of the black-white test score gap between small and regular classes. This is highly misleading as these gaps are measured with aggregate data 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. Thus, these raw differences between class types may be confounded by factors that vary across schools. In contrast, our approach directly tested whether black or economically disadvantaged students receive any additional achievement benefits from being in smaller classes and correctly exploited the experimental variation from randomization.

We next consider whether there is any evidence that minority or disadvantaged students benefit more or less than their classmates for each school input. We estimate the following equation that

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

allows the impacts of each input to the production process to vary either by race or free lunch status

$$A_{ij} = \beta' X_{ij} + \beta'_{CS} CS_{ij} + \beta'_{XCS} CS_{ij} X_{ij} + v_j + \varepsilon_{ij} \tag{6}$$

where $R_{ij} = 1$ indicates the group of students who we allow a differential response. The results are presented in Table 5. In columns 1-3, $R_{ij} = 1$ indicates whether a student is African American and $R_{ij} = 1$ indicates whether a student is on free lunch in columns 4 to 6. Notice that in all columns, the effects of class size interacted with either being black or being economically disadvantaged are statistically insignificant.[17] Further, none of the teacher characteristics have a significantly different impact for either group. We conclude that in Kindergarten there are no additional gains in achievement in any subject area from attending small classes for either disadvantaged or minority children.

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 by demonstrating that higher ability students gain the most from CSR while many low ability students do not benefit from these reductions. Second we find no significant benefits in reducing class size from 22 to 15 students. There is no clear evidence for a tipping point at which benefits to small classes accrue in all subject areas. While 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 than others in which case

[17] We also investigated specifications that included interaction terms between race and free lunch status separately with the school identifiers in addition to the other regressors in the estimating equation. These specifications allow the impacts of school effects v_j to vary across groups within schools. Our results are robust to the inclusion of these terms and are available from the authors by request.

policy would be most effective targeting specific populations and not mandating across the board reductions. Finally, we find significant heterogeneity in the effectiveness of CSR across schools and in parental and school behavioral responses.

Understanding why some schools were able but other schools were not able to translate smaller classes into gains in student achievement is essential for public policy.[18] Since treatments were not standardized across schools, uncovering the source as well as the extent of heterogeneity in treatment implementation is of critical importance for education policy. While there are many potential candidates to explain the heterogeneous returns it would be possible to investigate some of these candidates if the complete data set collected by Project STAR researchers were made available to the general research community. In conclusion, we suggest that the substantial heterogeneity in the impacts of class size should promote further investigation rather than the approval of additional policies that mandate class size reductions.

[18]There has been very little 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. Some of the strongest available empirical evidence is provided by Betts and Shkolnik (1999) who 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, experimental evidence from the education literature on teacher behavior across class sizes (16, 23, 30 or 37 students) is found in Shapson et al., (1980). Shapson and his colleagues 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 with the study because they had less work to do in terms of evaluating students' work, than did the teachers in class sizes of 30 and 37. They conclude that teachers need to be trained in instructional strategies for various size classes. Thus, the available evidence suggests that teaching practices did not vary with class size as hypothesized.

References

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

[2] Betts, Julian R. and and Jamie 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.

[3] Ding, Weili and Steven F. Lehrer (2004), "Estimating Dynamic Treatment Effects from Project STAR," *mimeo*, Queen's University.

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

[5] Grogger, Jeffrey and Lynn A. Karoly (2005), Welfare Reform: Effects of a Decade of Change, Cambridge, MA: Harvard University Press, 2005.

[6] Hanushek, Eric A. (1986), "The Economics of Schooling: Production and Efficiency in Public Schools," *Journal of Economic Literature*, 49, 1141 - 1177.

[7] Hanushek, Eric A. (1995), "Interpreting Recent Research on Schooling in Developing Countries," *World Bank Research Observer*, 10, 227 - 246.

