The Effect of Small Classes and Teacher Experience on Students' Performance: Empirical Evidence from Project STAR

by:
Marie-Hélène Brière

An essay submitted to the Department of Economics
in partial fulfillment of the requirements
for the degree of Master of Arts

Queen's University
Kingston, Ontario, Canada
August 2014

Acknowledgements

I would like to thank Professor Charles Beach, my essay supervisor, for his guidance on this paper, and his insightful suggestions and comments all through the process. Thanks go as well to PhD candidate Jeffrey Penney for assistance in finding the data and during the writing process. This research was supported by funding from the Social Sciences and Humanities Research Council.

Table of Contents

List of Tables

List of Graphs

Section 1: Introduction

In 1985, the Tennessee state legislature authorized and funded a large-scale research experiment in order to guide its policy decisions in public education. The project, named Project STAR (for Student Teacher Achievement Ratio) followed a cohort of students from 79 participating schools in the state from kindergarten to the third grade, collecting information on each student and teacher. Most significantly, the project randomly assigned the students in the experiment to three different class types: small (13-17 students), regular (22-25 students), and regular with a full-time teacher's aide. Students were also followed after the end of the project, when they returned to regular-sized classes in grade 4, and in subsequent years.

Such a randomized experiment is unique in its scale and has therefore been extensively analyzed in the literature. It led to positive findings regarding the effect of class size reductions on student achievement. The scale of the effect and its applicability over all grades remain contentious. However, there is general agreement on the beneficial effect of smaller classes in the earlier grades of the study. It also added to the discussion on the effect of teacher characteristics on student achievement. In general, the only observable teacher characteristic found to have an impact on student achievement is experience, although there is little consensus on how much of an impact experience has on students, both using Project STAR or other empirical evidence. Unobservable characteristics or non-measurable characteristics are those that have the most explanatory power on student outcomes. This will be discussed in further detail in Section 2 of this paper.

Steffen Mueller, in "Teacher Experience and the Class Size Effect – Experimental Evidence" (2013), included an interaction term between class size and teacher experience to determine whether the experience acquired by a teacher had any effect on the benefits of a smaller class. The present paper seeks to replicate the results of Mueller's study as well as expand his methodology to subsamples of students to determine if class size effects, teacher experience effects and their interaction vary between different groups of students. Examining Mueller's model is described in greater detail in Section 3.

Section 4 contains estimation results and a discussion of these results' implications. Section 5 addresses concerns with Project STAR. Then, Section 6 concludes.

Overall the findings of this paper confirm what is found in the literature. Specific groups of students, particularly black students and students from lower-income families, benefit more than their peers from being in a small class. Additionally, it is determined that a rookie teacher being in charge of their class less negatively impacts their academic outcomes. A policy looking to reduce class size on a large scale would have to keep these student differences as well as teacher effects in mind. Jepsen and Rivkin (2009) in "Class Size Reduction and Student Achievement: The Potential Tradeoff between Teacher Quality and Class Size" determine that a policy of class size reduction can be regressive as a result of teacher preferences. They found that schools with higher minority and low-income populations would, as a result of the policy, find themselves with a larger number of inexperienced teachers than higher-income or majority population schools, making the policy regressive at first. In order to yield an

effect that would reduce the equity gap, this paper finds that teacher experience would have to be taken into consideration by policy-makers.

Section 2: Literature Review

Class size

Reducing class sizes on a large scale would be an expensive endeavour; as such, governments that seek to enact such a policy ought to consult research regarding the effectiveness of a smaller class sizes at improving student outcomes. This cost concern leads Hanushek (1999) to review both non-experimental empirical evidence and the STAR experimental evidence to evaluate the effectiveness of smaller classes at raising student academic achievement.

A review of historical data from the United States and a comparative study of countries around the world lead Hanushek to the conclusion that a higher teacher-to-student ratio does not lead to any significant increase in student outcomes. If there are any effects, they are very small. It is important to stress, however, that the teacher-to-student ratio is not the same thing as class size. Class size represents the number of peers a student sits with in a room, while the teacher-to-student ratio measures the number of teachers to students in a school, or at times a school district, state, country, etc. The latter can increase if there are more specialized teachers to teach different subjects or if a larger number are dedicated to specialized education, while the number of students in a class can remain unchanged.

Upon his analysis of the STAR experiment's data Hanushek concludes that there is a beneficial effect on reading and mathematics scores. However, the absence of a widening gap between students assigned to a small class and those assigned to a regular class points, he states, to the absence of any benefit to continuous treatment,

and that rather, one year in a small class, early on, such as in kindergarten would be sufficient to see improvement in student test scores.

Rice (1999) analyzed how class-size reduction affects various classroom activities in high school and determined that smaller classes led to teachers spending more time working with small groups and more time leading whole group discussions. It also led to more time preparing for class and more time using innovative teaching techniques and less time spent keeping order in the classroom.

In the concluding report of Project STAR (Word et al. 1990), the conclusion is that the effects of a smaller class are most important in kindergarten and the first grade in reading and mathematics, though there is some benefit to class size reductions in the second and third grades as well. This is also consistent with the findings of Rivkin, Hanushek and Kain (2005) who, upon examining a dataset from Texas, find that the effect of a smaller class decreases as a student progresses on to higher grades.

However, Krueger (1999) found a positive effect of cumulative time being spent in a small class, though it is smaller than the initial effect of entering a small class. Krueger and Whitmore (2000), Chetty et al. (2011), and Finn and Achilles (1999) point to a long lasting impact of class size reduction, though it is not conclusive as to whether this is a result of longer exposure to a small class or just of a one-year (one-time) effect. Finn and Achilles, though, declare that there is no evidence supporting only a one-year treatment as opposed to a multiple-year exposure to smaller classes[1]. The absence of a widening gap, they contend, is due to the different learning challenges presented by

[1] "[...]the conclusion that 1 year of small classes is enough is not supported by any STAR results. The STAR analyses show that 3 to 4 years of small- class participation produce academic and behavior improvements that persist through Grade 7 and beyond." (Finn and Achilles, 1999, p.106).

each year. They found that a gap still existed between students who had been placed in a small class and those who had not beyond the duration of the STAR experiment, when all students were returned to regular sized classes in the fourth grade, until the seventh grade (no further observations had then been published). Nye et al. (1999) observed students until the eighth grade and found that the beneficial impact of exposure to small classes strongly persisted into grade 8, and that these lasting benefits were increasingly important the more years a student had spent in a small class.

Further into the future of these students, Krueger and Whitmore (2000) found that a higher proportion of students who had been in a small class had taken SAT or ACT college entrance exams. Chetty et al. (2011) followed the students in the sample into adulthood and linked higher class quality in kindergarten to higher adult outcomes, calculated using information on 401K, homeownership, income, marriage, etc. Class quality was positively affected by class size, as well as teacher experience.

Overall, there is general consensus regarding the benefits of smaller classes for students, though there is no agreement as to the magnitude of the impact of class reduction or what the optimal class size should be or how long the students should be placed in a small class. However, it would seem that the earlier grades are the most likely to yield larger benefits.[2] This is a period when students are "learning to learn", acquiring their basic skills and learning to interact with other children and to work in groups, and so these formative years are therefore most likely to respond to smaller class sizes.[3]

[2] Finn and Achilles (1999); Mosteller (1995); Rivkin, Hanushek and Kain (2005).
[3] Mosteller (1995).

Additionally, it appears that not all students benefit equally from being in smaller classes. It is well documented that minority students and students from lower-income areas or families reap greater benefit from smaller classes than their peers[4]. In regular classes, these students are commonly found to be lagging academically, so smaller classes can be a way to shorten the performance gap that exists between students. The difference in the effect on male and female students is more contentious, though the present study finds that there is a difference between the two genders.

Finn and Achilles (1999) link the benefits of smaller classes to increased student engagement. Since disengagement occurs more in minority students or students who attend inner-city schools (where a majority receive a free lunch), smaller classes are most likely to have a larger impact on these students. A student qualifies for a free lunch if his or her family income is below a state-established threshold. This implies that students receiving a free lunch are likely from lower-income families.

So, while a smaller class has beneficial impact on a student, the effect differs across students, and it is uncertain how long a student needs to be in a small class for a long-term effect to be present without cutting class size unnecessarily (when accounting for the costliness of such a policy). While Project STAR can be used to yield important information in that regard, it is important to keep in mind the variation in findings. In addition, experimental issues with Project STAR mean some effects may be mis-estimated.

[4] Finn and Achilles (1999); Krueger (1999); Krueger and Whitmore (2000); Mosteller (1995).

Teacher characteristics

Looking at the size of the class a student is in does not provide a complete picture of the variables that can affect a student's academic success. A student's teacher has an impact on student achievement, as well as family environment and background, possibly genetics, and other effects that cannot be measured. Even teacher characteristics cannot all be observed. Those that can be and that are observed in the STAR experiment are gender, race, teacher experience, education and on-the-job-training. These do not necessarily all explain teacher quality to any large degree, and it is teacher quality that is of interest. Rivkin, Hanushek and Kain (2005) used non-experimental empirical evidence to conclude that teacher quality actually had more of an impact on student achievement than class size, but that it is mostly unobservable.

Generally, though, they found that observable teacher characteristics have very little calculable impact on student outcomes. Other than experience, and that only to a small degree, observable teacher characteristics (education and professional development courses) were usually found to have no significant impact on test scores[5]. Most of the impact of teachers on student outcomes would be accounted for in a fixed effect, residual term, indicating that one cannot observe teacher quality in any particular characteristic presented by a teacher, though it has a significant impact on student achievement[6]. Teacher experience is the only characteristic that appears correlated to teacher quality. In their follow-up on adult outcomes of Project STAR

[5] Harris and Sass (2011).
[6] Rivkin, Hanushek and Kain (2005).

students, Chetty et al. (2011) found that a more experienced teacher in kindergarten was linked to higher subsequent adult earnings of the students.

While teacher experience has a small impact on student achievement in early grades, Harris and Sass (2011) find that most of the gains taking place in the first three years of a teacher's on-the-job experience. As such, in the later definition of a rookie teacher for the purposes of this paper, the cut-off of two or fewer years of experience is used as the definition of a rookie, since quality gains beyond two years of experience are likely to be minimal[7].

Class size and teacher interaction

In determining a policy to improve student achievement through class size reductions, it is important not to ignore the importance of teachers in such a policy. Should teacher quality have as large an impact on student outcomes, a class-size reduction policy should take teacher preferences into account. As Ehrenberg et al. (2001) point out, class-size reduction policies assume that teachers of equal quality are available to fill in the positions created by the policy, while it may be worthwhile to examine the effects of such a policy when that is not the case.

Jepsen and Rivkin (2009) do just that when they examine the policy enacted in California that reduced class size on a large-scale. They found that mandating smaller classes on such a mass-scale meant there was a sudden need for new teachers to fill new vacancies created by the policy. As such, there was an influx of inexperienced and

[7] These findings appear also Nye et al. (2004) who find that teacher quality gains are non-linear in experience, occurring more in the first few years of work.

uncertified teachers into those schools that enacted the policy. Lower experience and lack of certification meant negative impacts on students, on average, impacts that almost completely negated the benefits of the class size changes. In particular, schools in wealthier neighbourhoods were more likely to attract experienced teachers, and, as a result, most of the new teachers were hired in higher-poverty schools and schools with more minority students[8]. Consequently, the initial effect of the policy was found to be regressive. Since experience can be acquired over time, the policy was neutral in the longer term.

Mueller (2013) uses project STAR data to examine the interaction of teacher experience and class size. Mueller determines that more experienced teachers are those who manage to generate a class-size effect, while rookie teachers generate little to no benefit from smaller classes compared to regular classes. This is what the present study seeks to replicate and expand on. If a policy to reduce class size is likely to be regressive, it is important to determine the joint effect of teacher experience and class size on various sub-groups of students.

The findings of the present study indicate that students from lower-income families and black students, those most likely to be taught by rookie teachers following a policy of class-size reduction, are not as negatively impacted by a rookie teacher in a small class as their higher-income or white peers. However, if Jepsen and Rivkin's (2009) findings are correct, and senior teachers will concentrate in higher-income, predominantly white schools, then minority and lower income students will be at a disadvantage compared to white or high-income students since the positive marginal

[8] Achilles, Finn and Bain (1997) point out that studies on class size effects usually hold teacher quality constant and this means that such a policy would reduce the equity gap *if* applied equally to all students.

effect of a rookie teacher in a small class for an average black or low-income students is still smaller than the marginal effect of a small class with a senior teacher for white or high-income students. The distribution of teachers is therefore important information to policy-makers.

Section 3: Analytical Model[9]

The model in Mueller (2013) is based on Lazear (2001). Lazear posits that classroom instruction is a public good, and that it can suffer from being overcrowded. A disruption by one student affects all other students in the classroom, since the teacher has to stop instruction to handle the disruption.

Here, a student *i* in class *c* and school *s* obtains learning outcome L_{ics}, measured in test scores, as a function of quality class-time teaching which depends on class size *n,* teacher experience *E,* and other characteristics, $X_{ics,}$ like student gender, race and family income, teacher race, gender, and education as well as school specific characteristics.

Equation 1:

$$L_{ics} = p_{cs}^{n} \cdot q(n, E)_{cs} + X_{ics}$$

Quality classroom instruction is a function of the probability p_{cs}^{n} that a student is not disrupting the class, and the quality *q* of the teacher's instruction, which is a function of the number of students in the class *n* and the teacher's experience *E.* Consequently, the effect of class size is both direct and indirect, as determined by the first-order derivative of achievement with respect to class size,

Equation 2:

$$\frac{\partial L}{\partial n} = p^{n} \cdot \ln p \cdot q(n, E) + p^{n} \cdot \frac{\partial q(n, E)}{\partial n} \ .$$

Assuming that $\frac{\partial q(n,E)}{\partial n} \leq 0$, and since p≤1, then the above is negative. Assuming further that $\frac{\partial q(n,E)}{\partial E} \geq 0$, or that teachers with more experience provide higher quality instructions, as the literature indicates at the lower grade levels, and taking the first-

[9] This section comes from Section 3 of Mueller (2013).

order derivative of the above with respect to teacher experience to determine how teacher experience affects the sign and magnitude of the class size effect, it can be seen that a negative cross-derivative $\frac{\partial^2 L}{\partial n \partial E}$ would imply that the class size effect increases with teacher experience.

Equation 3

$$\frac{\partial^2 L}{\partial n \partial E} = p^n \left(\ln p \cdot \frac{\partial q(n,E)}{\partial E} + \frac{\partial^2 q(n,E)}{\partial n \partial E} \right) .$$

The sign of $\frac{\partial^2 L}{\partial n \partial E}$ depends on the sign of $\frac{\partial^2 q(n,E)}{\partial n \partial E}$. Should the latter be negative, one would then assume that class size reductions are more beneficial when teachers are more experienced, and the opposite should the sign of $\frac{\partial^2 q(n,E)}{\partial n \partial E}$ be positive.

