Estimating Heterogeneity in the Effect of Small Classes on Educational Achievement

Graham McKee

Faculty Advisors: Katharine Sims and Steven Rivkin

Submitted to the Department of Economics at Amherst College in partial fulfillment of the requirements for the degree of Bachelor of Arts with Honors

May 7, 2009

Acknowledgements

First and foremost I would like to thank Professors Sims and Rivkin for their generous advice and support throughout this process. I could not have finished (or started) this process without them. I would also like to thank Professor Reyes for her guidance in the initial stages of the thesis process and Fabian for his Stata help. Thank you to my fellow economics thesis writers (including Sam Grausz) for keeping me company for the past several months. Thank you to Pan, Sam, and all my friends and family for their support.

Abstract

In this thesis I analyze heterogeneity in the effect of small classes by estimating the small class quantile treatment effects. I use data from Project STAR, a Tennessee random assignment experiment, and address possible biases caused by non-random switching and attrition. I find that small classes have a positive effect on all quantiles of the test score distribution in kindergarten and first grade. Moreover, I find substantial heterogeneity in the effect, which is about two times as large on the upper end of the distribution as on the lower end. I also find mild evidence that the small class effect is larger for non-white and low income students conditional on achievement.

1. Introduction

Making education more efficient and equitable are two key goals of policymakers, as education is seen as a way both to increase output and to provide equal opportunities for economic success. In order to determine how to allocate scarce resources to meet these goals, policymakers since the 1960s have looked to empirical estimates of the effect of these inputs on various measures of achievement.[1]

Among these educational inputs, class size reduction has commonly been seen as an effective way to enhance outcomes, due in part to the conventional wisdom that smaller classes increase resources per student and that this translates into academic success. Given the desirability of school reform and the intuitive appeal of smaller classes, it is no surprise that much research has been done to estimate the class size effect. Starting with the Coleman Report (1966), empirical research has produced mixed results in both sign and magnitude (Krueger 1999). This research was hampered by the likely endogeneity of class size and academic achievement. In an effort to avoid this problem, researchers in Tennessee implemented Project STAR (Student/Teacher Achievement Ratio), an experiment in which students were randomly assigned to small or large classes and their academic performance was tracked.

Analysis of the STAR results, as well as the majority of previous class size research, has focused on estimating the average effect of smaller classes. The average effect gives only limited information about the distribution of effects except in the under the assumption that the effect is the same for all students. This is a strong a priori assumption. If small class size allows teachers to focus more on needy students (as in

[1] As noted in Krueger (1999), economists tend to look at income while education researchers tend to look at test scores.

Betts and Shkolnik (1999)) then the effect of smaller class size would be larger for these students. Similarly, if classes with more lower achieving students are more disruptive and smaller classes reduce opportunities for disruption (as in Lazear (2001)) then lower achieving students might benefit more from class size reductions. Alternatively, disruptive environments might hurt better students more, so reducing disruptions might have a larger benefit for higher performing students.

If the small class effect is varies by achievement, then the average effect does not answer questions about equity because it does not indicate who benefits from the reductions (Heckman and Smith (1997) and Levin (2001)). If policymakers are concerned with these equity questions, then the preferred educational policy depends on the distribution of benefits. For instance, policymakers might be particularly interested in class size's effect on lower achieving students. Class size could have a large average effect while giving little benefit to these students. In this case, the average effect would overstate the extent to which reductions achieve the equity goals of policymakers.

In this thesis, I use data from the STAR experiment to examine the heterogeneous effects of small classes.[2] I do so by estimating the class size effect on various quantiles of the conditional test score distribution. While several researchers[3] have estimated heterogeneous class size effects before, they use non-experimental data that necessitate more complicated and less reliable econometric techniques to avoid selection bias.

I address the aspects of the STAR that could bias the results. Attrition is the most important of these, and I adjust for it using propensity score weights. Contrary to the previous work on heterogeneous effects, I find significant positive effects of class size at

[2] The STAR data are available online at http://www.heros-inc.org/data.htm
[3] Edie and Showalter (1998), Levin (2001), and Ma and Koenker (2006)

each quantile. Further, I find significantly larger effects on the upper tail of the conditional test score distribution and significantly smaller effects on the lower tail.

This thesis proceeds as follows. Section 2 summarizes the previous research on class size in greater detail. Section 3 describes the STAR experiment and data. Section 4 introduces quantile treatment effects and discusses the difficulties in using the STAR to estimate them. Section 5 discusses the econometric model and presents the initial results. Section 6 discusses the interpretation of the results from section 5. Section 7 addresses the issue of sample attrition and presents results from an adjustment for non-random attrition. Section 8 addresses the possibility of differences in the quantile treatment effects based on race, income, and teacher experience. Section 10 concludes.

2. Background

The earliest empirical estimates of the class size effect used ordinary least squares and failed to account for the likely endogeneity of class size and performance. For instance, suppose wealthier or more motivated parents are more likely to enroll their students in schools with smaller classes. The same parents are more likely to encourage their children to work more diligently and learn outside of class. If this is the case, then parental characteristics are positively correlated with both class size and achievement. Since the parental characteristics are unobserved, they were not controlled for in the analyses and thus the estimates were likely biased upwards. Similarly, teacher quality may be determined in part by class size, with better teachers opting to teach smaller classes. This would also bias the results. Though an array of methodologies including

instrumental variables has been used adjust for selection bias, the literature has not generated a consensus on the magnitude (or even the sign) of the small class effect.[4]

When the Tennessee legislature funded the STAR experiment starting in the 1985-86 school year, researchers hoped to avoid the selection bias issue altogether. In the experiment, schools from across the state created small, regular, and regular with TA classes. Kindergarten students were randomly assigned to one of these class types through third grade and their achievement was measured with standardized tests at the end of each year. Teachers were also randomly assigned.

Krueger (1999) uses the data from this experiment to estimate the average effect of smaller classes on test scores. He explores the possible biases introduced by measurement error, non-random switching, and attrition. He finds these potential biases to be small, and accounting for them he finds that a class size reduction from 22 to 15 students raises the average test score by 5.37 percentiles (about .2 standard deviations). Krueger (2001) follows up this analysis by estimating the effect on later test results; he finds a smaller but still positive effect on middle school test performance and on the probability of taking a college entrance exam.

Krueger estimates the "average treatment effect" (ATE) of small classes. This parameter is the difference in average test scores between students in various class sizes. The average effect provides only limited information on the distribution of effects across individuals, except under the assumption that the effect is the same for all students. In the presence of heterogeneous program effects, the ATE does not answer many relevant evaluation questions.