[8] Hanushek, Eric A. (1999a), "The Evidence on Class Size," in Susan E. Mayer and Paul Peterson (ed.), *Earning and Learning: How Schools Matter,* Washington, D.C.: Brookings Institution, 131 - 168.

[9] Hanushek, Eric 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.

[10] Hotz, Joseph V., Guido Imbens, and Julie Mortimer (2005), "Predicting the Efficacy of Future Training Programs Using Past Experiences," *Journal of Econometrics*, 124, 241-270.

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

[12] Krueger, Alan B. (2003), "Economic Considerations and Class Size," *Economic Journal,* 113, F34 - F63.

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

[14] Manitoba Teachers' Society (2001) *Class Size: Less is More: The Manitoba Teachers' Society; Written Submission to Class Size and Composition,* Winnipeg MB: ProActive Information Services Inc.

[15] Mosteller, Frederick (1995), "The Tennessee Study of Class Size in the Early School Grades," *The Future of Children: Critical Issues for Children and Youths,* V, 113 - 127.

[16] Murnane, Richard J. (1975), "*Impact of School Resources on the Learning of Inner City Children*," Cambridge, Ballinger.

[17] Nye, Barbara, Larry V. Hedges and Spyros 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.

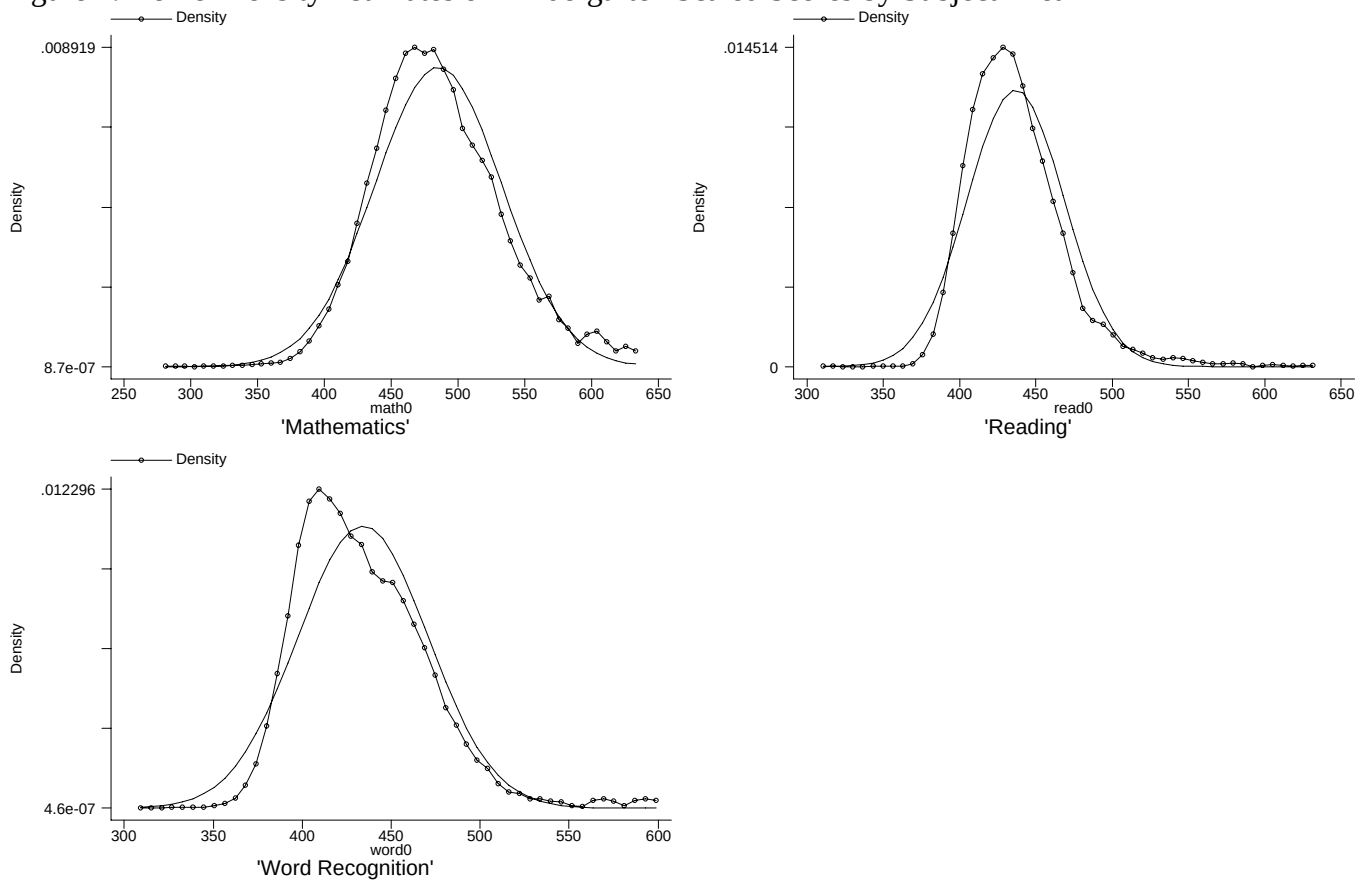
[18] Pate-Bain, Helen, Charles M. Achilles, Jayne Boyd-Zaharas and Bernard McKenna (1992), "Class Size Does Make a Difference," *Phi Delta Kappan,* 253 - 256.