It is also possible to assume that a teacher's experience affects the probability of a student disrupting the class *p(E)*. Adding this assumption to the model, I calculate the second order derivative as:

Equation 4[10]

$$\frac{\partial^2 L}{\partial n \partial E} = \left(p(E)^{n-1} \cdot \frac{\partial p(E)}{\partial E} \right) \left(\{ n \cdot \ln p(E) + 1 \} \cdot q(n,E) + n \cdot \frac{\partial q(n,E)}{\partial n} \right) + p(E)^n \left(\ln p \cdot \frac{\partial q(n,E)}{\partial E} + \frac{\partial^2 q(n,E)}{\partial n \partial E} \right) .$$

This is a correction of Mueller's (2013) equation, and it differs in two terms: the first term in parentheses and the last term in the second set of parentheses. The first term is originally stated as $\left(p(E)^n \cdot \frac{\partial p(E)^n}{\partial E} \right)$ by Mueller. The sign of this term is positive

[10] In Mueller (2013) : $\frac{\partial^2 L}{\partial n \partial E} = \left(p(E)^n \cdot \frac{\partial p(E)^n}{\partial E} \right) \left(\{ n \cdot \ln p(E) + 1 \} \cdot q(n,E) + n \cdot \frac{\partial q(n,E)}{\partial E} \right) + p(E)^n \left(\ln p \cdot \frac{\partial q(n,E)}{\partial E} + \frac{\partial^2 q(n,E)}{\partial n \partial E} \right)$. My own calculations did not concur with this equation.

regardless of the correction. The second change is the last term in the second set of parentheses, which Mueller writes as $\left(n \cdot \frac{\partial q(n,E)}{\partial E}\right)$ and assumes to be positive. However, $\left(n \cdot \frac{\partial q(n,E)}{\partial n}\right)$ is assumed to be negative in theory. Regardless of the corrections, it is possible that the sign of Equation 4 is positive, even if Equation 3 is negative, just as Mueller (2013) presents, though the sign condition is now different.

If $\left(\{n \cdot \ln p(E) + 1\} \cdot q(n,E) > (n) \cdot -\left\{\frac{\partial q(n,E)}{\partial n}\right\}\right)$ Equation 4 can be positive. As such, the following regression model seeks to estimate the effects of class size, teacher experience, and their joint effect on student test scores to determine how teacher experience affects the class-size effect on student performance:

Equation 5

$$Y_{icgs} = \beta_o + \beta_1 SMALL_{cgs} + \beta_2 ROOKIE_{cgs} + \beta_3 (SMALL_{cgs} \cdot ROOKIE_{cgs}) + \beta_k S_{icgs} + \beta_j T_{cgs} + \alpha_s + \gamma_g + \epsilon_{icgs} \ .$$

Separate regressions will be conducted for reading scores and math scores. Y_{icgs} is the test score for student *i* in grade *g,* class *c,* and school *s*. *S* is a vector of student characteristics, such as race, gender, and whether the student gets a free lunch. *T* is a vector containing teacher characteristics other than experience: education, race, and gender. A dummy variable for each school controls for school fixed effects, α_s, and a dummy for each grade controls for the growth in test scores over each grade γ_g.

Error terms, ϵ_{icgs}, will be correlated within students over time and within classes (teachers) in the cross-section. Mueller applies an estimation method, from Cameron, Gelbach and Miller (2011), which uses a non-nested two-way cluster structure to estimate the model. The randomness of the experimental design should also limit the impact of fixed effects at the individual level. However, the difference in

estimated standard errors between regular OLS with controls for school by grade fixed effects and the OLS model that accounts for two-way clustered errors by student and teacher (class) demonstrates that the latter is likely an important consideration. The method of two-way clustered errors estimation will be discussed in additional detail in Appendix A.

The STAR experiment data used in the present study is available from the Harvard University Henry A. Murray Research Archive.[11]

Explanatory variables

Small class

The intent of the STAR experiment was to assign each student to a class type, either a small class, with 13-17 students, or a regular class, with 22-25 students, which either had or didn't have a teacher's aide. There was some inconsistency in the actual size of the class due to students coming into schools later in the experiment and leaving partway through. As such, some small classes had more than 17 students and some regular classes had fewer than 22 students. The model uses "intent to treat" in assigning observations to small or regular class classification, so that a student observation is considered to be in a small class if that is the class type he or she was assigned to officially regardless of the actual number of people in the class. Krueger (1999) uses initial random assignment as an IV for class size to similar conclusions. The size of the sample minimizes the importance of variation in class size.

[11] The STAR data can be accessed from the Henry A Murray Research Archive at the following address: http://thedata.harvard.edu/dvn/dv/mra/faces/study/StudyPage.xhtml?studyId=18871

Some students also switched from small classes to regular classes within the same schools between grades. Mueller (2013) decided to remove all post-switching observations from the sample to be used in estimation. Mueller identifies 250 students who switched after the first grade and cuts them from the sample along with all subsequent observations for these students.

Table 1 presents my calculation of the number of students who switched from a small class to a regular class or vice versa between each grade. The difference between the number of switches I identify and that Mueller identifies is difficult to explain with the information at hand. Once all post-switching observations are removed, my sample for the regressions on math scores has 22,532 observations with complete information and 22,212 observations for the regressions on reading scores, compared with 21,748 and 21,443 respectively, which Mueller obtains.

Table 1: Within-school class-type switching

	Small to regular	Regular to small
Kindergarten to grade 1	89	214
Grade 1 to grade 2	27	163
Grade 2 to grade 3	46	175

In order to participate in Project STAR, schools had to have a minimum number of students (57) to create at least one class of each of the three types. Consequently, a little over one quarter of student observations in the sample ended up assigned to small classes, which is close to 13/57. A dummy variable is generated for students in a small class (see Table 2 and Table 3 below). A breakdown of students by class type is found in Tables 4 and 5. Tables are presented in pairs corresponding to the two student performance scores for math and reading

Teacher experience

As discussed earlier, teachers with fewer than three years of experience are designated as *rookie* teachers for the purpose of this model. Approximately one-eighth of the observations' teachers are rookies in this study (see Table 2 and Table 3). In addition, approximately 3.5% of all observations are both taught by rookies and assigned to small classes. It is important to note that each teacher is observed as many times as there were students in his or her class and these percentages do not, therefore, represent the teacher population accurately. As fewer students were assigned to small classes, those teachers assigned to small classes are observed fewer times than their colleagues teaching regular classes.

Other teacher characteristics

Other teacher characteristics controlled for in the model and observed by Project STAR are the teacher's gender, race, and education (highest degree obtained). A dummy variable is created for each characteristic. Race is grouped into white and non-white teachers (which groups teachers observed as black and other non-white), and education is grouped into teachers with at least a graduate degree and those without (having, in turn, either a bachelor's degree at most, or a specialist's degree). Summary statistics are found in Table 2 and Table 3. A breakdown of teacher race and highest attained degree are found in Table 4 and Table 5.

Tables 6 and 7 present the summary statistics for each regressor by grade. There is continuity in the distribution of student and teacher characteristics over each year; however, of note is the absence of any male teacher for the kindergarten grade. Their

number increases with grade, as does the higher proportion of non-white teachers in the higher level classes compared to kindergarten and grade 1.

Student characteristics

Student characteristics controlled for in the model are gender, race, and free-lunch status. Races and ethnicities observed in the sample are white, black, Asian, Hispanic, Native American and others. A dummy variable is used in the model for non-white students. Male and female students are differentiated as well with the use of a dummy, as are students on a free lunch and those not. Statistics on the dummy variables are found in Tables 2 and 3, while a breakdown of students by race is found in Tables 4 and 5.

Looking at the mean of the *Male student* variable, one can see that the male to female ratio is about even. This holds across all grades as well (see Tables 6 and 7). Approximately half the students receive a free lunch as well, though the distribution of students who receive a free lunch is very different between white and non-white students. Among white students, 33% receive a free lunch, while 80% of students who are not white receive a free lunch. In addition, while White students represent 66% of the sample, they represent only 44% of students who receive free lunches.

Table 2: Summary statistics of explanatory variables for math scores regression

Variable	Obs	Mean	Std. Dev.	Min	Max
Male student	22532	0.5178	0.4997	0	1
Non-White student	22532	0.3383	0.4731	0	1
Free lunch	22532	0.4955	0.5000	0	1
Male teacher	22532	0.0114	0.1062	0	1
Non-White teacher	22532	0.1880	0.3907	0	1
Graduate degree	22532	0.3692	0.4826	0	1
Rookie teacher	22532	0.1259	0.3318	0	1
Small class	22532	0.2749	0.4465	0	1
Small*rookie	22532	0.0352	0.1844	0	1

Table 3: Summary statistics of explanatory variables for reading scores regression

Variable	Obs	Mean	Std. Dev.	Min	Max
Male student	22212	0.5173	0.4997	0	1
Non-White student	22212	0.3406	0.4739	0	1
Free lunch	22212	0.4942	0.5000	0	1
Male teacher	22212	0.0117	0.1074	0	1
Non-White teacher	22212	0.1893	0.3918	0	1
Graduate degree	22212	0.3698	0.4827	0	1
Rookie teacher	22212	0.1248	0.3305	0	1
Small class	22212	0.2758	0.4469	0	1
Small*rookie	22212	0.0344	0.1824	0	1

Table 4: Breakdown of student, teacher and class characteristics, math scores

Variables	Observations	Percent
Student race		
White	14,910	66.17
Black	7,497	33.27
Asian	58	0.26
Hispanic	26	0.12
Native American	8	0.04
Other	33	0.15
Teacher's highest degree		
Bachelor's	14,003	62.15
Master's	8,112	36
Master's +	151	0.67
Specialist	210	0.93
Doctoral	56	0.25
Teacher's race		
White	18,296	81.2
Black	4,225	18.75
Asian	11	0.05
Class type		
Small class	6,195	27.49
Regular class	8,008	35.54
Regular + aide class	8,329	36.97

Table 5: Breakdown of student, teacher and class characteristics, reading scores

Variables	Freq.	Percent
Student race		
White	14646	65.94
Black	7441	33.5
Asian	58	0.26
Hispanic	26	0.12
Native American	8	0.04
Other	33	0.15
Teacher's highest degree		
Bachelor's	13813	62.19
Master's	8008	36.05
Master's +	149	0.67
Specialist	186	0.84
Doctoral	56	0.25
Teacher's race		
White	18,007	81.07
Black	4,194	18.88
Asian	11	0.05
Class type		
Small class	6127	27.58
Regular class	7897	35.55
Regular + aide class	8188	36.86

Table 6: Summary statistics of explanatory variables by grade, math scores

Variable	Obs	Mean	Std. Dev.	Min	Max
Kindergarten					
Male student	5809	0.5137	0.4999	0	1
Non-White student	5809	0.3298	0.4702	0	1
Free lunch	5809	0.4834	0.4998	0	1
Male teacher	5809	0	0	0	0
Non-White teacher	5809	0.1594	0.3661	0	1
Graduate degree	5809	0.3458	0.4757	0	1
Rookie teacher	5809	0.1377	0.3446	0	1
Small class	5809	0.3025	0.4594	0	1
Small*rookie	5809	0.0475	0.2128	0	1
Grade 1					
Male student	6079	0.5180	0.4997	0	1
Non-White student	6079	0.3404	0.4739	0	1
Free lunch	6079	0.5139	0.4998	0	1
Male teacher	6079	0.0041	0.0640	0	1
Non-White teacher	6079	0.1760	0.3809	0	1
Graduate degree	6079	0.3446	0.4753	0	1
Rookie teacher	6079	0.1597	0.3664	0	1
Small class	6079	0.2637	0.4407	0	1
Small*rookie	6079	0.0352	0.1843	0	1
Grade 2					
Male student	5348	0.5208	0.4996	0	1
Non-White student	5348	0.3528	0.4779	0	1
Free lunch	5348	0.4942	0.5000	0	1
Male teacher	5348	0.0116	0.1071	0	1
Non-White teacher	5348	0.2066	0.4049	0	1
Graduate degree	5348	0.3601	0.4801	0	1
Rookie teacher	5348	0.1208	0.3259	0	1
Small class	5348	0.2620	0.4397	0	1
Small*rookie	5348	0.0316	0.1750	0	1
Grade 3					
Male student	5296	0.5193	0.4997	0	1
Non-White student	5296	0.3304	0.4704	0	1
Free lunch	5296	0.4890	0.4999	0	1
Male teacher	5296	0.0321	0.1763	0	1
Non-White teacher	5296	0.2143	0.4104	0	1
Graduate degree	5296	0.4322	0.4954	0	1
Rookie teacher	5296	0.0793	0.2702	0	1
Small class	5296	0.2708	0.4444	0	1
Small*rookie	5296	0.0255	0.1576	0	1

Table 7: Summary statistics of explanatory variables by grades, reading scores

Variable	Obs	Mean	Std. Dev.	Min	Max
Kindergarten					
Male student	5728	0.5134	0.4999	0	1
Non-White student	5728	0.3280	0.4695	0	1
Free lunch	5728	0.4843	0.4998	0	1
Male teacher	5728	0	0	0	0
Non-White teacher	5728	0.1583	0.3651	0	1
Graduate degree	5728	0.3478	0.4763	0	1
Rookie teacher	5728	0.1356	0.3424	0	1
Small class	5728	0.3027	0.4595	0	1
Small*rookie	5728	0.0454	0.2082	0	1
Grade 1					
Male student	5902	0.5169	0.4998	0	1
Non-White student	5902	0.3462	0.4758	0	1
Free lunch	5902	0.5100	0.4999	0	1
Male teacher	5902	0.0042	0.0650	0	1
Non-White teacher	5902	0.1784	0.3829	0	1
Graduate degree	5902	0.3440	0.4751	0	1
Rookie teacher	5902	0.1596	0.3663	0	1
Small class	5902	0.2662	0.4420	0	1
Small*rookie	5902	0.0359	0.1861	0	1
Grade 2					
Male student	5354	0.5192	0.4997	0	1
Non-White student	5354	0.3521	0.4777	0	1
Free lunch	5354	0.4944	0.5000	0	1
Male teacher	5354	0.0116	0.1070	0	1
Non-White teacher	5354	0.2066	0.4049	0	1
Graduate degree	5354	0.3603	0.4801	0	1
Rookie teacher	5354	0.1203	0.3253	0	1
Small class	5354	0.2624	0.4400	0	1
Small*rookie	5354	0.0314	0.1744	0	1
Grade 3					
Male student	5228	0.5199	0.4997	0	1
Non-White student	5228	0.3365	0.4725	0	1
Free lunch	5228	0.4872	0.4999	0	1
Male teacher	5228	0.0329	0.1784	0	1
Non-White teacher	5228	0.2179	0.4128	0	1
Graduate degree	5228	0.4327	0.4955	0	1
Rookie teacher	5228	0.0782	0.2686	0	1
Small class	5228	0.2710	0.4445	0	1
Small*rookie	5228	0.0239	0.1528	0	1

Dependent variables

In order to determine the effect of various characteristics on student achievement, the STAR experiment collected test scores in each grade for reading and mathematics. Both curriculum based state testing (BSF) was administered as well as SAT standardized testing. SAT scores will be used for this model, as the scores are comparable across grades. Following Mueller (2013), the scores will first be transformed to a standardized mean of 0 and a standard deviation of 1.

A look at the distribution of test scores across various stratifications of the sample sheds some light on the nature of test scores and how they compare across different students. Statistics on test scores for various subgroups of students are presented in Tables 8 and 9, and kernel density graphs of the test score distributions are shown to illustrate these differences.

There are clear disparities in test scores between different groups of students. In particular, Black students have a lower mean test scores than White students, and also have a lower standard deviation. While males and females have very similar standard deviations in test scores, the mean test score for males is about as far from zero as the mean test score for females, but it is negative while the female mean test score is positive. This difference is more marked in reading scores than math scores. A similar relationship appears between scores for students receiving a free lunch and those not receiving a free lunch, those receiving a free lunch equally far from a zero mean as those not receiving a free lunch, but on the negative side as opposed to their peers. Since these male to female and free-lunch to no-free-lunch divides are about even, it is to be

expected that the two groups should average out to zero since the scores are normalized to zero mean.