[4] Hanushek (2003) and Krueger (2003) offer competing summaries of the class size literature. Hanushek claims the results are entirely inconclusive, while Krueger contests that a majority of studies show at least positive effects.

One way to address heterogeneity is to compute the ATE for various subgroups defined by observable characteristics such as income and race. This is the approach taken by Krueger in his STAR analyses. While one might expect the effect to differ based on these characteristics, the scenarios mentioned in section 1 suggest that the effect might be heterogeneous with respect to achievement. To the extent that demographic characteristics are correlated with achievement, the subgroup ATEs will give some sense of this heterogeneity. However, this is an indirect and less than ideal method for estimating heterogeneity that is based on achievement.

A more direct way to estimate heterogeneity with respect to achievement is to estimate the quantile treatment effects (QTEs). The QTE is the effect of a program on a given quantile of the achievement distribution. In other words, the QTE for a given quantile is the difference between that quantile of the treatment and control distributions. Just as quantiles complement the mean in describing the shape of a distribution, QTEs complement the ATE in describing a program's effect on the distribution of outcomes.

Ideally, experiments allow for straightforward estimates of the QTEs in the same way they allow for straightforward estimates of the ATE. However, previous estimates of the class size QTEs have not exploited the STAR data. Eide and Showalter (1998), Levin (2001), and Ma and Koenker (2006) use non-experimental data to estimate these quantile effects. Eide and Showalter (1998) use US data and find that the effect is not statistically different from zero at any quantile. However, they treat class size as exogenous and thus encounter the same problems as the earlier average effect research.

Levin (2001) estimates the quantile effects using an instrumental variable approach to get around the endogeneity issue. His instrument is a Dutch Ministry of

Education rule that relates funding for teachers to total school enrollment. He groups classes into six size categories and finds that the small-class effects for 4^{th}, 6^{th}, and 8^{th} grade are mostly insignificant. For math scores in 8^{th} grade, he finds a *negative* effect of smaller classes on the middle of the distribution; moving into the next larger class size category increases the .25 quantile of scores by 1.05 percentiles, the median of scores by 1.52 percentiles, and the .75 quantile of scores by .99 percentiles. For language scores, the effect is only significant for the .25 quantile, with movement into the next larger class size category increasing the .25 quantile of scores by .84 percentiles. Levin also finds that there is a positive and significant "peer effect" that is enhanced by smaller class size. However, he finds that this effect is only strong on the top of each test score distribution.

Ma and Koenker (2006) re-analyse the Dutch data using a methodology designed to measure the quantile effects of marginal class size reductions rather than creating a small number of class size categories. Their results are similar to Levin's.

Since these papers use non-experimental data, they focus on correcting for the problem of endogeneity of class size and educational performance. The STAR experiment provides an opportunity to estimate quantile effects while avoiding the most serious endogeneity issues that are the focus of these methodologies.

3. Data

Project STAR was legislated by the Tennessee government and carried out by researchers from the state's four universities (Tennessee State, Memphis State, the University of Tennessee, and Vanderbilt). In the experiment, students were randomly assigned students to small classes (13-17 students), regular classes (22-25), or regular

classes with teacher assistants from kindergarten through 3rd grade. Since my focus is on class size and since previous research has found that the effect of TAs is insignificant, I will only use the regular and small class data for my analysis. I also restrict my analysis to kindergarteners and first graders who were also in the STAR in kindergarten for reasons discussed in the next section.

Teachers were also randomly assigned to one of the three class types. Seventy-nine schools participated in the experiment. These schools are not a random sample, as they had to meet criteria such as size (large enough to have three classes in grades K-3) and location (the legislation required representation of inner city, urban, suburban, and rural schools). Because of this, inner city schools and minorities were overrepresented.

Achievement was measured by tests administered at the end of each year. Students took Stanford Achievement Tests (SATs) in math, reading, and word skills. I do not discuss the word skills results for two reasons. First, the test is not effective at differentiating students at the high end of the distribution: in first grade, over 10% of students have the highest score possible, which forces the estimates on all .9 quantile estimates to zero. Second, word skills is highly correlated with reading (sample correlation is .91 in both years). Thus the estimates are very similar for the two tests at all quantiles other than the .9th.

For my analysis, I re-scale the test scores in order to compare the effects of small classes in each grade. Standard deviation units are the most straightforward. However, the sample of students is different in each year due to an influx of new students in first grade and attrition following kindergarten. For comparability, then, I the units should be standard deviations of the scores of the same group of students for both years. To do this,

for both grades I scale the scores by subtracting the mean score of the group of students who have scores for both tests in both grades and then divide by the standard deviation of the scores of the same group of students. The units in the analysis (standard deviations of the scores of students who have scores for both years) are now consistent for both grades.

Figure 3.1 shows kernel density estimates of the regular and small class kindergarten and first grade distributions of each test. The small class distribution is higher than the regular class distribution for both tests in both years. This illustrates that small class size has a positive ATE. Since the treatment and control distributions have some differences in shape, we might suspect that the effect differs by quantile.

Table 3.1 presents summary statistics for the treatment and control groups of kindergarteners and first graders who were in STAR in kindergarten. There are no pre-experiment test scores, so it is impossible to determine if the kindergarten groups are similar along this dimension. However, the two kindergarten groups are similar in other characteristics. Only a slightly higher percent (.6%) of regular class students received free school lunch and only a slightly higher percent (1%) of small class students are white. Age is also very similar. The p-values in the right-most column confirm that the treatment and control groups are not statistically significantly different along these dimensions, suggesting that random assignment is a reasonable assumption for kindergarten. The treatment-control differences in test scores are approximately the ATEs.

Two aspects are noticeable about the first grade group. First, there is a high rate of attrition: 500 small class students (26%) and 668 regular class students (30%) leave STAR after kindergarten. Second, the treatment and control groups of these students

look different in observable characteristics both from each other and from the overall kindergarten treatment and control groups. From this, we should be concerned that attrition is non-random and systematically different for small and regular class students. I address this issue in section 6.

4. Treatment effect framework

As previously discussed, I will explore heterogeneity in the small class effect by estimating the quantile treatment effects. The relationship between the QTEs and the ATE is analogous to the relationship between the quantiles and the mean of a distribution. In subsection 4.1, I formally review the average treatment effect and introduce the quantile treatment effects. I also discuss how the ideal experiment identifies both parameters. Subsection 4.2 overviews the differences between the STAR and the ideal experiment and outlines how I will address them.

4.1. Treatment effects and their experimental identification

I define the treatment effects using the potential outcomes notation.[5] Let T denote participation status with T_i=1 if individual i (or, more realistically, i's parents) selects to be in a small class and T_i=0 otherwise. Let $Y_i(1)$ and $Y_i(0)$ denote i's outcomes in a small class and regular class, respectively. In evaluating the effect of small classes on any individual, we want to know $Y_i(1)$-$Y_i(0)$.