[19] Shapson, Stan M., Edgar N. Wright, Gary Eason and John Fitzgerald (1980), "An Experimental Study of the Effects of Class Size," *American Educational Research Journal*, 17, 141-152.

[20] Woolfolk, Anita E. (1990), "*Educational Psychology*," 4th ed. Boston, Allyn and Bacon.

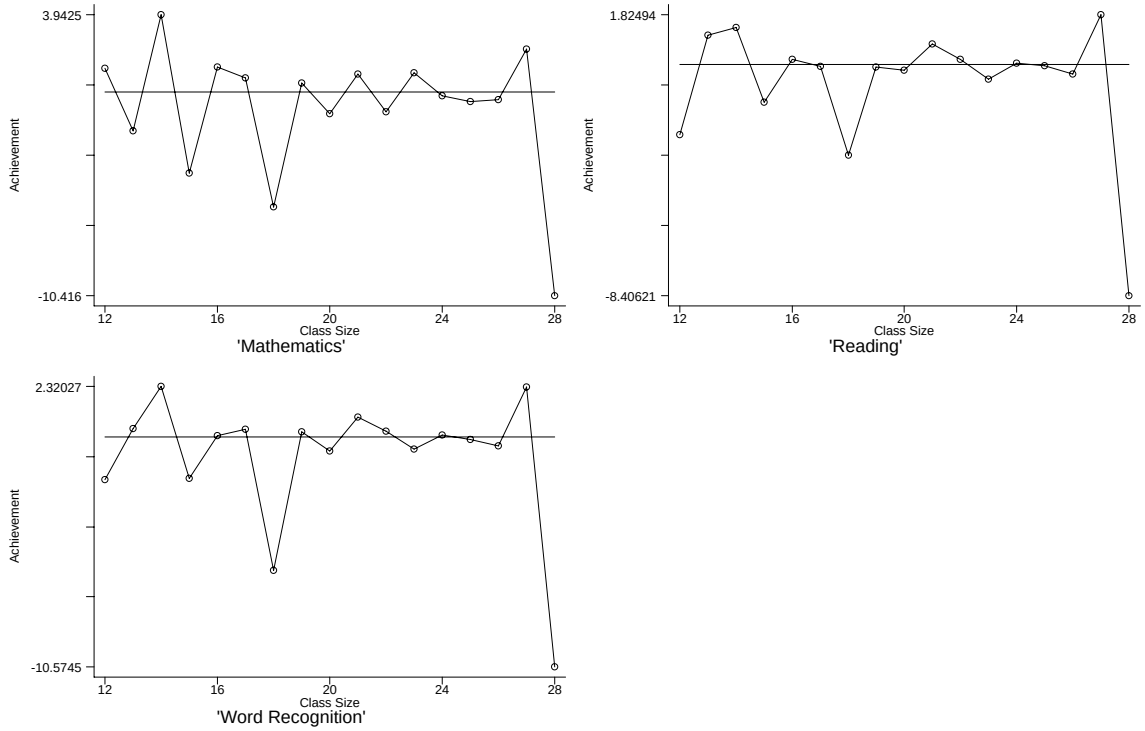
[21] Word, Elizabeth, John Johnston, Helen Bain, Dewayne B. Fulton, Jayne Boyd-Zaharias, Nan M. Lintz, Charles M. Achilles, John Folger and Carolyn Breda (1990), *Student/Teacher Achievement Ratio (STAR): Tennessee's K–3 Class-Size Study*, Nashville, TN: Tennessee State Department of Education.