Clearly, the scores improve from one year to the next, which is possibly due to the SAT test scores being comparable from grade to grade. It would be expected for students to improve over time.

Of note is the bimodal distribution of reading scores over the entirety of the sample, likely due to the uneven growth in test scores from one grade to the next. First grade reading scores are distributed with very fat tails, which would indicate that the pace in the learning of reading varies across students particularly in that grade. As this is the grade when reading tends to become an important part of the curriculum, this is of particular interest.

The difference between students of different races is marked, as is the difference between students who receive a free lunch and those who do not (Graphs 2, 4, 6, and 8). These two groups appear to have lower test scores than their peers. Graphs 9 and 10 show that small classes tend to yield better test scores and graphs 11 and 12 show that more experienced teachers tend to generate higher test scores as well. As such, it is of interest to determine how class-size effect and teacher experience might affect different groups of students to determine if policy makers could better allocate school resources to benefit students, to use smaller classes and experienced teachers where they are more effective.

The mathematical model of Lazear (2001) and Mueller (2013) also implies that a smaller class is not a perfect substitute for well-behaved students. Learning outcomes in a larger class with well-behaved students (with p=0.99 for example) would still be

higher than learning outcomes in a smaller class with less-well-behaved students (with p=0.97 for example). However, lower p students would see a larger benefit from a reduction in class size. As such, if minority students or students on free lunch tend to be less engaged, in other words, tend to have a lower p, then one could expect to see larger marginal effects from a class size reduction.

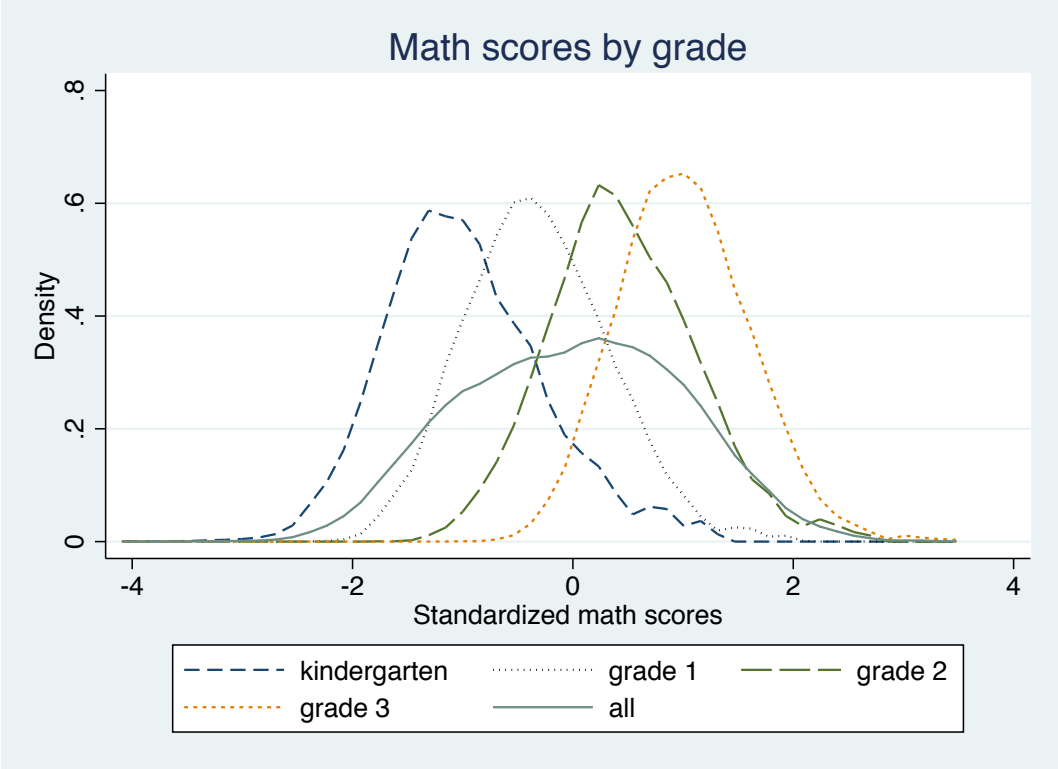
Table 8: Distribution of math scores across various subsamples

Math scores	Obs.	Mean	Std. Dev.	Min	Max
By grade					
Kindergarten	5809	-0.9917	0.7196	-3.9665	1.1253
Grade 1	6079	-0.3170	0.6514	-2.2190	1.8785
Grade 2	5348	0.4433	0.6655	-1.6616	2.5564
Grade 3	5296	1.0039	0.5933	-0.9686	3.3549
By student race					
White	14910	0.1225	1.0006	-3.4844	3.3549
Black	7497	-0.2467	0.9497	-3.9665	3.0234
Asian	58	0.2376	1.0822	-1.9328	3.0234
Hispanic	26	0.2145	0.9548	-1.3302	1.9087
Native American	8	-0.8707	0.7683	-2.0985	0.3118
Other	33	0.3351	1.3146	-2.5203	3.3549
By gender					
Male	11668	-0.0120	1.0189	-3.9665	3.3549
Female	10864	0.0128	0.9792	-3.4844	3.3549
By free lunch or not					
Free lunch	11165	-0.1954	0.9718	-3.9665	3.3549
No free lunch	11367	0.1919	0.9901	-2.8065	3.3549
By class type					
Small class	6195	0.0664	1.0176	-3.9665	3.0234
Regular class	8008	-0.0722	0.9823	-3.4844	3.3549
Regular + aide class	8329	0.0200	0.9995	-3.1982	3.3549

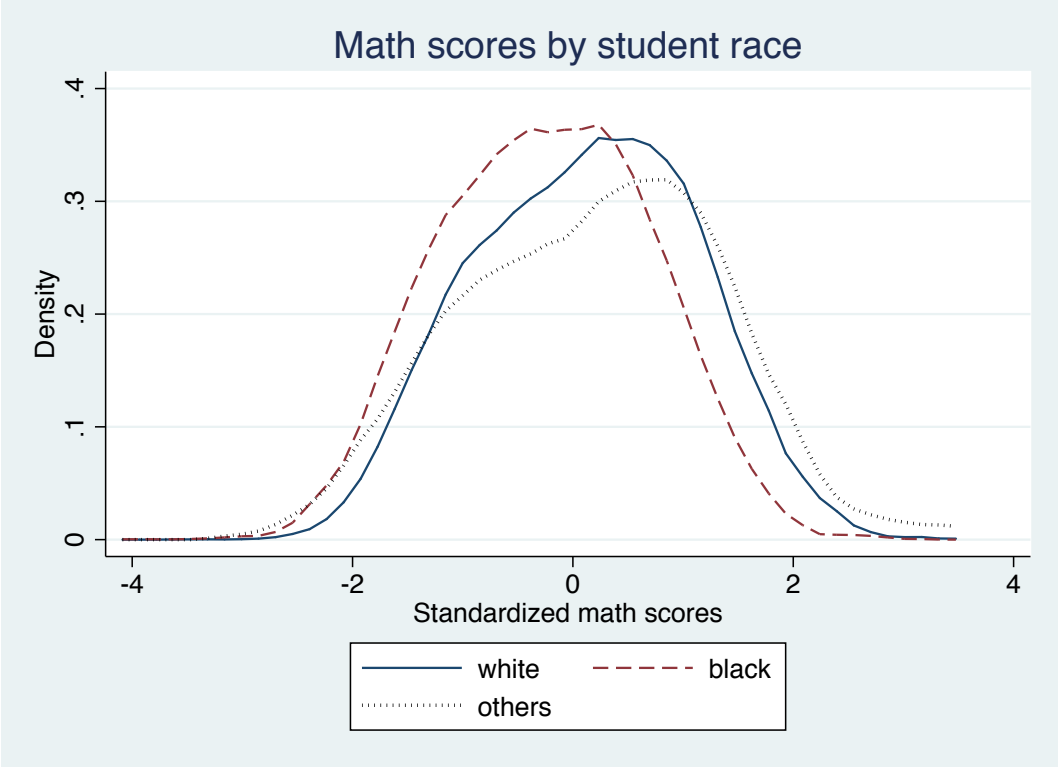
Table 9: Distribution of reading scores across various subsamples

Reading scores	Obs.	Mean	Std. Dev.	Min	Max
By grade					
Kindergarten	5728	-1.2323	0.3923	-2.7353	1.1146
Grade 1	5902	-0.1968	0.6813	-1.6371	1.4107
Grade 2	5354	0.5859	0.5664	-0.8474	2.4102
Grade 3	5228	0.9723	0.4722	-0.4649	2.9408
By student race					
White	14646	0.1035	1.0232	-2.7353	2.9408
Black	7441	-0.2079	0.9164	-2.3528	2.9408
Asian	58	0.3097	1.1086	-1.6865	2.6693
Hispanic	26	0.3153	0.9922	-1.3903	1.7685
Native American	8	-0.6947	0.7050	-1.7482	0.0534
Other	33	0.3211	1.1079	-1.8839	2.1757
By gender					
Male	11490	-0.0465	0.9953	-2.7353	2.9408
Female	10722	0.0498	1.0027	-2.3528	2.9408
By free lunch or not					
Free lunch	10978	-0.1789	0.9485	-2.7353	2.9408
No free lunch	11234	0.1748	1.0180	-2.0567	2.9408
By class type					
Small class	6127	0.0431	1.0278	-2.0567	2.9408
Regular class	7897	-0.0711	0.9770	-2.7353	2.9408
Regular + aide class	8188	0.0363	0.9972	-2.0320	2.9408