The evaluation problem is that we can only observe $Y_i(1)$ or $Y_i(0)$, but not both, for any individual. We can, however, attempt to estimate E[Y(1)|T=1] - E[Y(0)|T=1];

[5] See Djebbari and Smith (2008) for a footnote discussion on attributing the potential outcomes framework.

this is the average treatment effect.[6] The ATE can be estimated by constructing a comparison group of non-participants and comparing the average of the marginal distributions $E[Y(1)|T=1] - E[Y(0)|T=0]$. Due to self-selection into the program, however, T and Y will likely be correlated, so that $E[Y(0)|T=0]$ and $E[Y(0)|T=1]$ are not equal and the results are biased. The difference between the terms is the selection bias.

The quantile treatment effects are similar to the ATE. For a given quantile p, the QTE is defined as $p(Y(1)|\ T=1) - p(Y(0)|\ T=1)$. In the non-experimental setting, $p(Y(0)|\ T=1)$ is unobservable. If one were to use $p(Y(0)|\ T=0)$ in its place, the difference between the two would be the selection bias.

If the selection decision is a function of observable characteristics, then we can construct a group of non-participants such that the selection bias is zero. On the other hand, if selection is a function of unobservable characteristics, as class size is likely to be (see previous sections), then selection bias remains a concern. Ideally, a random-assignment experiment such as the STAR overcomes this issue. Since all individuals in the experiment select to participate, $T_i=1$ for all i. By randomly not treating individuals, the ideal experiment provides marginal distributions of $(Y(0)|T=1)$ and $(Y(1)|T=1)$. The ATE and the QTEs can thus be estimated by comparing the averages and quantiles of these two distributions.

Even with unbiased estimation, the QTEs are difficult to interpret because the relationship between the QTEs and the effect on individuals at each quantile is unclear (Heckman and Smith (1997)). The difference in the quantiles tells us how small class

[6] In previous literature, this effect has instead been labeled the "average treatment effect on the treated" (ATT or ATET), in contrast to the "overall average treatment effect." Since the second is a special case of the first, I will refer to the average treatment effects simply as the ATE and quantile treatment effects as the QTEs.

size affects the shape of the distribution; it does not necessarily tell us the small class effect for an individual at a given quantile of the untreated distribution. For example, a student at the upper tail of the treatment distribution might (counterfactually) be in the middle of the control distribution. Similarly, if the entire distribution shifts upwards as a result of treatment, we cannot conclude that no individual was negatively affected by treatment. If the two distributions overlap, it is possible that an individual at the lower tail of the treatment distribution would have a higher rank in the control distribution in such a way that her score was negatively affected by treatment.

More generally, the marginal distributions Y(0) and Y(1) give us only limited information about the joint distribution (Y(0),Y(1)). The difference between the pth quantiles of the conditional distributions will only yield the effect on a student at that quantile if a student would be in the pth quantile regardless of treatment status; that is, if $Y_i(1)$ is at the same quantile as $Y_i(0)$ for all i. This independence of rank from treatment status is known as rank preservation.

Though rank preservation is a strong assumption, in the case of small classes we might intuitively expect a fairly high dependence between the rank of $Y_i(1)$ and that of $Y_i(0)$ for any individual i. A student at the lower tail of the distribution in regular classes would not likely jump to the top, or even the median, of the distribution in small classes. If rank preservation does hold, then the QTEs are the effects on individuals at a given quantile rather than simply the effect on that quantile of the distribution.

4.2. How the STAR differs from the ideal experiment

So far, this section has discussed how the QTEs are identified with an *ideal* experiment. As with many social experiments, there are several ways in which the STAR

experiment diverges from this ideal. These issues were first addressed in Krueger (1999), and I discuss them in this section.

First, there was re-randomization between regular and regular with TA classes after Kindergarten. Krueger points out that if constancy of peers is an important determinant of achievement, the estimates might be biased after kindergarten. Second, there is the possibility of non-random switching between class types after the initial assignment. This issue appears to be trivial for kindergarten, as Krueger calculates that only .3% of students switched class type between their assignment and the start of the school year. However, there was switching after Kindergarten. 108 students (6.25% of small class students) switched from small to regular classes while 126 (6.30% of regular class students) switched from regular to small classes. Students who switch to regular classes have an average kindergarten reading SAT of -.28 while students who switch to small classes have an average kindergarten reading SAT of -.09, compared to the overall kindergarten average of -.12. This suggests that switching is non-random and thus first grade class size is non-random. Using first grade class type as the independent variable will therefore bias the first grade results. Instead, I will use kindergarten class type as the independent variable for my first grade analysis, thereby estimating the reduced form of the model. Switching will bias the average effect estimate downwards.[7]

A third issue with the STAR experiment is that students join and leave the experiment every year. 26% of students leave after kindergarten, and by 3rd grade 50% of the original kindergarten students have left. These students are replaced with new students who join each year. Since the problems of attrition and entry become

[7] Depending on where in the distribution the students switch, the estimates at each quantile will be biased differently. See section 6 for further discussion.

compounded every year, and since there is evidence that attrition between kindergarten and first grade and between kindergarten and later grades is a different phenomenon, I choose not to estimate the effects for second and third grade.[8] I still estimate the effect in first grade for students who were in STAR in kindergarten in order to see if there is "value added" from additional years of small classes.

A fourth way in which the STAR diverges from an ideal experiment is in its treatment population. The QTEs are estimates of the effects on a particular treatment population (the population for whom T=1). The question is, for which population do we want to know the effects of class size reductions, and for which population does the STAR identify the QTEs? If policymakers (say, in the US) want to reduce class sizes state-wide or nationally, then the parameter of interest is the effect on the overall student population in Tennessee or the US, respectively. Non-experimental analyses, on the other hand, estimate the effect on the population of students who select into small classes. This population likely is not representative of the overall students population; the problem of selection bias arises precisely because the two groups differ $((Y(0)|T=1) \neq (Y(0)))$.

The STAR experiment potentially has a benefit over non-experimental data in that its treatment sample is better representative of the overall student population than is the non-experimental treatment population. This is because school choice is not entirely determined by class size. School characteristics such as teacher quality and peers are probable determinants of school choice. In the non-experimental setting, therefore, the small class treatment population also receives many other "treatments" associated with

[8] This is revealed by a comparison of probit regressions for the probability of leaving immediately after kindergarten against those for the probability of leaving after 1st or 2nd grade. The coefficients on kindergarten test score and small class type differ in size and magnitude.

better schools. STAR schools do not include these other benefits, so parents are not likely to be indifferent between STAR schools and non-experiment schools with small classes. Moreover, since assignment is random, parents know that their child has less than a 50% chance of actually participating in a small class. If the decision to enroll in STAR schools is completely independent of the fact that small classes are introduced, and if these schools are a representative sample, then the treatment sample is representative of the overall population: $Y(0)|T_{STAR} = Y(0)$. In this case, the QTEs from the STAR data would be the *overall* QTEs.