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



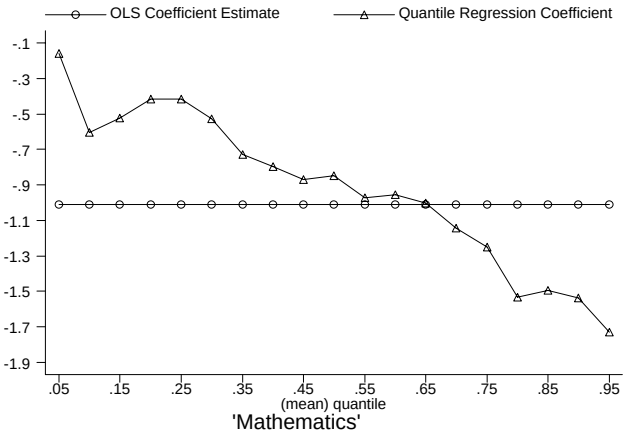
Note: In each figure, the density function of the scaled test score data is presented with by the curve connected by dots. The smooth line that does not contain any dot represents the Normal density curve.

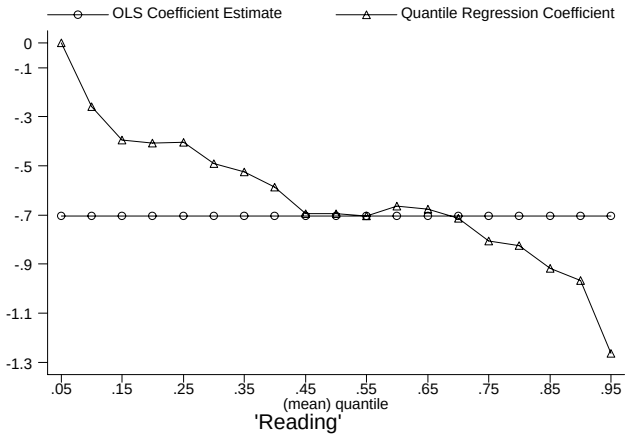
Figure 2: Nonparametric Estimate of the Effect of Kindergarten Class on Kindergarten Achievement



Note: In each figure, the straight line represents a zero impact.

Figure 3: Quantile Regression and OLS Estimates of the Impact of Class Size on Kindergarten Achievement





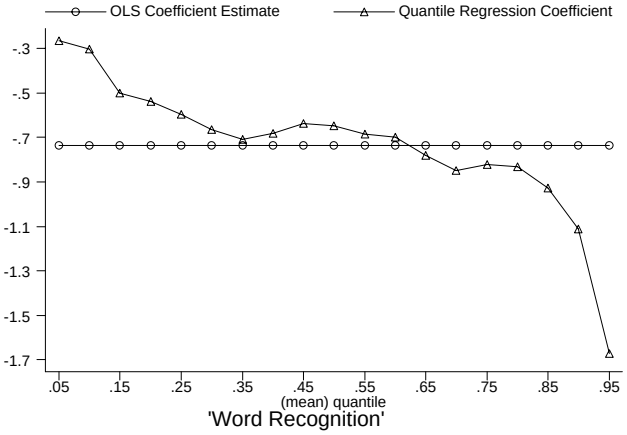


Table 1: Summary Statistics of the Project STAR Kindergarten Sample

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
Teacher is Not White	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: Flexible Estimates of the Class Size Effect on kindergarten Achievement