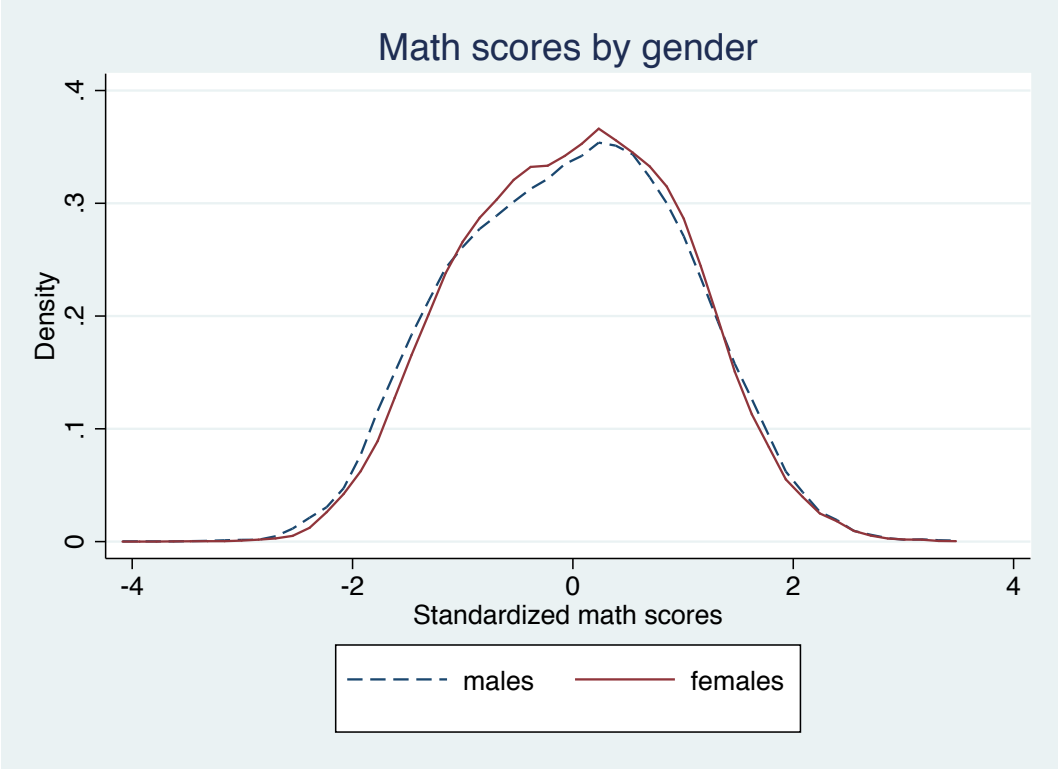
Graph 1: Math scores by grade



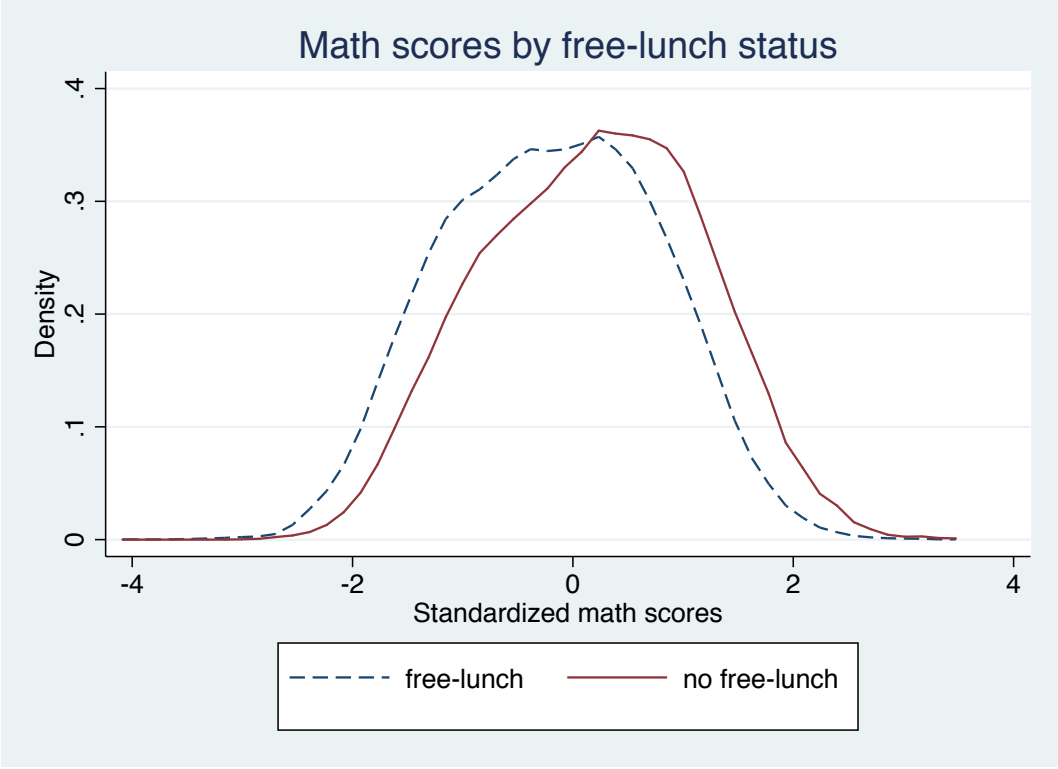
Graph 2: Math scores by student race



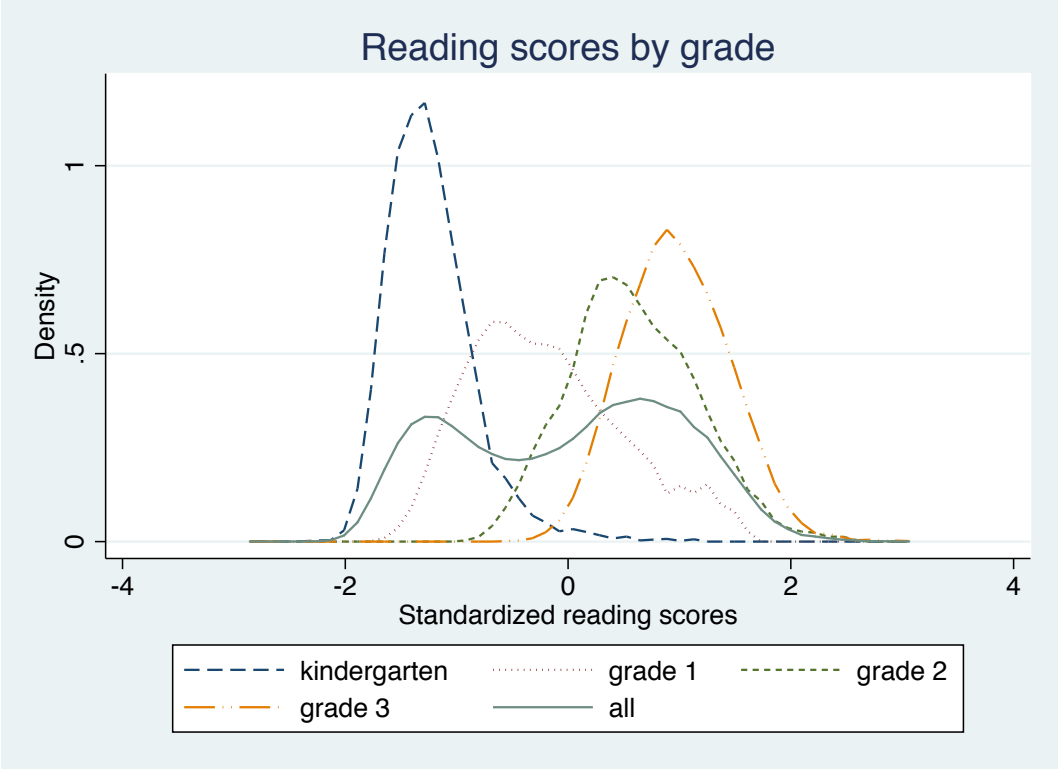
Graph 3: Math scores by gender



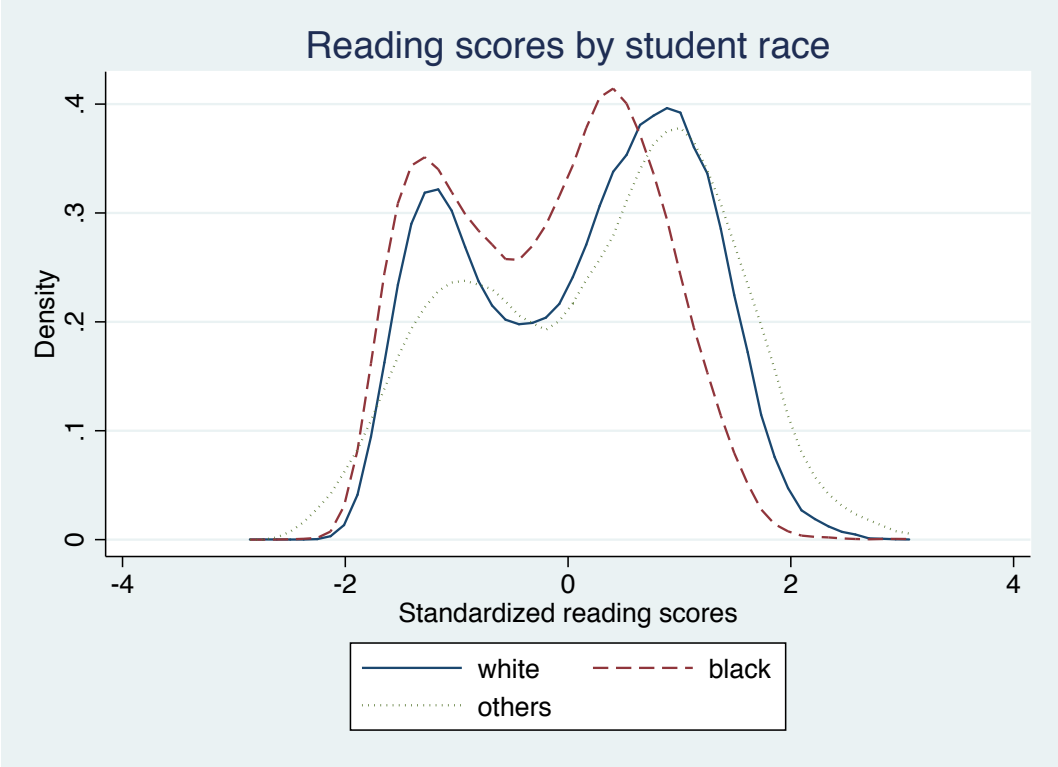
Graph 4: Math scores by free-lunch status



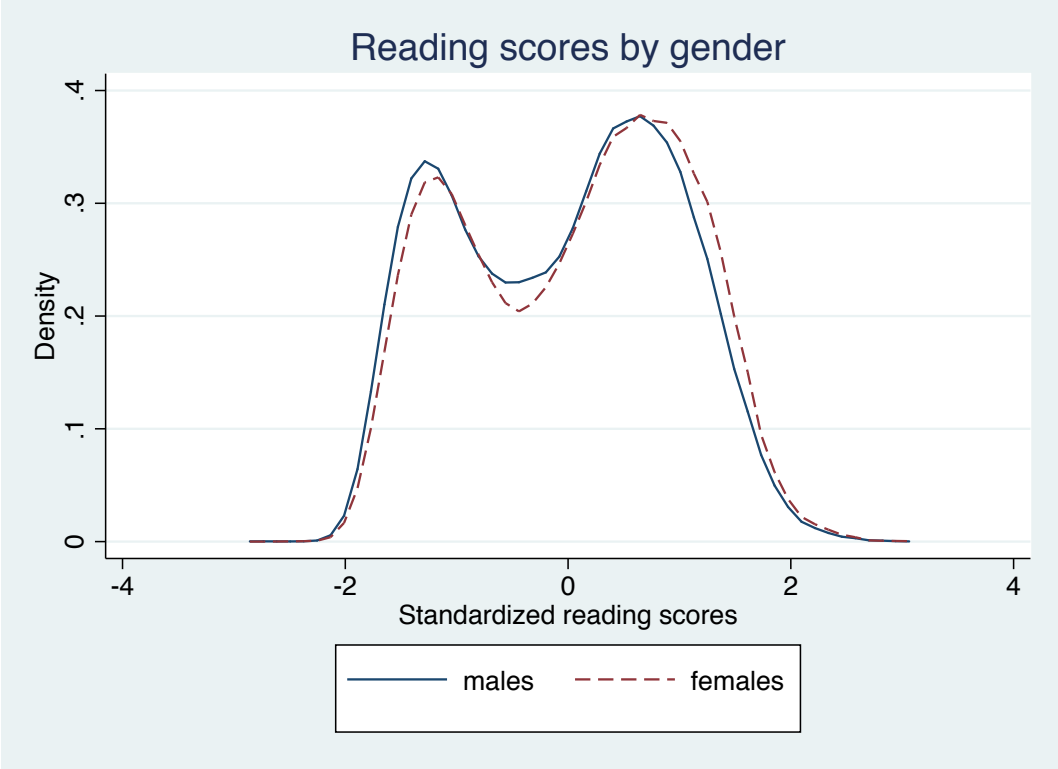
Graph 5: Reading scores by grade



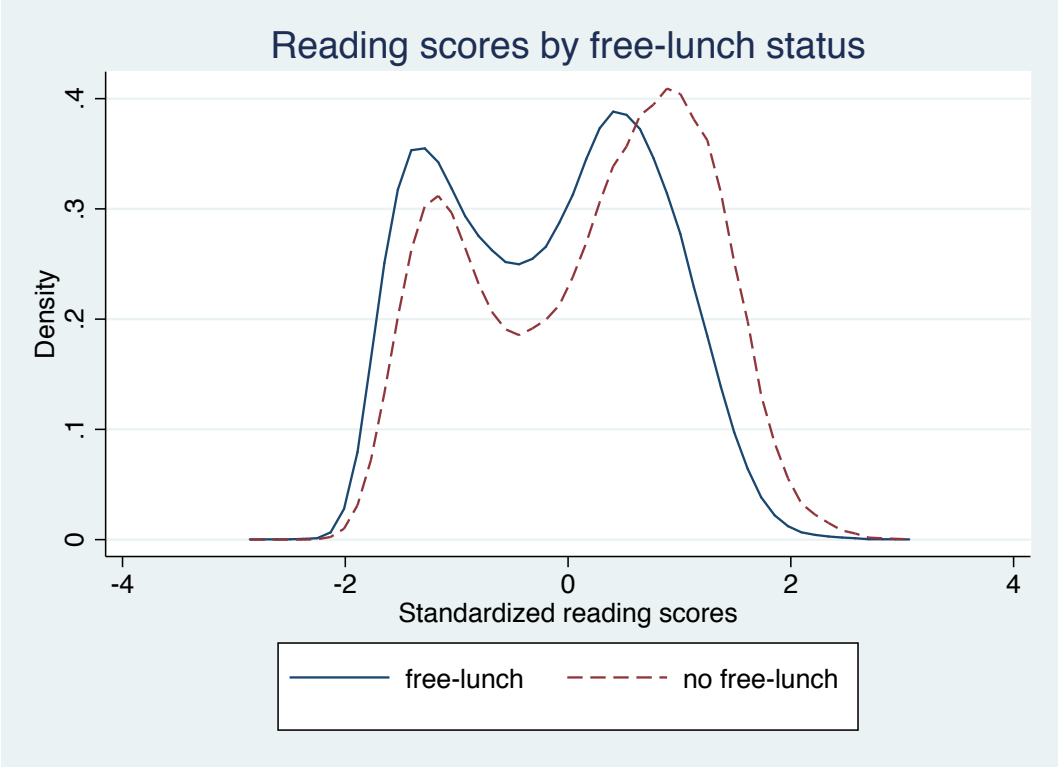
Graph 6: Reading scores by student race



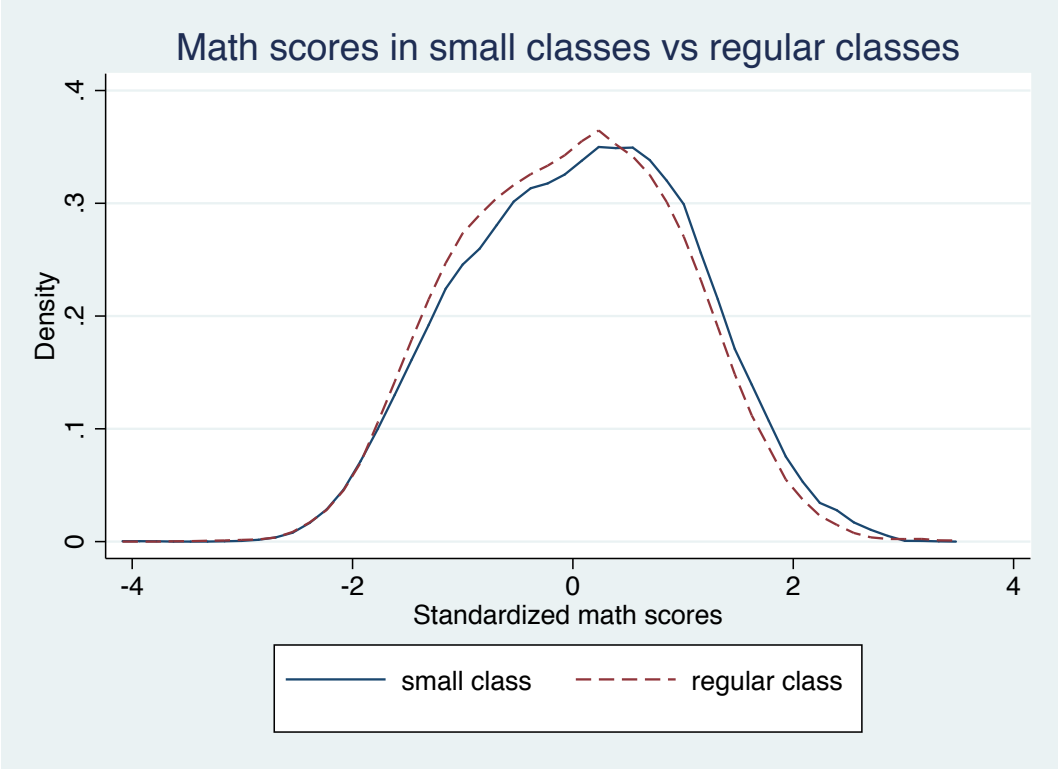
Graph 7: Reading scores by gender



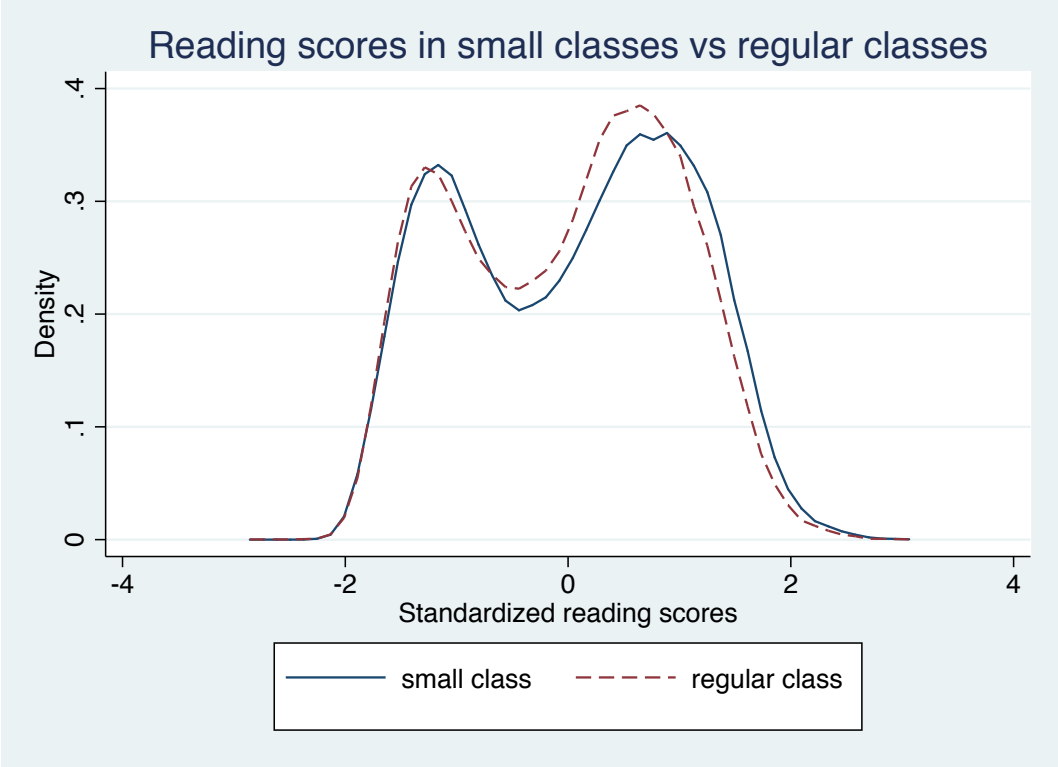
Graph 8: Reading scores by free-lunch status