Unfortunately, the STAR experiment does not quite meet these assumptions in three ways. First, it is possible that the potential of small-class enrollment influenced parents' decisions. Second, since kindergarten was optional in Tennessee at the time, the sample I use in my analysis is probably not representative. Third, as we have already seen, the schools are not representative of either the state or of the country. Therefore the estimates QTEs are not quite the overall QTEs and not quite the QTEs on the population that selects into small classes in the non-experimental setting.

5. Estimation and results

5.1. Estimating the QTEs using quantile regression

I estimate the QTEs using the quantile regression method (QRM).[9] Whereas linear regression estimates the conditional mean of the response variable, quantile regression estimates conditional quantiles of the response variable.

[9] See Hao and Naiman (2007) and Koenker (2001) for further information on quantile regression.

For a given quantile τ, the QRM minimizes weighted deviations of dependent variable y_i from the conditional τth quantile ξ, which is a function of covariates x_i. The deviations are weighted by the $\rho_\tau(.)$ operator, which is defined as

$$\rho_\tau(u) = \tau * u \quad \text{if } u>0$$
$$(1-\tau)*u \quad \text{if } u<0$$

The QRM estimates β, the vector of coefficients describing the relationship between x and ξ, the conditional τth quantile of y. Thus it minimizes (over β)

$$\sum \rho_\tau(y_i - \xi(x_i, \beta)).$$

For a given τ, the QRM will estimate the effect that being assigned to a small class has on the τth quantile of the test score distribution. The QRM thus provides estimates of the QTEs. As discussed above, the random assignment experiment should provide the marginal distributions, in which case I can obtain unbiased estimates of the QTE with the straightforward specification

$$Y_i = \beta_0^{(\tau)} + \beta_1^{(\tau)} SMALL_i + \varepsilon_i^{(\tau)},$$

where SMALL is an indicator for small class size in kindergarten and the pth quantile of ε is zero.[10] To capture the effects across the distribution, I estimate the above specification for each τ in {.1, .25, .5, .75, .9}.

I also estimate specifications that control for observable characteristics for two reasons. First, there is the possibility that assignment was not completely random. Though Krueger (1999) shows that both student and teacher assignment are not strongly predicted by any observable characteristics, this is still consistent with a small degree of non-randomness in assignment. Second, even with random assignment, there will be some sample correlation between small class and observable variables. Because of this, I

[10] Since heterogeneous effects require heteroskedastic residuals, I use bootstrapped standard errors for the estimated coefficients.

can increase the precision of the SMALL coefficient estimates by controlling for these observable characteristics. Therefore for each τ I also estimate

$$Y_i = \beta_0^{(\tau)} + \beta_1^{(\tau)} X_i + \beta_2^{(\tau)} SMALL_i + \varepsilon_i^{(\tau)},$$

where X_i is a vector of school, teacher, and demographic characteristics for individual i. My main specifications also include school fixed effects because randomization occurred within schools.

Ignoring attrition for now, the interpretation of the results is the following. The coefficient on SMALL at each quantile is the estimate of the QTE for that quantile. For kindergarten, this will be the one-year effect of small classes on performance. For first grade, this effect will be the cumulative effect of two years of small classes on performance. That is, it will include the effects from the first and second years. To isolate the effect of the second year, I include previous year's test score as a control in one set of regressions. For these regressions, the coefficient on SMALL is the marginal effect of the second year of small classes.

5.2. Results

Table 5.1 displays the estimates of the coefficient on SMALL. The units are standard deviations of the test scores of students who have scores for both years. Table 5.2 shows the p-values from Wald tests that the coefficients on SMALL for the two quantiles in each row are equal.

Table 5.1a presents the kindergarten results. Column 1 shows the results for math score with no controls or school effects, Column 2 the results with controls but no school effects, Column 3 the results with school effects and no other controls, and Column 4 the results with school effects and other controls. Each column also displays the

corresponding OLS estimate of the ATE, which are roughly equal to Krueger's (1999) estimates. In each of these columns, the coefficients are positive and significant for each estimated quantile (the 10th percentile of math scores is the one exception). This suggests that smaller class sizes are beneficial for all kindergarten students in math.

Additionally, with few exceptions, in the first four columns the point estimates and confidence intervals shift upwards as quantile increases. Comparing columns 1 and 2 and columns 3 and 4, the results do not change significantly when controlling for observable student, teacher, and school characteristics. For reference, Table 5.3 shows a summary of the coefficient estimates on the controls from the regression in Column 4 of Table 5.1. The student controls include student gender, race, age, and whether the student received free lunch, repeated kindergarten, or received special education or special instruction.[11] There is also a variable for how many days of the school year the student was present. The teacher controls include indicators for whether the teacher was in her first three years, whether she held an advanced degree, and whether she was black. The school controls include indicators for school location: inner city, suburban, and urban. Rural is the omitted indicator.

Column 3 shows that without these controls (but with school effects), the estimates range from .11 standard deviations for the 10th percentile to .26 standard deviations for the 90th percentile. When the controls are included, the estimates have a similar range from .11 standard deviations for the 10th percentile to .31 standard deviations for the 90th percentile. Table 5.2 shows that in both cases the estimates for bottom three quantiles are significantly different from the top two. This means that the

[11] There are no data on repeat, special education, or special instruction for first grade.

effect of small classes is significantly larger on the top of the distribution than on the middle and bottom.

Figure 5.1 provides a graphical view of the results from the Column 4 specification estimated for each percentile. Small class coefficient estimate is on the y axis and quantile is on the x axis. The grey band is the bootstrapped 90% confidence interval, the dashed line is the OLS estimate, and the dotted lines are the bounds of the 90% confidence interval of the OLS estimate. The upward slope shows that the estimate increases as quantile increases; in other words, the small class effect on math scores is larger on the top of the distribution than on the bottom.

Columns 5 through 8 of Table 5.1a show that small classes have similar effects on the distribution of reading scores. All of the estimates are positive and significant and range from around .10 standard deviations for the 10^{th} percentile to between .2 and .28 standard deviations for the 90^{th} percentile. Without school effects, the only statistically significant difference is between the 10^{th} and 75^{th} percentile estimates with no controls. With school effects, the results are more heterogeneous and significantly different for the bottom two and top two quantiles. As with math, class size has larger effects on the higher end of the distribution and smaller effects on the lower end of the distribution. This is contrary to Levin's (2001) result that small classes have different effects for different tests.