	Mathematics	Reading	Word Recognition
Class Size = 12	-4.206	-3.219	-0.264
	(7.199)	(5.025)	(6.021)
Class Size = 13	1.549	1.437	1.748
	(4.938)	(2.997)	(3.744)
Class Size = 14	15.080**	6.025	9.772**
	(5.534)	(3.402)	(4.429)
Class Size = 15	-0.584	2.893	2.391
	(5.667)	(3.935)	(3.445)
Class Size = 16	10.392**	7.512**	6.985*
	(4.024)	(3.019)	(3.315)
Class Size = 17	6.817	4.341	4.304
	(4.371)	(2.982)	(3.265)
Class Size = 18	-3.374	-4.136	-3.004
	(4.974)	(2.805)	(4.818)
Class Size = 19	-1.132	0.956	1.653
	(4.882)	(3.095)	(3.831)
Class Size = 20	-7.199	-0.584	1.161
	(6.185)	(3.446)	(3.374)
Class Size = 21	-4.381	-4.447	-2.722
	(4.301)	(3.144)	(3.651)
Class Size = 23	-4.784	-2.974	-2.214
	(3.737)	(2.429)	(2.485)
Class Size = 24	0.614	-0.368	-0.246
	(3.714)	(2.516)	(2.776)
Class Size = 25	1.544	-0.249	-0.778
	(5.510)	(3.532)	(4.055)
Class Size = 26	-16.743*	-9.893	-12.100*
	(7.556)	(5.744)	(6.635)
Class Size = 27	8.705	3.192	4.208
	(7.394)	(6.177)	(7.667)
Class Size = 28	-3.202	-4.294	6.190
	(6.637)	(3.564)	(4.970)
N	5810	5729	5789

Note: Standard errors corrected at the classroom level in parentheses. Regression equation includes information on school identifiers, children's gender, race and free lunch status, as well as teacher's race, education and years of experience.
* Significant at 5%; ** Significant at 1%

Table 3: Are Attritors Different from Non-Attritors?

Subject Area	Mathematics	Reading	Word Recognition
Kindergarten Class Size	-1.252**	-0.795**	-0.902**
	(0.306)	(0.182)	(0.209)
White or Asian Student	20.183**	8.446**	8.300**
	(2.769)	(2.004)	(2.526)
Female Student	2.578	3.341**	2.478*
	(1.365)	(1.074)	(1.296)
Student on Free Lunch	-13.688**	-12.233**	-13.895**
	(1.695)	(1.191)	(1.483)
Years of Teaching Experience	0.334	0.262*	0.337*
	(0.220)	(0.124)	(0.135)
White Teacher	-1.425	-1.927	-1.945
	(4.423)	(3.116)	(3.556)
Teacher Has Master Degree	-1.962	-1.506	-0.820
	(2.396)	(1.412)	(1.719)
Attrition Indicator	-32.800**	-20.236**	-23.016**
	(7.221)	(4.583)	(5.754)
Attrition Indicator Interacted with Kindergarten Class Size	0.670*	0.285	0.431
	(0.310)	(0.198)	(0.240)
Attrition Indicator Interacted with White or Asian Student	-3.622	-0.117	-0.968
	(2.756)	(1.829)	(2.377)
Attrition Indicator Interacted with Female Student	5.552*	2.915	3.720
	(2.079)	(1.455)	(1.734)
Attrition Indicator Interacted with Student on Free Lunch	-5.301*	-0.544	0.468
	(2.400)	(1.561)	(1.897)
Attrition Indicator Interacted with Years of Teaching Experience	0.190	0.079	-0.059
	(0.211)	(0.130)	(0.164)
Attrition Indicator Interacted with White Teacher	1.495	2.421	0.783
	(3.520)	(2.150)	(2.700)
Attrition Indicator Interacted with Teacher Has Master Degree	-1.095	1.042	1.701
	(2.513)	(1.589)	(1.879)
Number of Observations,	5810	5729	5789
R-Squared	0.304	0.294	0.258
Joint Effect of Attrition on Constant and Coefficient Estimates	42.22**	33.19**	26.28**
	[0.000]	[0.000]	[0.000]
Joint Effect on All Coefficient Estimates but not Constant	3.08**	1.39	1.58
	[0.003]	[0.207]	[0.135]
Effect of Attrition on Constant Alone	20.63**	19.50**	26.28**
	[0.000]	[0.000]	[0.000]