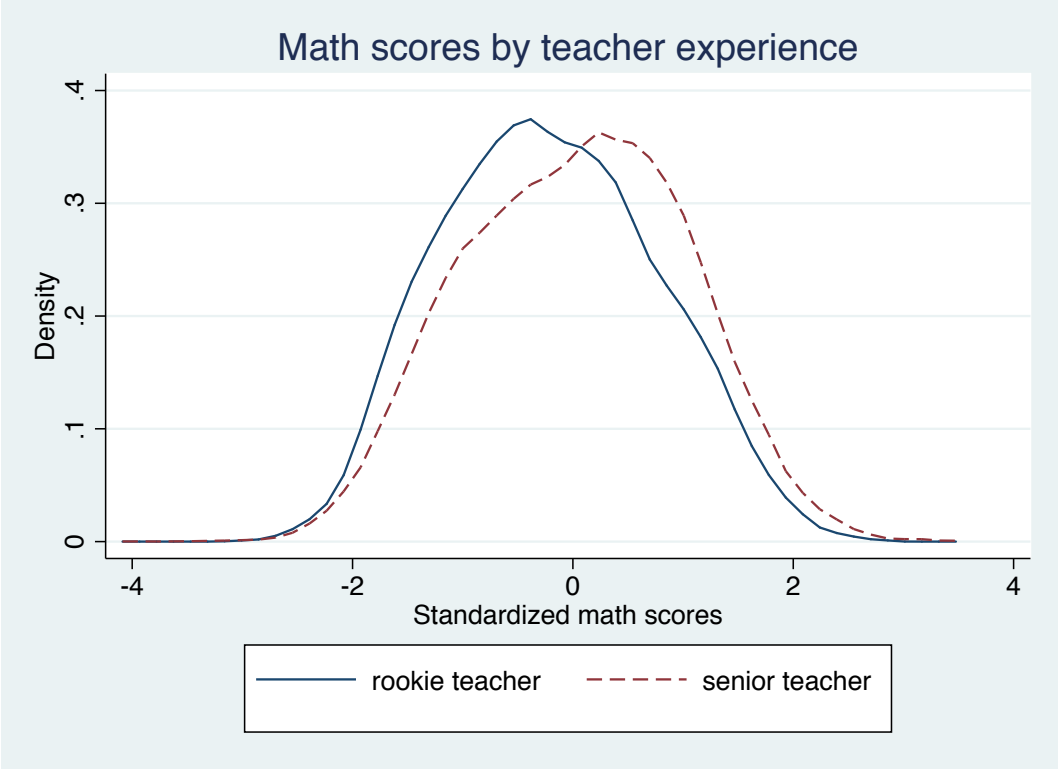
Graph 9: Match scores in small classes vs. regular classes



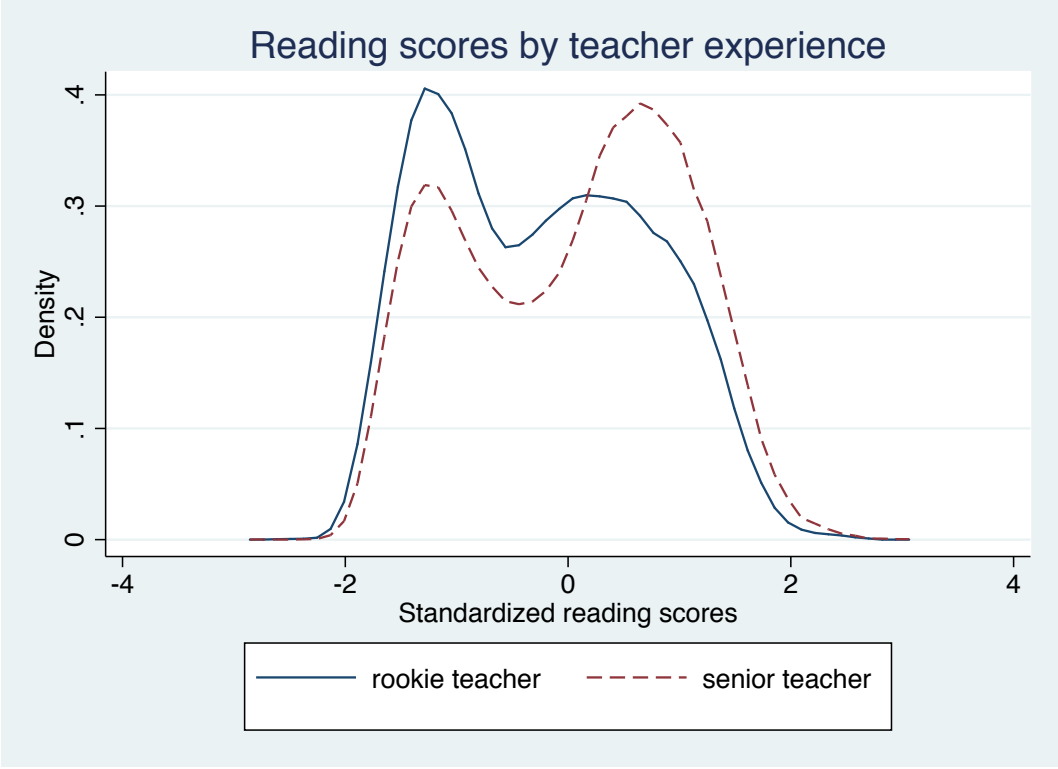
Graph 10: Reading scores in small classes vs. regular classes



Graph 11: Math scores by teacher experience



Graph 12: Reading scores by teacher experience



Section 4: Regression Estimation Results

Mueller's (2013) results indicate that "only senior [teachers] generate class size effects and that the class size effect likely comes through an increase in teaching quality per unit of instructional time" (Mueller, p. 47). The first purpose of the present study is to replicate this finding and then to examine the results when the same theory is applied to specific subgroups of students. Just like Mueller (2013), the OLS regression here controls for school fixed effects with dummy variables for each school, α_s. Controlling for school fixed effects accounts for the fact that the randomization of students and teachers was done at the school level, and this controls for school-specific characteristics. Additionally, data are pooled over all grades, with grades controlled for by dummy variables for each grade, γ_g. Again, the equation to be estimated (Equation 5) is as follows:

$$Y_{icgs} = \beta_o + \beta_1 SMALL_{cgs} + \beta_2 ROOKIE_{cgs} + \beta_3 (SMALL_{cgs} \cdot ROOKIE_{cgs}) + \beta_k S_{icgs} + \beta_j T_{cgs} + \alpha_s + \gamma_g + \epsilon_{icgs} \quad .$$

The coefficient estimates reported by Mueller are presented in Table 10 while those I estimate by running the same regression on the sample I obtained following the same methodology as Mueller are reported in Table 11. The coefficient estimates for the teacher characteristics and student characteristics (T_{cgs} and S_{icgs}) are also reported. Mueller (2013) did not report these control variable coefficients. The dummy variables' coefficient estimates are reported for the grades, but not for the 79 schools in the

sample. Kindergarten is the default grade and a dummy for each of grades 1 through 3 is in the model and reported in the estimation results tables.

In order to determine how different subsamples of students were affected by a reduction in class size or to teachers with different amounts of experience, the same regression is also run on different subsamples of students. Mueller (2013) did not provide any stratified regression results except for the unrestricted quantile regression performed on students at each decile of test scores. In addition to the latter, in this study, students are separated into three different racial subgroups, white, black, and others; separated by gender; separated according to whether or not they received a free lunch; separated by grade; and separated by grade and gender. Tables 12 to 19 report the regression estimates for each subsample.

For analysis, marginal effects of a small class and a rookie teacher where intervention terms are present are calculated and presented in Tables 20 and 21.

Tables 22 and 23 present the results of the unrestricted quantile regression for each decile of student achievement. The method used is that developed by Firpo, Fortin and Lemieux (2009)[12]

Table 10: OLS regression results from Mueller (2013)

Variable	Math	Reading
Small	0.162*** (0.021)	0.136*** (0.016)
Rookie	0.036 (0.033)	-0.005 (0.024)
Small*Rookie	-0.143*** (0.052)	-0.125*** (0.039)

(note: standard errors in parentheses)
*** p<0.01, ** p<0.05, * p<0.1

[12] The unconditional quantile regression Stata files are available on Dr. Nicole M. Fortin's website: http://faculty.arts.ubc.ca/nfortin/datahead.html (last accessed Jul. 30, 2014).

Table 11: OLS regression results

Whole sample		
VARIABLES	Math	Reading
small	**0.153*** **	**0.125*** **
	(0.021)	(0.016)
rookie	**0.029**	**-0.001**
	(0.033)	(0.024)
small*rookie	**-0.143*** **	**-0.120*** **
	(0.052)	(0.039)
male student	-0.033***	-0.105***
	(0.011)	(0.009)
non-white student	-0.224***	-0.098***
	(0.021)	(0.018)
free lunch	-0.278***	-0.255***
	(0.013)	(0.011)
male teacher	0.015	0.000
	(0.058)	(0.045)
non-white teacher	0.030	0.022
	(0.029)	(0.020)
teacher grad.	0.021	0.010
	(0.018)	(0.013)
grade 1	0.680***	1.042***
	(0.024)	(0.018)
grade 2	1.433***	1.817***
	(0.025)	(0.016)
grade 3	1.995***	2.202***
	(0.025)	(0.016)
constant	-1.135***	-1.256***
	(0.079)	(0.057)
observations	22,532	22,212
R-squared	0.646	0.767

(note: standard errors in parentheses)
*** p<0.01, ** p<0.05, * p<0.1

Table 12: OLS regression results by student race

By race	Black		White		Other	
VARIABLES	math	reading	math	reading	math	reading
small	**0.196*** **	**0.171*** **	**0.136*** **	**0.106*** **	**-0.071**	**0.130**
	(0.038)	(0.027)	(0.023)	(0.018)	(0.183)	(0.199)
rookie	**0.038**	**-0.005**	**0.034**	**0.013**	**0.312**	**0.574**
	(0.055)	(0.033)	(0.036)	(0.027)	(0.366)	(0.401)
small*rookie	**-0.092**	**-0.088**	**-0.212*** **	**-0.177*** **	**-0.300**	**-0.677**
	(0.083)	(0.055)	(0.056)	(0.042)	(0.482)	(0.553)
male student	-0.0508***	-0.0961***	-0.022	-0.109***	-0.194	0.022
	(0.018)	(0.013)	(0.014)	(0.012)	(0.159)	(0.177)
free lunch	-0.230***	-0.200***	-0.280***	-0.262***	-0.368	-0.128
	(0.023)	(0.019)	(0.015)	(0.013)	(0.254)	(0.236)
male teacher	-0.125*	-0.097	0.065	0.032	-0.536**	-0.845***
	(0.076)	(0.079)	(0.060)	(0.047)	(0.241)	(0.232)
non-white teacher	0.0833**	0.0587***	-0.0888**	-0.0512**	-0.156	-0.355
	(0.033)	(0.022)	(0.039)	(0.025)	(0.281)	(0.266)
teacher grad.	0.050	0.035	0.013	0.005	-0.289**	-0.274***
	(0.036)	(0.023)	(0.019)	(0.014)	(0.141)	(0.100)
grade 1	0.560***	0.810***	0.736***	1.155***	0.892***	1.324***
	(0.048)	(0.029)	(0.023)	(0.018)	(0.155)	(0.138)
grade 2	1.296***	1.628***	1.502***	1.911***	1.818***	1.987***
	(0.051)	(0.028)	(0.025)	(0.017)	(0.186)	(0.179)
grade 3	1.920***	2.091***	2.028***	2.253***	2.490***	2.368***
	(0.055)	(0.030)	(0.025)	(0.016)	(0.209)	(0.156)
constant	-1.312***	-1.276***	-1.167***	-1.311***	-1.367***	-1.365***
	(0.098)	(0.070)	(0.084)	(0.074)	(0.249)	(0.268)
Observations	7,497	7,441	14,910	14,646	125	125
R-squared	0.637	0.79	0.641	0.759	0.839	0.849

(note: standard errors in parentheses)
*** p<0.01, ** p<0.05, * p<0.1

Table 13: OLS regression results by student gender

By gender	Male		Female	
VARIABLES	math	reading	math	reading
small	**0.167*** **	**0.119*** **	**0.136*** **	**0.131*** **
	(0.025)	(0.019)	(0.024)	(0.019)
rookie	**0.014**	**-0.011**	**0.049**	**0.010**
	(0.036)	(0.029)	(0.037)	(0.025)
small*rookie	**-0.108***	**-0.073**	**-0.183*** **	**-0.172*** **
	(0.058)	(0.044)	(0.057)	(0.043)
non-white student	-0.252***	-0.117***	-0.189***	-0.0775***
	(0.029)	(0.024)	(0.030)	(0.025)
free lunch	-0.263***	-0.247***	-0.291***	-0.261***
	(0.018)	(0.015)	(0.018)	(0.015)
male teacher	0.015	-0.006	0.019	0.013
	(0.076)	(0.061)	(0.057)	(0.048)
non-white teacher	0.029	0.015	0.030	0.029
	(0.031)	(0.022)	(0.032)	(0.022)
teacher grad.	0.003	0.005	0.0392*	0.017
	(0.020)	(0.014)	(0.020)	(0.015)
grade 1	0.740***	1.015***	0.616***	1.068***
	(0.025)	(0.019)	(0.026)	(0.020)
grade 2	1.487***	1.799***	1.372***	1.833***
	(0.027)	(0.017)	(0.028)	(0.019)
grade 3	2.052***	2.196***	1.933***	2.206***
	(0.027)	(0.017)	(0.027)	(0.018)
constant	-1.169***	-1.297***	-1.141***	-1.340***
	(0.090)	(0.057)	(0.091)	(0.082)
Observations	11,668	11,490	10,864	10,722
R-squared	0.652	0.763	0.645	0.774

(note: standard errors in parentheses)
*** p<0.01, ** p<0.05, * p<0.1

Table 14: OLS regression results by free lunch

By free lunch status	Free lunch		No free lunch	
VARIABLES	math	reading	math	reading
small	**0.158*** **	**0.135*** **	**0.147*** **	**0.115*** **
	(0.027)	(0.020)	(0.026)	(0.019)
rookie	**0.027**	**0.000**	**0.041**	**0.001**
	(0.043)	(0.029)	(0.035)	(0.027)
small*rookie	**-0.115***	**-0.0958** **	**-0.189*** **	**-0.155*** **
	(0.064)	(0.047)	(0.060)	(0.045)
male student	-0.0293**	-0.0968***	-0.0345**	-0.112***
	(0.014)	(0.011)	(0.017)	(0.014)
non-white student	-0.207***	-0.0836***	-0.230***	-0.0984***
	(0.028)	(0.022)	(0.032)	(0.028)
male teacher	0.050	0.067	-0.022	-0.052
	(0.075)	(0.061)	(0.064)	(0.050)
non-white teacher	0.0728**	0.0570***	-0.055	-0.030
	(0.033)	(0.022)	(0.036)	(0.024)
teacher grad.	0.031	0.0344**	0.013	-0.009
	(0.025)	(0.017)	(0.020)	(0.014)
grade 1	0.644***	0.890***	0.712***	1.189***
	(0.033)	(0.023)	(0.025)	(0.018)
grade 2	1.407***	1.715***	1.456***	1.909***
	(0.036)	(0.022)	(0.026)	(0.017)
grade 3	1.978***	2.143***	2.007***	2.252***
	(0.036)	(0.021)	(0.026)	(0.016)
constant	-1.352***	-1.386***	-1.291***	-1.495***
	(0.077)	(0.058)	(0.108)	(0.093)
Observations	11,165	10,978	11,367	11,234
R-squared	0.634	0.772	0.637	0.76