Tables 5.1b and 5.1c show the estimates of the two year and marginal first grade effects, respectively. Again, the estimates are those for the coefficients on small class size in kindergarten. Columns 1 and 5 shows the estimates with no controls, columns 2 and 6 the estimates with controls, columns 3 and 7 the estimates with no controls except

kindergarten test score, and columns 4 and 8 the estimates with controls including kindergarten test score. The estimates in Table 5.1c are mostly positive and somewhat significant, showing that there is "value added" from an extra year of small classes. There is much less heterogeneity in this marginal effect, but given the smaller magnitude, this might be due to the larger relative standard errors of the estimates. The estimates in Table 5.1b show that the heterogeneity persists in the cumulative effect. The similarity in magnitude between the cumulative and kindergarten effects despite the first grade value added suggests that some of the kindergarten effect "wears off" before the end of first grade, though some of this is due to the switching between class types after kindergarten.

Taken together, the estimates in Table 5.1 provide a picture of substantial heterogeneity in the effect of small classes. Though the estimates are not always significantly different at different quantiles, they show an overall trend of a larger effect on the top of the distribution and a smaller effect on the bottom of the distribution; the effect on the top is between two and three times that on the bottom. If rank preservation holds, the results show that small classes are better for better students. The OLS results fail to capture this heterogeneity.

6. Attrition

The STAR design flaws and attrition potentially threaten the internal validity of the results in the early STAR analyses. Krueger (1999) shows that the first several potential problems are minor. Attrition, on the other hand, has not yet been addressed and therefore it presents the most serious threat to the validity of the above results. Krueger (1999) uses a (self-described) crude imputation to test the sensitivity of the

results to attrition. With this imputation, his estimates of the average effect do not differ significantly. However, attrition poses a more complex challenge to the QTE results than to the ATE results. For this reason, I address the issue presently and attempt to do so more thoroughly. Subsection 6.1 outlines the effects of attrition. Subsection 6.2 presents the weighting method used to adjust for attrition and presents adjusted first grade results.

6.1. Effects of non-random attrition

If attrition is random, it will not bias the results. However, Table 3.1 suggests that attrition is not random. There are two different types of non-random attrition that confound the results in different ways.

The first type is nonrandom attrition that is uncorrelated with class size. This kind of attrition does not introduce bias. However, it does change the estimated parameter by changing the population for which the QTEs are estimated. As discussed in section 4.2, the STAR treatment sample is hopefully close to representative of the overall student population. Attrition that is uncorrelated with class size increases the differences between the treatment sample and the overall population, increasing the difficulty of evaluating the estimates' external applicability. Further complicating the interpretation, there is no clear relationship between the QTE on a given quantile of the post-attrition distribution and the QTE on the same quantile of the unobserved no-attrition distribution since the direction and magnitude of the difference will be determined by the distribution of QTEs for the no-attrition distribution. For example, suppose the effects on the 40^{th} and 50^{th} percentiles of the no-attrition distribution are larger than the effect on the 46^{th} percentile. Now suppose students leave from the below the 40^{th} percentile of the treatment and control distributions in such a way that the 46^{th} percentiles of each original

no-attrition distribution become the 40^{th} percentiles of the distributions of remaining students and the original 50^{th} percentiles become the new 46^{th} percentiles. Without adjusting for attrition, the estimate of the 40^{th} percentile effect will be smaller than the no-attrition 40^{th} percentile effect while the estimate of the 46^{th} percentile effect will be larger than the no-attrition 46^{th} percentile effect.

The second type of nonrandom attrition is that which systematically differs by class type. This kind of attrition does bias the estimates for each quantile. As with the first type of nonrandom attrition, the direction of the bias at each quantile is unclear, and the quantile estimates will not necessarily be biased in the same direction or magnitude. If attrition is highly correlated with low test scores and regular class size, as we might infer from Table 3.1, then the magnitude of the bias at different quantiles will differ. In this case it is possible that that attrition is driving the observed heterogeneity in the first grade results.

To test for non-random attrition of both types, I run a probit regression of an indicator for attrition on observed student, teacher, and school characteristics. Table 6.1 shows the significant estimates from this regression. The coefficients reported are the marginal effects of the given variable on the probability of leaving. With class size constant, the coefficients on the other variables describe the first type of nonrandom attrition, while the coefficient on class size describes the second type of nonrandom attrition. As expected, lower test scores significantly predict attrition: a one standard deviation decrease in math score is associated with a 6 percentage point increase in the probability of leaving. Class size also significantly predicts attrition. Small class students are less likely to leave, other factors constant.

These estimates fully characterize attrition (and can thus be used to correct for it) only if attrition is not a function of unobservable characteristics. I argue that this is not an unreasonable assumption. Given random initial assignment, attrition that is the same for both class types is most likely driven by ability or effort, which will be reflected in test score and attendance. Attrition that differs by class type for any reason will be reflected in the coefficient of the small class indicator.

6.2. Correcting for attrition bias

Following the literature dealing with attrition in both the linear regression and the quantile regression framework, I use propensity score weighting to adjust for the potential attrition bias. If selection on observables holds, and if the probit specification is correct, I can use the results from the probit to obtain corrected estimates for the first grade QTEs by assigning inverse probability weights.[12] To do this, for each individual i I predict

$$P_i(attrit|X_i)$$

using the estimates from the probit. I then construct weight

$$w_i = 1/(1-(P_i(attrit|X_i)).$$

For quantile τ, the weighted quantile regression minimizes (over β)

$$\sum w_i \rho_\tau(Y_i - \xi(X_i, \beta)).$$

The idea of weighting is that we want to recover the outcome distribution of the original pre-attrition sample. Non-random attrition biases the results by changing the quantiles of the observed outcome distributions and weighting corrects for this. For instance, suppose systematic attrition from the lower tail of the distribution causes the

[12] This is the method used, for example, in Maitra and Vahid (2006). Other possible selection correction methods include Heckman-type correction.

20^{th} percentile to shift to the 15^{th} percentile. With propensity score weighting, the students at the lower tail who remain in the experiment will receive more weight since they were more likely to leave, thus filling in the distribution at the lower tail and shifting the 15^{th} percentile back towards the 20^{th} percentile.

Table 6.2 displays the results of the weighted attrition regressions. Columns 1 through 4 show the estimates of the small class effect on math scores while columns 5 through 8 show the estimates of the effect on reading scores. Columns 1 and 5 show the un-weighted estimates for comparison. Columns 2 and 6 show the weighted results. Columns 3, 4, 7, and 8 show the un-weighted and weighted results when controlling for kindergarten test score. All of these regressions include controls and school effects.