Note: Standard errors corrected at the classroom level in () parentheses. Probability > F are in the [] parentheses.

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

Table 4: Does The Impact of Class Size Vary by Student or Teacher Characteristics?

	Mathematics	Reading	Word Recognition	Mathematics	Reading	Word Recognition
Current Class Size	-1.365 (0.715)	-1.160 (0.487)*	-1.368 (0.588)*	N/A	N/A	N/A
Small Class Indicator	N/A	N/A	N/A	10.525 (6.024)	9.362 (4.475)*	10.164 (5.209)
Female Student	-1.513 (5.535)	4.622 (3.764)	1.812 (4.356)	7.795 (1.227)**	5.681 (0.900)**	5.629 (1.142)**
Student is White/Asian	23.740 (12.457)	8.657 (9.413)	5.759 (10.755)	16.255 (3.171)**	7.544 (2.106)**	7.044 (2.285)**
Current Free Lunch Status	-23.798 (6.907)**	-14.160 (5.214)**	-16.223 (5.822)**	-20.148 (1.788)**	-14.919 (1.204)**	-16.045 (1.408)**
Current Teacher is Non -White	20.867 (27.279)	8.184 (11.643)	-0.790 (14.073)	-3.127 (4.814)	-1.464 (3.739)	-1.697 (4.212)
Current Teacher has a Master's	5.919 (13.738)	-3.498 (8.025)	-0.844 (9.578)	-4.307 (2.958)	-0.484 (1.967)	0.047 (2.370)
Teacher Years of Experience	-0.907 (0.929)	-0.780 (0.647)	-0.776 (0.762)	0.582 (0.262)*	0.430 (0.165)*	0.414 (0.192)*
Class Size or Small Class* Female Student	0.389 (0.267)	0.036 (0.186)	0.155 (0.217)	-4.617 (2.259)*	-1.057 (1.716)	-2.157 (1.959)
Class Size or Small Class* White/Asian Student	-0.344 (0.582)	-0.052 (0.437)	0.052 (0.499)	1.505 (4.494)	0.416 (3.607)	-0.431 (3.833)
Class Size or Small Class* Free Lunch Status	0.180 (0.339)	-0.029 (0.250)	0.009 (0.281)	0.023 (3.090)	0.616 (2.157)	0.090 (2.320)
Class Size or Small Class* Teacher is Non -White	-1.015 (1.294)	-0.402 (0.584)	-0.015 (0.710)	12.015 (9.572)	5.216 (4.726)	2.960 (5.641)
Class Size or Small Class* Teacher has a Master's	-0.454 (0.698)	0.112 (0.416)	0.029 (0.498)	5.174 (5.943)	-1.444 (3.509)	-0.028 (4.268)
Class Size or Small Class* Teacher Years of Experience	0.067 (0.044)	0.054 (0.032)	0.054 (0.038)	-0.466 (0.410)	-0.412 (0.299)	-0.326 (0.342)
Constant	505.453 (14.983)**	457.247 (10.551)**	460.021 (12.662)**	474.532 (3.563)**	430.758 (2.435)**	429.022 (2.633)**
R-Squared	0.27	0.27	0.23	0.27	0.27	0.23
N	5810	5729	5789	5810	5729	5789

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

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

Table 5: Does The Impact of Education Production Function Inputs Vary by Race or Free Lunch Status?