(note: standard errors in parentheses)
*** p<0.01, ** p<0.05, * p<0.1

Table 15: OLS regression results by grade

By grade	Kindergarten		Grade 1		Grade 2		Grade 3	
VARIABLES	Math	Reading	Math	Reading	Math	Reading	Math	Reading
small	**0.142*** **	**0.0758*** **	**0.169*** **	**0.174*** **	**0.162*** **	**0.125*** **	**0.106*** **	**0.105*** **
	(0.034)	(0.017)	(0.033)	(0.027)	(0.033)	(0.024)	(0.028)	(0.021)
rookie	**-0.0489**	**-0.0205**	**-0.0670***	**-0.0169**	**0.00944**	**-0.0834** **	**-0.0364**	**0.00577**
	(0.069)	(0.032)	(0.039)	(0.036)	(0.058)	(0.040)	(0.092)	(0.053)
small*rookie	**-0.109**	**-0.0245**	**-0.0448**	**-0.200** **	**-0.191***	**-0.0735**	**-0.0081**	**-0.053**
	(0.097)	(0.056)	(0.088)	(0.083)	(0.100)	(0.072)	(0.106)	(0.070)
male student	-0.0988***	-0.0675***	-0.00409	-0.138***	-0.0161	-0.114***	-0.00399	-0.0949***
	(0.017)	(0.010)	(0.015)	(0.015)	(0.016)	(0.014)	(0.016)	(0.013)
non-white student	-0.246***	-0.0902***	-0.267***	-0.0964***	-0.205***	-0.0914***	-0.117***	-0.0861***
	(0.035)	(0.019)	(0.029)	(0.029)	(0.030)	(0.028)	(0.033)	(0.025)
free lunch	-0.304***	-0.183***	-0.272***	-0.344***	-0.291***	-0.288***	-0.246***	-0.210***
	(0.020)	(0.011)	(0.020)	(0.022)	(0.018)	(0.016)	(0.019)	(0.014)
male teacher			0.179*	0.219***	0.124	0.0101	-0.140**	-0.0872*
			(0.096)	(0.070)	(0.120)	(0.085)	(0.056)	(0.050)
non-white teacher	0.00184	0.00234	0.0902*	0.0840**	-0.0279	-0.028	-0.0555	0.00721
	(0.059)	(0.034)	(0.048)	(0.039)	(0.044)	(0.035)	(0.046)	(0.030)
teacher grad.	-0.0429	-0.00462	0.0145	-0.0115	-0.0231	-0.00348	0.0378	0.0264
	(0.032)	(0.016)	(0.030)	(0.026)	(0.030)	(0.022)	(0.028)	(0.021)
constant	-1.135***	-1.400***	-0.652***	-0.302***	0.408***	0.700***	0.983***	1.029***
	(0.140)	(0.037)	(0.064)	(0.040)	(0.107)	(0.052)	(0.095)	(0.060)
Observations	5,809	5,728	6,079	5,902	5,348	5,354	5,296	5,228
R-squared	0.267	0.263	0.289	0.299	0.263	0.264	0.228	0.201

(note: standard errors in parentheses)

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

Table 16: OLS regression results by grade and gender, kindergarten

By gender and grade	Kindergarten Male		Kindergarten female	
VARIABLES	math	reading	math	reading
small	**0.159*** **	**0.0742*** **	**0.121*** **	**0.0769*** **
	(0.038)	(0.019)	(0.040)	(0.022)
rookie	**-0.088**	**-0.023**	**-0.010**	**-0.018**
	(0.070)	(0.032)	(0.085)	(0.042)
small*rookie	**-0.043**	**0.006**	**-0.197***	**-0.068**
	(0.108)	(0.061)	(0.110)	(0.066)
non-white student	-0.226***	-0.0933***	-0.259***	-0.0860***
	(0.052)	(0.027)	(0.050)	(0.030)
free lunch	-0.261***	-0.162***	-0.349***	-0.203***
	(0.028)	(0.015)	(0.030)	(0.016)
non-white teacher	-0.045	-0.028	0.056	0.038
	(0.062)	(0.037)	(0.075)	(0.039)
teacher grad.	-0.0951***	-0.018	0.007	0.007
	(0.037)	(0.020)	(0.038)	(0.020)
constant	-1.215***	-1.436***	-1.184***	-1.467***
	(0.127)	(0.027)	(0.176)	(0.057)
Observations	2,984	2,941	2,825	2,787
R-squared	0.279	0.27	0.268	0.264

(note: standard errors in parentheses)
*** p<0.01, ** p<0.05, * p<0.1

Table 17: OLS regression results by grade and gender, grade 1

By gender and grade	Grade1 male		Grade1 female	
VARIABLES	math	reading	math	reading
small	**0.185*** **	**0.151*** **	**0.200*** **	**0.150*** **
	(0.037)	(0.030)	(0.034)	(0.037)
rookie	**-0.104** **	**-0.022**	**-0.006**	**-0.025**
	(0.047)	(0.046)	(0.041)	(0.048)
small*rookie	**-0.035**	**-0.211** **	**-0.182** **	**-0.053**
	(0.116)	(0.093)	(0.084)	(0.086)
non-white student	-0.325***	-0.140***	-0.053	-0.197***
	(0.039)	(0.041)	(0.044)	(0.041)
free lunch	-0.255***	-0.330***	-0.356***	-0.286***
	(0.027)	(0.028)	(0.030)	(0.027)
male teacher	0.240**	0.313***	0.077	0.078
	(0.096)	(0.062)	(0.127)	(0.130)
non-white teacher	0.089	0.0930**	0.072	0.0978*
	(0.058)	(0.045)	(0.045)	(0.050)
teacher grad.	-0.010	-0.015	-0.008	0.040
	(0.036)	(0.029)	(0.033)	(0.035)
constant	-0.632***	-0.379***	-0.385***	-0.684***
	(0.109)	(0.048)	(0.099)	(0.040)
Observations	3,149	3,051	2,851	2,930
R-squared	0.297	0.296	0.312	0.301

(note: standard errors in parentheses)
*** p<0.01, ** p<0.05, * p<0.1

Table 18: OLS regression results by grade and gender, grade 2

By gender and grade	Grade2 male		Grade2 female	
VARIABLES	math	reading	math	reading
small	**0.153*** **	**0.123*** **	**0.168*** **	**0.122*** **
	(0.040)	(0.028)	(0.036)	(0.029)
rookie	**-0.009**	**-0.112** **	**0.026**	**-0.057**
	(0.062)	(0.046)	(0.067)	(0.045)
small*rookie	**-0.174***	**0.022**	**-0.203**	**-0.162**
	(0.099)	(0.069)	(0.129)	(0.099)
non-white student	-0.228***	-0.0852**	-0.187***	-0.110***
	(0.042)	(0.039)	(0.044)	(0.040)
free lunch	-0.297***	-0.276***	-0.283***	-0.294***
	(0.026)	(0.025)	(0.026)	(0.023)
male teacher	0.201	0.073	0.027	-0.069
	(0.124)	(0.125)	(0.158)	(0.137)
non-white teacher	-0.044	-0.021	-0.014	-0.034
	(0.051)	(0.040)	(0.047)	(0.037)
teacher grad.	-0.016	-0.008	-0.032	0.001
	(0.034)	(0.027)	(0.036)	(0.028)
constant	0.523***	0.670***	0.274***	0.613***
	(0.157)	(0.068)	(0.053)	(0.086)
Observations	2,785	2,780	2,563	2,574
R-squared	0.267	0.248	0.283	0.292

(note: standard errors in parentheses)
*** p<0.01, ** p<0.05, * p<0.1

Table 19: OLS regression results by grade and gender, grade 3

By gender and grade	Grade3 male		Grade3 female	
VARIABLES	math	reading	math	reading
small	**0.112*** **	**0.0897*** **	**0.0958*** **	**0.119*** **
	(0.034)	(0.025)	(0.033)	(0.025)
rookie	**-0.072**	**-0.047**	**0.012**	**0.073**
	(0.117)	(0.066)	(0.076)	(0.057)
small*rookie	**0.070**	**0.064**	**-0.080**	**-0.175** **
	(0.137)	(0.078)	(0.098)	(0.087)
non-white student	-0.160***	-0.116***	-0.075	-0.054
	(0.038)	(0.030)	(0.047)	(0.038)
free lunch	-0.239***	-0.212***	-0.248***	-0.202***
	(0.025)	(0.020)	(0.025)	(0.020)
male teacher	-0.199***	-0.113**	-0.068	-0.047
	(0.057)	(0.051)	(0.064)	(0.059)
non-white teacher	-0.042	-0.014	-0.075	0.020
	(0.057)	(0.038)	(0.049)	(0.033)
teacher grad.	0.034	0.022	0.051	0.035
	(0.036)	(0.027)	(0.032)	(0.024)
constant	1.017***	0.991***	0.922***	0.948***
	(0.080)	(0.053)	(0.137)	(0.095)
Observations	2,750	2,718	2,546	2,510
R-squared	0.233	0.2	0.257	0.22

(note: standard errors in parentheses).
*** p<0.01, ** p<0.05, * p<0.1.

Table 20: Marginal effects of a small class, math scores

Math	Small class		Rookie teacher		Small class and rookie teacher	
Subsample	Marginal effect	Join F-test p-value	Marginal effect	Join F-test p-value	Marginal effect	Join F-test p-value
All	0.1350	0.0000	-0.0099	0.0150	0.039	0.000
Black	0.1797	0.0000	0.0138	0.0000	0.142	0.000
White	0.1147	0.0000	-0.0260	0.5377	-0.042	0.000
Others	-0.0924	0.7423	0.2136	0.6897	-0.059	0.818
Male	0.1534	0.0000	-0.0151	0.1258	0.073	0.000
Female	0.1129	0.0000	-0.0023	0.0046	0.002	0.000
Free lunch	0.1405	0.0000	-0.0035	0.1747	0.070	0.000
No free lunch	0.1280	0.0000	-0.0127	0.0050	-0.001	0.000
Kindergarten	0.1270	0.0001	-0.0819	0.1042	-0.016	0.000
G1	0.1521	0.0000	-0.0996	0.0755	0.028	0.000
G2	0.0972	0.0000	-0.0614	0.1068	-0.086	0.000
G3	0.1654	0.0002	-0.0791	0.7213	0.057	0.001
K'n male	0.1801	0.0000	-0.1145	0.1539	0.046	0.000
K'n female	0.1428	0.0080	-0.0408	0.0436	0.072	0.006
G1 male	0.1326	0.0000	-0.0402	0.0264	-0.020	0.000
G1 female	0.1241	0.0001	-0.0557	0.5258	-0.030	0.000
G2 male	0.1430	0.0007	-0.0271	0.0752	-0.010	0.001
G2 female	0.1050	0.0000	-0.0385	0.2816	0.062	0.000
G3 male	0.1176	0.0005	-0.0530	0.8291	0.110	0.002
G3 female	0.0895	0.0126	-0.0099	0.5821	0.028	0.033

Joint F-tests on all relevant coefficients being equal to zero.
Shaded test results are not significantly different from zero.

Table 21: Marginal effects of a small class, reading scores

Reading	Small class		Rookie teacher		Small class and rookie teacher	
Subsample	Marginal effect	Join F-test p-value	Marginal effect	Join F-test p-value	Marginal effect	Join F-test p-value
All	0.1100	0.0000	-0.0342	0.0006	0.0039	0.0000
Black	0.1624	0.0000	-0.0304	0.0000	0.0778	0.0000
White	0.0745	0.0000	-0.0324	0.1264	-0.0576	0.0000
Others	0.0813	0.4611	0.3519	0.3514	0.0270	0.5513
Male	0.1099	0.0000	-0.0304	0.0452	0.0358	0.0000
Female	0.1095	0.0000	-0.0376	0.0000	-0.0306	0.0000
Free lunch	0.1205	0.0000	-0.0262	0.0380	0.0387	0.0000
No free lunch	0.0997	0.0000	-0.0431	0.0002	-0.0391	0.0000
Kindergarten	0.0725	0.0000	-0.0279	0.5185	0.0308	0.0000
G1	0.0752	0.0000	-0.0217	0.0087	0.0571	0.0000
G2	0.0688	0.0000	-0.0353	0.0135	-0.0082	0.0000
G3	0.1584	0.0000	-0.0711	0.6659	-0.0429	0.0000
K'n male	0.1220	0.0001	-0.0863	0.7292	-0.0824	0.0001
K'n female	0.1756	0.0022	-0.0614	0.2557	0.0116	0.0030
G1 male	0.1136	0.0000	-0.1027	0.0076	-0.0319	0.0000
G1 female	0.1266	0.0000	-0.1061	0.0555	0.0328	0.0000
G2 male	0.1021	0.0000	-0.0990	0.0261	-0.0970	0.0000
G2 female	0.0988	0.0001	-0.0083	0.0483	0.0578	0.0000
G3 male	0.0948	0.0001	-0.0298	0.7037	0.1071	0.0003
G3 female	0.1055	0.0000	0.0249	0.1267	0.0172	0.0000

Joint F-tests on all relevant coefficients being equal to zero.
Shaded test results are not significantly different from zero.

Table 22: Unconditional quantile regression (math scores)

Decile	1	2	3	4	5	6	7	8	9
small	**0.099*** **	**0.140*** **	**0.164*** **	**0.153*** **	**0.160*** **	**0.171*** **	**0.152*** **	**0.145*** **	**0.135*** **
	(0.020)	(0.018)	(0.018)	(0.017)	(0.017)	(0.017)	(0.017)	(0.019)	(0.023)
rookie	**-0.012**	**0.000**	**-0.003**	**-0.010**	**0.014**	**0.039**	**0.0456***	**0.0753*** **	**0.0783*** **
	(0.033)	(0.031)	(0.030)	(0.028)	(0.026)	(0.024)	(0.024)	(0.025)	(0.028)
small*rookie	**-0.153** **	**-0.197*** **	**-0.132** **	**-0.144*** **	**-0.152*** **	**-0.194*** **	**-0.158*** **	**-0.162*** **	**-0.150*** **
	(0.064)	(0.057)	(0.053)	(0.048)	(0.045)	(0.042)	(0.043)	(0.044)	(0.051)
male student	-0.079***	-0.081***	-0.065***	-0.038***	-0.033**	-0.018	-0.018	0.001	0.030*
	(0.017)	(0.016)	(0.015)	(0.014)	(0.014)	(0.013)	(0.014)	(0.015)	(0.018)
non-white student	-0.200***	-0.210***	-0.269***	-0.263***	-0.259***	-0.266***	-0.261***	-0.249***	-0.184***
	(0.024)	(0.021)	(0.020)	(0.018)	(0.018)	(0.017)	(0.018)	(0.018)	(0.021)
free lunch	-0.264***	-0.315***	-0.310***	-0.316***	-0.296***	-0.278***	-0.285***	-0.288***	-0.293***
	(0.019)	(0.017)	(0.017)	(0.016)	(0.015)	(0.015)	(0.016)	(0.017)	(0.020)
male teacher	0.016	0.025	0.0891***	0.069	0.0979*	0.029	-0.015	-0.113	0.098
	(0.014)	(0.027)	(0.032)	(0.044)	(0.053)	(0.066)	(0.084)	(0.098)	(0.128)
non-white teacher	0.0987***	0.0814***	0.0967***	0.0585***	-0.027	-0.029	-0.060***	-0.068***	-0.104***
	(0.025)	(0.023)	(0.022)	(0.021)	(0.020)	(0.020)	(0.020)	(0.021)	(0.023)
teacher grad.	-0.003	-0.026	-0.016	-0.016	0.007	0.0275*	0.012	0.000	0.031
	(0.017)	(0.016)	(0.016)	(0.015)	(0.015)	(0.015)	(0.015)	(0.016)	(0.020)
grade 1	1.358***	1.367***	1.170***	0.914***	0.631***	0.357***	0.206***	0.123***	0.0694***
	(0.032)	(0.029)	(0.027)	(0.024)	(0.020)	(0.016)	(0.014)	(0.011)	(0.008)
grade 2	1.572***	1.959***	2.169***	2.182***	1.878***	1.425***	1.052***	0.822***	0.541***
	(0.030)	(0.024)	(0.021)	(0.020)	(0.020)	(0.020)	(0.021)	(0.020)	(0.021)
grade 3	1.567***	1.992***	2.330***	2.550***	2.515***	2.283***	2.156***	1.835***	1.561***
	(0.030)	(0.024)	(0.019)	(0.015)	(0.014)	(0.015)	(0.019)	(0.023)	(0.031)
constant	-2.216***	-1.976***	-1.722***	-1.411***	-0.949***	-0.458***	-0.005	0.452***	0.942***
	(0.032)	(0.028)	(0.024)	(0.021)	(0.018)	(0.016)	(0.016)	(0.016)	(0.017)
Observations	22,532	22,532	22,532	22,532	22,532	22,532	22,532	22,532	22,532
R-squared	0.229	0.341	0.421	0.488	0.495	0.458	0.409	0.313	0.197