The weighted results are very similar to the un-weighted results. There is somewhat more heterogeneity in the point estimates in columns 2 and 6 than in columns 1 and 5, but given the standard errors these differences are not significant. These results suggest that attrition is not inflating the effect of class size on any quantile and is not driving the originally-observed heterogeneity in the first grade effects.

7. Variation in the QTEs by subgroup

In this section, I extend the analysis of heterogeneity by estimating the QTEs separately for subgroups defined by observable characteristics. In subsection 7.1, I estimate the QTEs for subgroups defined by race and income to examine the relationship between heterogeneity in the full sample QTEs and heterogeneity in the subsample ATEs. In subsection 7.2, I estimate the effects for experienced and inexperienced teachers to test whether teacher quality causes the small class effect.

7.1. Effects by race and income

As mentioned is Section 2, Krueger (1999) addresses heterogeneity by estimating the ATEs separately for different subgroups defined by demographic characteristics. He finds that the ATEs for non-whites and free lunch recipients are higher than those for whites and non-free lunch recipients, respectively, where free lunch is used as a proxy for income. He observes that these results "suggest that the lower achieving students benefit the most from attending smaller classes" (524). Assuming rank switching is not drastic, the results in sections 5 and 6 do not support this conclusion; to the contrary, they suggest that the lower achieving students benefit the *least* from smaller classes. The larger average effect for non-whites and free lunch recipients despite their lower average achievement suggests that the small class effects might vary with respect to these demographic characteristics *conditional on achievement*.[13] If this is the case, then the QTEs for non-whites need not follow the same patterns as the QTEs for whites. To explore this possibility, I estimate the QTEs separately for white and non-white students and free-lunch and non-free lunch recipients.

Since the quantile regressions are now run for each subgroup separately, the QTEs might differ simply because the quantiles of the subgroup distributions differ. Table 7.1, which displays the quantiles of the regular class kindergarten test distributions by subgroup, shows that the quantiles are not the same. Thus even if estimates in the previous two sections hold for each subgroup, the QTEs for whites and non-whites will differ at any given quantile. Since the non-white distribution is lower than the white

[13] Note that if this is the case then rank preservation will not hold for the overall population, though it might hold within subgroups. For example, if the effect is smaller on white students, then a white student at the 25^{th} percentile of the non-treatment distribution will be at a lower rank in the treatment distribution than a non-white student at the 25^{th} percentile of the non-treatment distribution.

distribution, the estimate at a given quantile should be smaller for non-whites if the effects only vary with respect to achievement.

Table 7.2 shows the estimates for the coefficient on small class size for each subgroup.[14] In Table 7.2a, the first column under each test shows the estimates for whites and the second column shows the estimates for non-whites. The effects on math scores for both groups roughly increase as quantile increases, ranging from .07 to .33 for whites and .07 to .26 for non-whites. For each of the other tests, the non-white estimates are almost all larger than the white estimates. Except on kindergarten reading scores, the pattern of larger effect on higher quantile persists. This pattern suggests that the small class effect increases with respect to achievement for both white and non-white students, though the differences in estimates across quantiles are rarely statistically significant due to the large standard errors corresponding to the reduced sample sizes.

Table 7.2b shows the results for free lunch and non-free lunch recipients. The overall trends are similar to those in Table 7.2a. Given a .42 sample correlation between free lunch receipt and non-white ethnicity in kindergarten, this is not surprising. The main difference is the relatively small variation in the different quantile estimates for free lunch recipients in kindergarten. The estimates for math range from .13 to .19 while the estimates for reading range from .15 to .21. Moreover, there is no trend with respect to quantile. These results suggest that the small class effect on kindergarten scores is similar for all free lunch recipients regardless of rank in the distribution. In first grade, free lunch recipients experience substantially more heterogeneous effects that increase as quantile increases. For both tests, the effect ranges from around .11 for the 10^{th}

[14] The regressions include controls but do not include school effects due to the large demands already put on the data with the smaller sample sizes. However, the estimates are similar with school effects.

percentile to around .40 for the 90th percentile. For non-white students, the familiar positive relationship between effect and quantile holds up for each test in each year. Again, sample size limits the statistical significance of many of these differences.

Comparing the first grade free lunch and non-free lunch recipient estimates, small classes appear to have a larger effect on the latter group. This is also supported by the OLS estimates of the average effect, and the relationship also holds for the average effects for kindergarten reading. Similarly, Table 7.2a suggests that non-white students receive a larger benefit than white students for reading tests in kindergarten and both tests in first grade. These estimates are in line with Krueger's findings that the average effects are larger for non-white and free lunch students despite their lower distribution of achievement.[15]

If rank preservation holds within the subgroups, the results in Table 7.2 provide the following two tentative conclusions. First, lower income and non-white students tend to benefit more from smaller classes. Second, within the subgroups, small classes are better for higher achieving students. In other words, the small class effects differs both with respect to achievement conditional on race or income and with respect to race or income conditional on achievement.

7.2. Teacher quality and the small class effect

The positive and heterogeneous effects of small classes raise two immediate questions. First, *why* do small classes have an impact? Second, why is the impact larger for better students? Lazear (2001) develops a model of classroom disruptions and posits that small classes have positive effects by reducing these disruptions. He predicts that

[15] Krueger's analysis uses averages of the three test scores, so he does not have comparable results for the kindergarten math scores.

small classes are better for more disruptive groups of students. Sections 5 and 6 do not support this theory insofar as disruptive classroom behavior is correlated with lower achievement.[16] Since there are no measures of classroom disruption, I can not test this theory more directly with the STAR data.

Teaching potentially provides an alternative answer to both of these questions. Instruction time per student is the one certain variable small classes augment. In a study on the effect of class size on teacher behavior, Betts and Shkolnik (1999) find that class size reductions cause teachers to tailor their instruction to meet individual needs. If differentiated instruction is one of the mechanisms through which class size has an effect, then teacher quality might be an important determinant of the impact of small class size, since better teachers presumably are more able to tailor instruction to individual needs. If schools with higher achieving students also have better teachers, then the high achieving students might benefit more from small classes as a result of teacher quality.