	Mathematics	Reading	Word Recognition	Mathematics	Reading	Word Recognition
Current Class Size	-1.021 (0.292)**	-0.654 (0.172)**	-0.704 (0.205)**	-1.082 (0.308)**	-0.662 (0.195)**	-0.711 (0.233)**
Female Student	5.396 (1.391)**	5.325 (1.003)**	5.046 (1.240)**	5.432 (1.397)**	5.373 (1.011)**	5.072 (1.247)**
Student is Black	-28.617 (11.094)*	-10.498 (8.058)	-12.163 (8.922)	-22.778 (3.532)**	-11.005 (2.372)**	-10.547 (2.715)**
Current Free Lunch Status	-22.400 (1.580)**	-16.127 (1.055)**	-17.633 (1.228)**	-25.894 (7.916)**	-14.952 (5.514)**	-17.777 (6.252)**
Current Teacher is Non -White	-9.685 (5.566)	-4.365 (3.378)	-5.944 (4.015)	-2.946 (5.004)	-1.949 (3.047)	-3.301 (3.621)
Teacher has a Master's	-3.126 (2.236)	-1.291 (1.327)	-0.628 (1.587)	-3.810 (2.324)	-2.541 (1.473)	-1.465 (1.758)
Teacher Years of Experience	0.306 (0.229)	0.213 (0.135)	0.169 (0.158)	0.442 (0.224)*	0.311 (0.150)*	0.267 (0.173)
Black* Current Class Size	-0.068 (0.489)	-0.235 (0.348)	-0.204 (0.384)	N/A	N/A	N/A
Black* Female Student	3.057 (2.542)	0.016 (1.569)	-0.292 (1.858)	N/A	N/A	N/A
Black* Current Free Lunch Status	9.827 (2.962)**	6.384 (2.016)**	7.258 (2.379)**	N/A	N/A	N/A
Black* Teacher is Non -White	13.964 (5.914)*	6.187 (3.864)	6.975 (4.304)	N/A	N/A	N/A
Black* Teacher has a Master's	2.665 (4.333)	1.089 (2.976)	2.144 (3.262)	N/A	N/A	N/A
Black* Teacher Years of Experience	0.254 (0.393)	0.224 (0.252)	0.319 (0.266)	N/A	N/A	N/A
Free Lunch* Current Class Size	N/A	N/A	N/A	0.122 (0.348)	-0.126 (0.243)	-0.099 (0.279)
Free Lunch* Female Student	N/A	N/A	N/A	3.019 (2.556)	0.042 (1.568)	-0.250 (1.872)
Free Lunch* Black	N/A	N/A	N/A	7.540 (3.437)*	5.225 (2.291)*	5.901 (2.670)*
Free Lunch* Teacher is Non -White	N/A	N/A	N/A	5.849 (4.834)	3.589 (2.924)	4.010 (3.201)
Free Lunch* Teacher has a Master's	N/A	N/A	N/A	2.783 (2.829)	3.491 (1.875)	3.157 (2.139)
Free Lunch* Teacher Years of Experience	N/A	N/A	N/A	-0.028 (0.279)	-0.017 (0.183)	0.076 (0.195)
R-Squared	0.27	0.27	0.23	0.27	0.27	0.23
N	5810	5729	5789	5810	5729	5789

Note: Standard errors corrected at the classroom level in parentheses. Regression equation includes information on school identifiers.
* Significant at 5%; ** Significant at 1%