(note: standard errors in parentheses)
*** p<0.01, ** p<0.05, * p<0.1

Table 23: Unconditional quantile regression (reading scores)

Decile	1	2	3	4	5	6	7	8	9
small	**0.067*** **	**0.091*** **	**0.107*** **	**0.148*** **	**0.132*** **	**0.144*** **	**0.152*** **	**0.154*** **	**0.139*** **
	(0.012)	(0.014)	(0.017)	(0.020)	(0.017)	(0.015)	(0.015)	(0.017)	(0.018)
rookie	**-0.016**	**-0.066*** **	**-0.093*** **	**-0.053**	**-0.054***	**-0.006**	**0.026**	**0.045** **	**0.019**
	(0.020)	(0.021)	(0.030)	(0.035)	(0.028)	(0.023)	(0.022)	(0.022)	(0.022)
small*rookie	**-0.0954** **	**-0.0953** **	**-0.113** **	**-0.176*** **	**-0.164*** **	**-0.158*** **	**-0.158*** **	**-0.161*** **	**-0.154*** **
	(0.040)	(0.040)	(0.055)	(0.059)	(0.049)	(0.041)	(0.039)	(0.039)	(0.039)
male student	-0.0688***	-0.0810***	-0.110***	-0.166***	-0.137***	-0.125***	-0.111***	-0.109***	-0.106***
	(0.011)	(0.011)	(0.014)	(0.017)	(0.014)	(0.013)	(0.013)	(0.013)	(0.014)
non-white student	-0.091***	-0.091***	-0.124***	-0.213***	-0.253***	-0.272***	-0.262***	-0.245***	-0.178***
	(0.014)	(0.014)	(0.019)	(0.023)	(0.019)	(0.016)	(0.016)	(0.016)	(0.016)
free lunch	-0.166***	-0.203***	-0.289***	-0.393***	-0.349***	-0.306***	-0.301***	-0.295***	-0.276***
	(0.012)	(0.012)	(0.016)	(0.019)	(0.016)	(0.014)	(0.014)	(0.015)	(0.016)
male teacher	0.002	0.0177*	0.0606**	0.074	0.058	-0.062	-0.045	-0.028	-0.098
	(0.008)	(0.010)	(0.024)	(0.059)	(0.063)	(0.069)	(0.073)	(0.084)	(0.091)
non-white teacher	0.061***	0.035**	-0.017	0.008	-0.017	-0.051***	-0.074***	-0.085***	-0.065***
	(0.015)	(0.016)	(0.022)	(0.025)	(0.020)	(0.018)	(0.018)	(0.018)	(0.018)
teacher grad.	-0.008	-0.014	0.009	0.015	-0.004	0.009	0.010	-0.013	-0.022
	(0.011)	(0.012)	(0.015)	(0.018)	(0.015)	(0.014)	(0.014)	(0.014)	(0.015)
grade 1	1.131***	1.959***	2.842***	2.161***	1.006***	0.567***	0.307***	0.218***	0.125***
	(0.020)	(0.021)	(0.027)	(0.030)	(0.020)	(0.014)	(0.011)	(0.010)	(0.009)
grade 2	1.152***	2.176***	3.807***	4.119***	2.685***	1.624***	1.154***	0.798***	0.465***
	(0.020)	(0.019)	(0.016)	(0.017)	(0.018)	(0.018)	(0.018)	(0.017)	(0.017)
grade 3	1.147***	2.169***	3.803***	4.334***	3.272***	2.396***	1.940***	1.480***	1.079***
	(0.020)	(0.019)	(0.016)	(0.011)	(0.010)	(0.013)	(0.016)	(0.020)	(0.023)
constant	-2.097***	-2.483***	-3.070***	-2.528***	-1.236***	-0.383***	0.134***	0.634***	1.115***
	(0.021)	(0.023)	(0.022)	(0.020)	(0.015)	(0.013)	(0.013)	(0.013)	(0.014)
Observations	22,212	22,212	22,212	22,212	22,212	22,212	22,212	22,212	22,212
R-squared	0.301	0.555	0.684	0.660	0.615	0.510	0.420	0.285	0.165

(note: standard errors in parentheses)
*** p<0.01, ** p<0.05, * p<0.1

Class size and teacher experience effects

Class size effect

In all cases, for the entire sample and for each subsample, the linear coefficient estimate for small classes is positive, as is the marginal effect of a small class size, confirming Mueller's (2013) hypothesis. A smaller class size appears to have a positive effect on student test outcomes. This is not the case for the subsample of students who are neither White nor Black. However, this subsample is very small in comparison and the coefficient estimates are not individually significantly different from zero, and the null hypothesis that the two coefficients dependent on the *SMALL* variable are jointly equal to zero cannot be rejected.

The class-size effect is particularly large for males, black students and students who receive a free lunch in comparison with their counterparts, and is more pronounced in the first and second grades. The larger beneficial marginal effect of class size reductions on black students and those who receive a free lunch is in line with results reported in the literature, confirming that class size reductions can allow students who are disadvantaged or more likely to be disengaged in the classroom (which the literature concurs is often the case for students from lower-income families and for African American students) to bridge the gap between their test results and those of other students. The larger class effect in the first and second grade can likely be linked to these years marking the beginning of more serious and structured instruction in both reading and mathematics. The marginal effect in kindergarten is not negligible either. Particularly in mathematics, being in a small class would raise average test scores by approximately one-seventh of a standard deviation.

The unconditional quantile regression results imply that students who are lower on the achievement distribution benefit less from a small class than their peers with higher academic achievement in reading. In mathematics, those who are in the middle of the distribution show the largest positive effect of being in a smaller class. As such in reading small classes might increase the achievement gap between low and high achieving students, while in mathematics, more students would achieve higher academic outcomes, while students already scoring low would fall even further behind.

Teacher experience effect

Over the entirety of the sample, the linear coefficient estimates for a rookie teacher compares to the estimates obtained by Mueller in that the linear coefficient estimates are not significantly different from zero. The interesting conclusion from this is that in a regular class, there is no statistically significant teacher experience effect. This implies that gains from teacher experience in smaller classes (see next sub-section) are likely not a result of learning from experience, as this would likely translate over to regular classes, but rather a gain from higher quality of teaching as Mueller states. The marginal effect, however, of a rookie teacher, controlling for class size is significantly different from zero, negative, though very small.

The important difference between the marginal effect and coefficient estimate on the *ROOKIE* dummy variable is that the former measures the effect of a rookie teacher not conditional on class size while the latter, alone, looks at the effect of a rookie teacher in a regular-size class.

The marginal effect of a rookie teacher is significantly different from zero in more cases than the linear coefficient estimate, implying that the difference in the effect of rookie teachers compared to senior teachers is most likely more important in small classes than in regular classes.

The marginal effect of a rookie teacher on test scores in the STAR sample is in almost all cases negative (when it is significantly different from zero) except for black students in reading, but quite small in terms of standard deviations on student test scores. Those subsamples that are most responsive (negatively) to teacher experience compared to their counterparts are white students, students not receiving a free lunch, and, to a small extent, female students. These are students who tend to be ahead in scores on average. This is particularly interesting because it means that students who are most likely to find themselves being taught by rookie teachers following a class-size reduction policy implementation are less negatively affected by less experienced teachers.

The other interesting exception to the negative or null effect of rookie teachers is in the top achieving quantiles in mathematics scores. It appears that rookie teachers have a very small positive effect on these high achieving students unconditional of class type.

Teacher experience and the class size effects together

So, the effect of having a rookie teacher is likely to either have no effect or a very small negative effect on students on average. Having a rookie teacher in a small class is likely to have a smaller marginal effect on students than having a senior teacher in a small

class. By examining the whole sample, Mueller (2013) concludes that experienced teachers are the only ones who capitalize on the class size effect, who generate the largest class size effect. This is confirmed by looking at the marginal effect of a small class with a rookie teacher (the sum of the linear coefficient estimates for a small class, a rookie teacher, and for a rookie teacher in a small class) is very close to zero. However, this result varies across subsamples.

The marginal effect of a rookie teacher in a small class is consistently smaller than the marginal effect of a small class, without consideration for teacher experience, and certainly smaller than the marginal effect of a senior teacher in a small class (calculated by the coefficient on the *SMALL* variable). Of particular interest are the cases in which the marginal effect of a rookie teaching a small class is negative, which is the case, in particular, for white students and students who do not receive a free lunch, and of female students in reading.

However, for certain other groups of students, rookie teachers still generate a significant positive class effect from a class reduction, though in no case do they have as large a positive effect as a senior teacher in a small class would on average. The marginal effect of being in a small class with a rookie teacher remains positive, though smaller than the linear coefficient for the small class dummy variable. This is the case for African Americans, students who receive a free lunch, and male students.

Interestingly, those subgroups who benefit most from being in a small class, African American students, males, and students who receive a free lunch still reap some benefits from being in a small class, even with a less experienced teacher. This is likely due to the benefit of smaller classes being so large that the teacher's impact is small in

comparison, but requires more theoretical and empirical investigation. Meanwhile those students who are on average less at a disadvantage academically, and who show smaller class size reduction comparative gains, are often better off in a regular class with any teacher than in a small class with a teacher who has fewer than three years of experience. If these students are already as engaged as they, on average, can be, or at the very least more engaged than low-income students, male students, or African American students, perhaps small classes can only have so much of an impact, and teacher experience is more of a maintenance factor of classroom engagement rather than a generating factor.

Tennessee implemented class size reductions following the study in higher-poverty schools scoring low on state tests, and the reductions were found to yield positive effects[13]. No information could be found on the teachers that were allocated to these schools, and it is possible to assume that either experienced teachers were at the head of these classrooms, or that gains could have been even larger had senior teachers been assigned to these classes, since even rookie teachers are found to generate a large class size effect among black students and lower-income students.

Student characteristics

Student characteristics that were controlled for in the estimation of the model were student gender, race, and whether or not the student qualified for a free lunch. Receiving a free lunch means a student's family income was below a certain threshold, and so this variable can be used to identify students from lower income families. The

[13] Project Challenge, (Mosteller 1995)

distribution of students on free lunch was not even between Caucasian students and others. Of those identified as non-White for the purpose of the regressions, approximately 80.5% received a free lunch, while only 35.2% of White students received a free lunch.

The effect of student characteristics is very present in the literature. It is generally acknowledged that boys have lower results than girls, though less so in math and in the sciences. The presence of a racial gap is also well documented. African American students, who make up the large majority of students identified as non-White for the purpose of these regressions, are usually found to have lower test scores than their Caucasian peers. Finally, students from families with lower incomes are known to score lower than their peers from higher-earning families.

These findings are corroborated by the results of the regressions performed here on the STAR experiment's sample. The coefficient estimates on the dummy variables for a non-Caucasian student, male student and a student receiving a free lunch were all negative and significantly different from zero. The coefficient estimates on the male dummy were much smaller when the scores being used for the analysis were math scores rather than reading scores, as the literature suggests.

What is interesting to note is the fact that, since both the coefficient estimates receiving a free lunch and the race dummy are negative and over 80% of non-Caucasian students receive a free lunch compared to approximately 35% of Caucasian students, minority students may be, on average, at a double disadvantage academically. However, in the regression on the different racial subsamples, it can be seen that receiving a free

lunch has a larger negative estimated marginal effect on a White student than on an African American student.

The effect of race is not even for males and females, however. Looking at the regressions separated by gender subgroups, one sees that the non-White student dummy variable yields smaller coefficient estimates in the female subsample than the male subsample, which implies that two girls differing only in race are likely to have much more similar scores than two boys differing only in race. However, receiving a free lunch affects girls more negatively than boys, all else held equal.

Teacher characteristics

The teacher characteristics controlled for in the estimation of the model, collected through the STAR experiment, other than experience, are gender, race, and highest degree obtained.

Male vs. female teacher

The gender of a teacher appears to have little impact on student test results. In all of the regressions the gender dummy variable's coefficient estimate was not significantly different from zero, except in the case of first-grade students (male students in reading and math and both male and female in reading) and third-grade students (males students in reading and math and both male and female in math). Interestingly, the male teacher coefficient is negative for third grade male students but not significantly different from zero for female students. However, in grade 1, the male teacher coefficient is positive for male students. This is difficult to explain, though the very

small number of male teachers in the experiment may have an impact on results. Since no kindergarten teacher in the sample was male, there is no way to know the effect of a male teacher on student outcomes in kindergarten.

Teacher's race

Looking at the subsample of African American students and comparing it to the subsample of Caucasian students, one can see that a non-Caucasian teacher has a positive marginal effect on the test results of African American students, and a negative marginal effect on the test results of Caucasian students. The magnitudes of the effects are almost equal, while the sign is different. This suggests that students respond better to a teacher of their own race. Interestingly, while 44% or so of Black students had a Black teacher, 94% of White students had a White teacher, suggesting that classes are likely to be predominantly Black or White.

In addition, students receiving a free lunch were estimated to benefit from a non-White teacher in both math and reading scores, while students not receiving a free lunch were estimated to have lower math scores if their teacher was not Caucasian. I would assume that the high proportion of black students receiving a free lunch is a likely explanation of this.

Teacher's education

In the estimation of this model a dummy variable is used to control for a teacher having a graduate degree, be it a *Master's*, what was registered as *Master's+*, or a *Doctorate*. Overall, teachers with graduate degrees were estimated to have no effect on student

score that is significantly different from zero. Further observation of the estimation subsamples indicates that students receiving a free lunch in reading are estimated to benefit by a very small amount from having a teacher with a graduate degree. Male kindergarten students were estimated to have lower scores in mathematics with a teacher who'd attained a graduate degree. These findings are in line with the findings in the literature that posit that a teacher's education is a very poor indicator of their quality of instruction.

Section 5: Further Statistical Tests

Heteroskedasticity

As discussed in Section 3, the correlation of error terms with student and class (teacher) is controlled for using the clustering method of Cameron, Gelbach and Miller (2011). If it is further assumed that the assignment has been carried out along the random design as planned, there is no reason to assume that heteroskedasticity should be corrected for in other variables. In addition, with the standardization of the dependent variables, which centred them to a mean of 0 and a standard deviation of 1 and limited it to a range of approximately -4 to3.5, using robust standard errors would be unnecessary.