One variable that might successfully measure teacher quality is teacher experience. The previous regressions control for an indicator that the teacher has zero to two years of experience. I explore potential teacher contributions to small-class effectiveness by estimating the QTEs separately for new and experienced teachers. Table 7.3 displays the results. The first column under each test shows the small class coefficient estimates for students with experienced teachers and the second column shows the coefficient estimates for students with new teachers. The small class effect for new teachers is significantly negative for kindergarten math scores, ranging from -.19 to -.24. For kindergarten reading scores, the effect on the 10^{th} percentile is larger for new

[16] Krueger (2003) hypothesizes that Lazear's model implies that small classes are better for better students. However, this is not necessarily implied by the disruption model since high achieving students might benefit more from reduced disruptions.

teachers, though the estimates for the other quantiles are small and insignificant. In first grade, the point estimates are actually larger for new teachers and within the new teacher subgroup they increase as quantile increases. However, these first grade differences both between subgroups and within the new teacher subgroup are statistically insignificant.

These results do not point to a clear difference in effect between new teachers and experienced teachers. This is not surprising given the small sample size. Also, it is highly unlikely that experience is the only predictor of teacher quality, so the results from this crude analysis should not be taken as proof that teacher quality does not explain heterogeneity in the small class effect.

8. Conclusion

Heterogeneity in class size effects has been overlooked relative to mean effects. This thesis shows that heterogeneity is a key feature of class size reductions; the magnitude of the small class effect in kindergarten and first grade is smaller at the lower end and larger at the upper end of the distribution. Although this result contradicts earlier studies of the heterogeneity in class size effect, the randomized STAR experiment provides straightforward identification of this effect with relatively few assumptions, and the lack of consensus on the heterogeneity should encourage further empirical research.

The pattern of heterogeneity found in this thesis has potentially important policy implications. Higher educational achievement, even in early grades, translates into higher earnings. Krueger (2003) examines this link between test score and income and provides a rough cost-benefit analysis of class size reductions from 22 to 15 students using his STAR estimates. He calculates that the benefits are approximately equal to the

costs. Since the benefits are not evenly distributed, however, the reduction might pass a cost-benefit test for the high end of the distribution while failing one for the low end. Traditionally, one justification of the usefulness of the mean effect has been that the government can transfer these later gains, thus cancelling out any inequality that arises from heterogeneous small class effects (Heckman and Smith (1997)). As Heckman and Smith point out, the assumption that the government can and will transfer these gains later is unrealistic. Because of these distributional effects, heterogeneity in the small class effect is a phenomenon that deserves further analysis and explanation so that policymakers can more fully understand their options for enhancing school efficiency and equity.

Figure 3.1
Kernel densities of math and reading standardized scores for treatment (small class) and control (regular class) groups

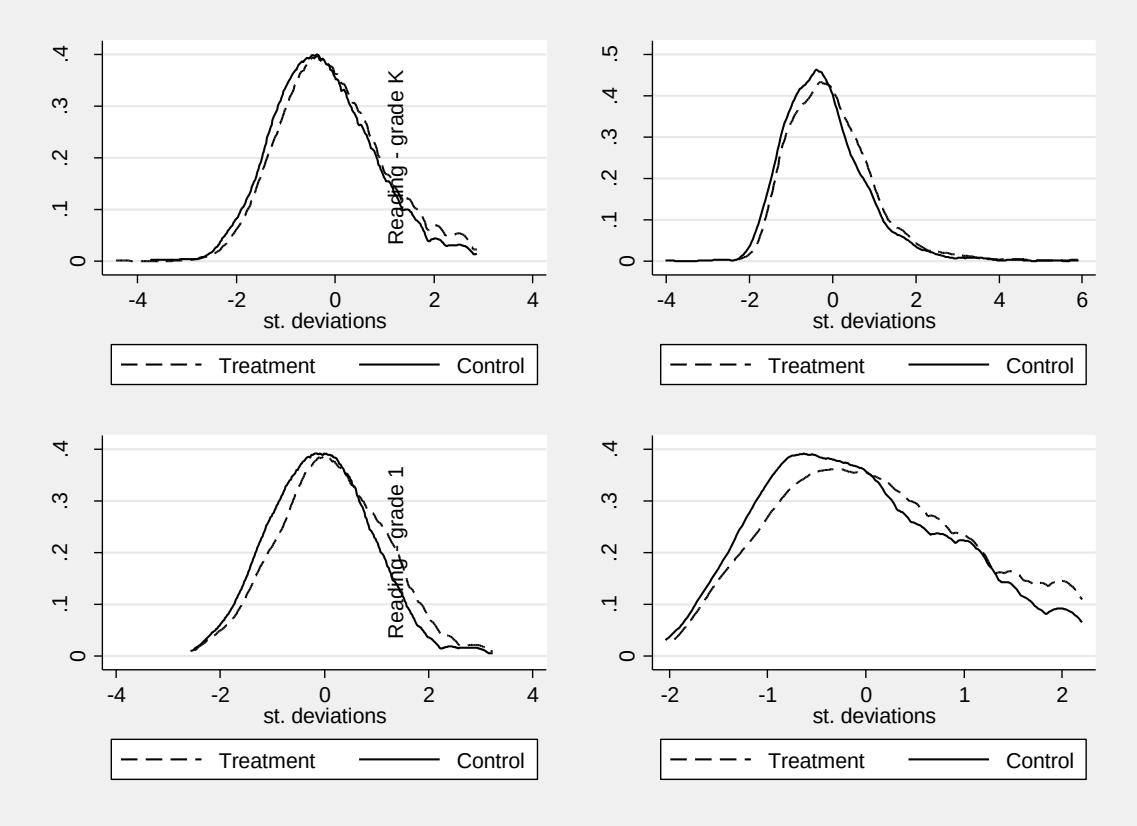


Table 3.2
Summary statistics by class type

	small			regular			pr(|T|>t)
	Obs	Mean	SD	Obs	Mean	SD	
Grade K							
Free lunch	1892	47.1%	-	2187	47.7%	-	0.68
Non-white	1900	31.9%	-	2194	32.9%	-	0.49
Age	1897	5.48	0.35	2190	5.46	0.35	0.17
Class size	1900	15.12	1.50	2194	22.38	2.15	0.00
Math SAT	1762	-0.05	1.07	2032	-0.22	1.03	0.00
Reading SAT	1739	-0.02	1.03	2006	-0.21	0.98	0.00
Grade 1							
Free lunch	1376	47.5%	-	1481	44.4%	-	0.10
Non-white	1400	30.0%	-	1526	28.8%	-	0.51
Age	1400	6.48	0.35	1526	6.47	0.35	0.57
Class size	1400	16.14	2.54	1526	22.44	2.85	0.00
Math SAT	1373	0.15	1.03	1496	-0.06	0.97	0.00
Reading SAT	1342	0.13	1.02	1461	-0.05	0.98	0.00
Grade K attritors							
Free lunch	498	53.8%	-	666	56.7%	-	0.32
White	500	62.8%	-	668	58.0%	-	0.09
Age	497	5.45	0.35	664	5.42	0.36	0.17
Class size	500	15.23	1.41	668	22.65	2.05	0.00
Math SAT	450	-0.43	1.10	600	-0.55	1.05	0.00
Reading SAT	436	-0.32	1.05	590	-0.51	0.94	0.00

The right-most column contains p-values for t-tests of the null hypothesis that the small and regular class averages are equal. The 1st grade summary statistics are for the group of students who were also in STAR in kindergarten.