Randomness of the experiment

A series of regressions can confirm random assignment of students and teachers to the different types of class. In each case, I control for schools since by design the experiment randomized assignment at the school level.

Randomness of class assignment for student

Running a multinomial logit regression of all three class types using student characteristics as explanatory variables helps determine if there was non-randomness in the assignment of students to either a small or regular class with an aide as opposed to a regular class without an aide. This regression is run over all student observations

when they first enter the experiment, then separate regressions are run for those students who enter in each grade. The regression controls for schools with dummy variables as well. Running a simple logit regression model to determine if selection occurred between small classes and regular classes (either with or without an aide) yields very similar results to those of the multinomial logit regression, with F-tests yielding the same conclusions. Coefficient estimates for the multinomial regressions are presented in Table 23 along with p-values for tests of joint significance for all the student characteristics.

It would appear that students were randomly assigned to class types in all grades but the first grade. This is likely due to students leaving after kindergarten disproportionately between small and regular classes and their positions having to be filled by entering students. In any case, the experiment was random by nature, and individual coefficients are not statistically significant at 95% confidence so I assume that this is not overly concerning. Chetty et al. (2011) obtained additional data such as earnings (both the students' future earnings and their parents'), college completion, home ownership, etc. and determined that assignment was random on these additional controls as well.

Table 24: Test for random assignment of students entering the experiment

	All entering students		First observed in k'n		First observed in G1		First observed in G2		First observed in G3	
	small class	Regular +aide	small class	regular+ aide	small class	regular+ aide	small class	regular+ aide	small class	regular+ aide
Non-white	0.053	-0.004	-0.045	0.043	0.350*	-0.236	0.195	0.117	0.296	0.084
	(0.092)	(0.083)	(0.133)	(0.127)	(0.211)	(0.164)	(0.260)	(0.228)	(0.283)	(0.262)
Male	0.019	0.011	0.011	0.028	-0.248*	-0.073	0.134	0.015	0.173	0.178
	(0.048)	(0.044)	(0.064)	(0.062)	(0.131)	(0.100)	(0.142)	(0.121)	(0.149)	(0.134)
Free-lunch	-0.039	0.007	-0.014	0.066	-0.294*	0.010	0.007	-0.208	-0.213	0.087
	(0.058)	(0.053)	(0.078)	(0.075)	(0.157)	(0.122)	(0.171)	(0.150)	(0.171)	(0.164)
Constant	-0.594**	-0.134	-0.149	-0.086	-18.1	-0.065	-0.490	0.002	-2.43**	-0.975*
	(0.252)	(0.203)	(0.350)	(0.335)	(2207)	(0.367)	(0.517)	(0.478)	(1.057)	(0.560)
Obs.	11,690	11,690	6,256	6,256	2,197	2,197	1,724	1,724	1,513	1,513
P-value for test of joint significance	0.9792		0.9039		0.0453		0.6630		0.3882	

(notes: 1. base outcome: regular class without a full-time aide
2. standard errors in parentheses)
*** p<0.01, ** p<0.05, * p<0.1

Random assignment of teachers

Each teacher in the sample taught only one class at one grade level during the experiment. In order to test whether each was assigned randomly to a small, regular or regular with teacher's aide class, a similar multinomial logit regression is run as as that above on the sample of teachers in which each teacher is only observed once rather than on the sample of teachers counted as many times as there are students in their class. The estimation results are presented in Table 24 along with p-values for tests of joint significance for all the teacher characteristics.

The individual coefficient on a rookie teacher is statistically significant, but the null hypothesis that all coefficients are jointly not significant cannot be rejected at 95%

confidence. The small number of teachers per school who participated in the experiment and the experiment being random by design leads me to believe this is sufficient to assume random assignment of teachers.

Table 25: Test of random assignment of teachers

	small class	regular + aide
VARIABLES		
Teacher grad.	-0.155	0.240
	(0.155)	(0.162)
Non-white teacher	0.057	-0.072
	(0.214)	(0.228)
Male teacher	-0.259	-0.454
	(0.635)	(0.702)
Rookie teacher	-0.300	-0.553**
	(0.212)	(0.240)
Constant	0.081	-0.141
	(0.713)	(0.715)
Observations	1,308	1,308
P-value for test of joint significance	0.0627	

(notes: 1. base outcome: regular class without a full-time aide
2. standard errors in parentheses)
*** p<0.01, ** p<0.05, * p<0.1

Randomness of student attrition

Generating a dummy variable for students who leave the sample before the end of the experiment and running logit regressions over the whole sample and again for each grade other than the third grade, I find that there is non-random attrition occurring in all grades. Table 25 presents coefficient estimates and p-values for tests of joint significance.

Table 26: Test of random attrition among students

	All attrition	Post k'n attrition	Post-g1 attrition	Post-g2 attrition
Mathematics score	-0.694***	-0.471***	-0.703***	-0.231***
	(0.040)	(0.069)	(0.083)	(0.084)
Reading score	-0.150***	-0.415***	-0.517***	-0.424***
	(0.038)	(0.132)	(0.080)	(0.100)
Non-white student	-0.356***	-0.551***	-0.399***	-0.153
	(0.068)	(0.133)	(0.121)	(0.133)
Male student	0.049	0.051	0.159**	-0.078
	(0.036)	(0.062)	(0.066)	(0.071)
Free lunch	0.291***	0.228***	0.328***	0.378***
	(0.043)	(0.078)	(0.081)	(0.089)
Small class	-0.126***	-0.116*	-0.0355	-0.0818
	(0.041)	(0.069)	(0.076)	(0.083)
Constant	-1.473***	-1.705***	-1.290***	-1.177***
	(0.167)	(0.334)	(0.283)	(0.348)
Observations	22,143	5,725	5,681	5,285
P-value for test of joint significance	0.000	0.000	0.000	0.000

(notes: standard errors in parentheses)
*** p<0.01, ** p<0.05, * p<0.1

It appears that students with lower grades are significantly more likely to leave the sample, as are non-white students and students who do not receive a free lunch. This has the potential to bias the results of the study, which does not account for attrition. It will not bias results at the kindergarten level, however.

Ehrenberg et al. (2001) and Nye et al. (1999) found that attrition would not have made smaller classes look more beneficial than studies had found them to be since attrition patterns were the same in both small classes and regular classes, and those students who left small classes were on average the higher achieving students.

Ding and Lehrer (2010) conduct a more in-depth analysis of the effect of class size-reductions on student achievement while accounting for non-ignorable attrition and switching. Their analysis questions the durability and persistence of class effects. Their results are more in line with Hanushek (1999), suggesting that class-size reductions are likely most effective in kindergarten and grade 1. I ignore switching in this paper by removing all post-switching observations, though it is likely that a different approach could yield somewhat different results.

The Hawthorne effect

Hanushek (1999), among others, pointed out that the STAR experiment likely led to teachers altering their behaviour in order to make small classes look beneficial, since they knew the intent of STAR was to guide policy-makers. However, Krueger (1999) found no evidence of either Hawthorne or John Henry effects. He made use of the variation in the size of classes classified as regular classes and found that, even though it would have been in the interest of teachers to make those smaller "regular" classes appear less effective, the class-size effect remained significant.

Representativeness of the STAR sample

Additionally, Hanushek (1999) demonstrates that the STAR sample is not representative of the population of Tennessee or of the United States. This is so, and it was done purposefully, as Nye et al. (1999) point out, to represent all possible

conditions present in the United States. Of course this is not entirely the case, since races other than black or white were not very well represented. However, the results cannot be completely ignored on the basis that the sample does not completely represent the population; it is something that must be kept in mind when applying the findings of STAR to policies that affect different populations of students.

Section 6: Conclusion

This paper sought to determine if the data obtained by Project STAR supported the hypothesis that a policy to reduce class size on a large scale would indeed be regressive. However, the data seems to point to the negative effects of rookie teachers being more significant for white students and students who do not receive a free lunch. If Jepsen and Rivkin's (2009) findings are correct that a mass reduction in class size would cause most rookie teachers to be concentrated in low-income or minority schools, then those students attending these schools would fall further behind and the gap documented between minority and majority students, as well as that between low- and high-income students, would widen as those new teachers would not fully maximize the possible gains from class-size reduction as their senior colleagues might. However, if the share of senior and rookie teacher is even between all strata of students, then white students, female students, and students with family income high enough not to qualify for a free lunch would be worse off at first if their new teachers are inexperienced, inferring from the STAR findings presented here. Unless a state can find enough experienced teachers to fill all newly opened positions, or gradually implement the policy and reallocate senior teachers to new smaller classes (since rookie teachers have only a very small negative impacts on regular classes across the board), the policy can be expected to create winners and losers in the short run.

Ignoring the interaction of teacher experience and class size, a policy to reduce class size may be assumed to reduce the test score gaps present between black and white students, as well as between low- and high-income students. If this data can be

assumed to be representative of younger students everywhere, the findings of this paper, in conjunction with the findings of Mueller (2013) and Jepsen and Rivkin (2009) impose the caveat that small classes will not benefit everyone if teacher characteristics and their effects on students are not accounted for.

Bibliography

Achilles, Charles M., Jeremy D. Finn, and Helen P. Bain, 1997. "Using Class Size to Reduce the Equity Gap". Educational Leadership (December 1997/January 1998), 40-43.

Cameron, A. Colin, Jonah B. Gelbach, and Douglas L. Miller, 2011. "Robust Inference with Multiway Clustering". American Statistical Association Journal of Business and Economic Statistics 29 (2), 238-249.

Chetty, Raj, John N. Friedman, Nathaniel Hilger, Emmanuela Saez, Diane Whitmore Schanzenbach, and Danny Yagan, 2011. "How Does your Kindergarten Classroom Affect your Earnings? Evidence from Project STAR". The Quarterly Journal of Economics 126 (4), 1593-1660.

Ding, Weili and Steven F. Lehrer, 2010. "Estimating Treatment Effects from Contaminated Multiperiod Education Experiments: the Dynamic Impacts of Class Size Reduction". Review of Economics and Statistics 92 (1), 31–42.

Ehrenberg, Ronald G., Dominic J. Brewer, Adam Gamoran, and Douglas Willms, 2001. "Class Size and Student Achievement". Psychological Science in the Public Interest 2 (1), 1-30.

Finn, Jeremy D. and Charles M. Achilles, 1999. "Tennessee's Class Size Study: Findings, Implications, Misconceptions". Educational Evaluation and Policy Analysis 21 (2), 97-109.

Firpo, Sergio, Nicole M. Fortin, and Thomas Lemieux, 2009. "Unconditional Quantile Regression". Econometrica 77 (3), 953–973.

Hanushek, Eric 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.

Harris, Douglas N. and Tim R. Sass, 2011. "Teacher Training, Teacher Quality and Student Achievement". Journal of Public Economics 95, 798-812.

Jepsen, Christopher and Steven Rivkin, 2009. "Class Size Reduction and Student Achievement. The Potential Tradeoff between Teacher Quality and Class Size". Journal of Human Resources 44 (1), 223-250.

Krueger, Alan B. and Diane M. Whitmore, 2000. "The Effect of Attending a Small Class in the Early Grades on College-Test Taking and Middle School Test Results: Evidence from Project STAR". The Economic Journal 111, 1-28.

Krueger, Alan B., 1999. “Experimental Estimates of Education Production Functions”. Quarterly Journal of Economics 114 (2), 497-532.

Lazear, Edward P., 2001. “Educational Production”. The Quarterly Journal of Economics 116 (3), 777-803.

Mosteller, Frederick, 1995. “The Tennessee Study of Class Size in the Early School Grades”. The Future of Children 5 (2), 113-127.

Mueller, Steffen, 2013. “Teacher Experience and the Class Size Effect – Experimental Evidence”. Journal of Public Economics 98, 44-52.

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.

Nye, Barbara, Spyros Konstantopoulos, and Larry V. Hedges, 2004. “How Large Are Teacher Effects?”. Educational Evaluation and Policy Analysis 26 (3), 227-257.

Rice, Jennifer King, 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.

Rivkin, Steven G., Eric A. Hanushek, and John F. Kain, 2005. “Teachers, Schools, and Academic Achievement”. Econometrica 73 (2), 417-458.

Rothstein, Jesse, 2010. “Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement”. The Quarterly Journal of Economics, 175-214.

Word, Elizabeth et al., 1990. “The State of Tennessee’s Student/Teacher Achievement Ratio (STAR) Project: Final Summary Report 1985-1990”. http://d64.e2services.net/class/STARsummary.pdf, last accessed (Jul 27, 2014).

Appendix A

(This section follows Cameron et al. 2011)

We can logically expect clustering of errors at the student level over time and between classes (identified by teacher) in the cross-section.

Clustering is important to consider since it can lead to under-estimated standard errors. It is impossible to account for student and teacher fixed effect with dummy variables because of the sheer size of the sample, and de-meaning at the student level is problematic with the amount of attrition and late entry that occurs, and the resulting unbalanced panel data.

Cameron et al. build from one-way clustered standard errors up to multi-way clustered standard errors. In a case where shocks would affect a group as a whole, say if a classroom is undergoing construction, or one teacher is more effective than another for no measurable reason, there is error independence across clusters:

$$E\left[u_{ig}, u_{jg} \middle| x_{ig}, x_{jg'}\right] = 0\,, \qquad \text{unless g=g'}$$

for i≠j, where g is the cluster and *X*'s are corresponding regressors.

Without clustering and with all usual assumptions holding:

$$\widehat{Var}\left(\hat{\beta}\right) = s^2(X'X)^{-1}$$

and s is the estimated standard deviation of the error. In the case of clustering, this expression should be inflated by:

$$\tau \cong 1 + \rho_{xj}\rho_u(\bar{N}_g - 1)$$

where ρ_{xj} measures within-cluster correlation of the x_j and ρ_u is within-cluster error correlation. $\overline{N}_g$ is the average cluster size. The adjusted variance can then be represented as:

$$Var(\hat{\beta}) = (X'X)^{-1}\Omega(X'X)^{-1}$$

with Ω estimated by $\hat{B}$:

$$\hat{B} = \sum_{g=1}^{G} X_g^1 \, \hat{u}\hat{u}' X_g$$

$$= X'(\hat{u}\hat{u}' \cdot S^G)X$$

where S^G is *NxN* with element *ij* =1 if *i* and *j* are in the same cluster g. This keeps only those elements that are in the same cluster in the calculation.

In a situation as in this paper, where errors are correlated in two groups along two dimensions, labelled G and H, in this case student over time and class in the cross-section, it is assumed that for i≠j,

$$E\left[u_{igh}, u_{jg'h'} \middle| x_{igh}, x_{jg'h'}\right] = 0 \text{ , unless g=g' or h=h'}$$

and so the elements of $\hat{u}\hat{u}'$ that are correlated in any direction are kept in the calculation for estimated variance:

$$\widehat{Var}(\hat{\beta}) = (X'X)^{-1}\hat{B}(X'X)^{-1}$$

with:

$$\hat{B} = X'(\hat{u}\hat{u}' \cdot S^{GH})X$$

where S^{GH} is *NxN* with element *ij* =1 if *i* and *j* share either a cluster *g* in dimension G or a cluster *h* dimension H, if i and j ϵ $(g \cup h)$.

Cameron et al. created a Stata ado file that estimates variance of OLS estimators with multi-way clustering, and this is the file used for the estimates in this paper.