Table 5.1

Estimates of coefficients on SMALL from quantile regressions by grade

5.1a. Kindergarten								
	Math				Reading			
quantile	1	2	3	4	5	6	7	8
.10	**.108**	**.142**	**.108**	**.109**	**.095**	**.131**	**.095**	**.120**
	(.084)	*(.046)*	*(.047)*	*(.044)*	*(.046)*	*(.034)*	*(.042)*	*(.040)*
.25	**.216**	**.097**	**.108**	**.149**	**.159**	**.150**	**.159**	**.165**
	(.060)	*(.032)*	*(.039)*	*(.042)*	*(.042)*	*(.029)*	*(.036)*	*(.031)*
.50	**.129**	**.156**	**.151**	**.166**	**.159**	**.170**	**.159**	**.201**
	(.067)	*(.038)*	*(.041)*	*(.039)*	*(.038)*	*(.036)*	*(.036)*	*(.031)*
.75	**.151**	**.168**	**.237**	**.272**	**.222**	**.181**	**.222**	**.189**
	(.041)	*(.058)*	*(.043)*	*(.042)*	*(.054)*	*(.048)*	*(.044)*	*(.039)*
.90	**.259**	**.270**	**.259**	**.307**	**.191**	**.210**	**.286**	**.287**
	(.082)	*(.081)*	*(.066)*	*(.062)*	*(.088)*	*(.067)*	*(.076)*	*(.064)*
OLS	**.167**	**.161**	**.191**	**.200**	**.185**	**.173**	**.211**	**.217**
	(.034)	*(.033)*	*(.031)*	*(.030)*	*(.033)*	*(.031)*	*(.030)*	*(.029)*
School Effects	no	no	yes	yes	no	no	yes	yes
Controls	no	yes	no	yes	no	yes	no	yes
N	3794	3779	3794	3779	3745	3730	3745	3730

5.1b. 1st grade								
	Math				Reading			
quantile	1	2	3	4	5	6	7	8
.10	**.091**	**.073**	**.091**	**.158**	**.142**	**.092**	**.142**	**.115**
	(.070)	*(.051)*	*(.052)*	*(.052)*	*(.047)*	*(.041)*	*(.045)*	*(.040)*
.25	**.228**	**.158**	**.160**	**.162**	**.213**	**.183**	**.195**	**.149**
	(.051)	*(.043)*	*(.039)*	*(.044)*	*(.038)*	*(.034)*	*(.047)*	*(.044)*
.50	**.205**	**.174**	**.205**	**.160**	**.284**	**.239**	**.213**	**.217**
	(.060)	*(.047)*	*(.046)*	*(.044)*	*(.057)*	*(.051)*	*(.044)*	*(.044)*
.75	**.342**	**.287**	**.274**	**.232**	**.373**	**.302**	**.337**	**.279**
	(.072)	*(.046)*	*(.046)*	*(.042)*	*(.081)*	*(.057)*	*(.061)*	*(.047)*
.90	**.388**	**.295**	**.320**	**.265**	**.444**	**.266**	**.390**	**.280**
	(.077)	*(.075)*	*(.063)*	*(.064)*	*(.137)*	*(.072)*	*(.088)*	*(.068)*
OLS	**.246**	**.206**	**.215**	**.187**	**.245**	**.225**	**.226**	**.213**
	(.036)	*(.034)*	*(.033)*	*(.032)*	*(.037)*	*(.034)*	*(.033)*	*(.033)*
School Effects	no	no	yes	yes	no	no	yes	yes
Controls	no	yes	no	yes	no	yes	no	yes
N	3044	3001	3044	3001	2979	2934	2979	2934

5.1c. 1st grade controlling for grade K score								
	Math				Reading			
quantile	1	2	3	4	5	6	7	8
.10	**.103**	**.018**	**.070**	**.083**	**.124**	**.053**	**.090**	**.059**
	(.045)	*(.044)*	*(.050)*	*(.049)*	*(.042)*	*(.044)*	*(.041)*	*(.038)*
.25	**.108**	**.111**	**.097**	**.061**	**.116**	**.101**	**.077**	**.085**
	(.053)	*(.037)*	*(.036)*	*(.036)*	*(.039)*	*(.035)*	*(.032)*	*(.033)*
.50	**.130**	**.124**	**.093**	**.061**	**.130**	**.126**	**.069**	**.051**
	(.037)	*(.035)*	*(.036)*	*(.033)*	*(.037)*	*(.042)*	*(.036)*	*(.044)*
.75	**.160**	**.152**	**.095**	**.057**	**.181**	**.180**	**.164**	**.133**
	(.040)	*(.042)*	*(.043)*	*(.039)*	*(.041)*	*(.053)*	*(.041)*	*(.049)*
.90	**.169**	**.136**	**.095**	**.101**	**.198**	**.199**	**.215**	**.207**
	(.071)	*(.053)*	*(.054)*	*(.052)*	*(.075)*	*(.064)*	*(.059)*	*(.063)*
OLS	**.136**	**.114**	**.087**	**.068**	**.160**	**.144**	**.127**	**.117**
	(.031)	*(.030)*	*(.028)*	*(.028)*	*(.031)*	*(.029)*	*(.028)*	*(.028)*
School Effects	no	no	yes	yes	no	no	yes	yes
Controls	no	yes	no	yes	no	yes	no	yes
N	2857	2817	2857	2817	2758	2716	2758	2716

Dependent variable: score for test type given in first row. Bootstrapped SEs in parentheses. Controls include all available individual, teacher, and school characteristics. The regressions with "school effects" include indicators for each school.

Table 5.2
P-values from tests for differences in SMALL coefficients for the two quantiles in the left-most columns
Numbered columns correspond to columns in Table 5.1

5.2a. Kindergarten									
		Math				Reading			
quantiles		1	2	3	4	5	6	7	8
.10	.25	.263	.312	1.000	.328	.216	.528	.105*	.237
.10	.50	.842	.800	.450	.244	.215	.305	.190	.080**
.10	.75	.627	.689	.028**	.003**	.048**	.338	.017**	.214
.10	.90	.159	.152	.054*	.008**	.304	.305	.013**	.017**
.25	.50	.251	.091*	.401	.650	1.000	.496	1.000	.265
.25	.75	.378	.177	.020**	.013**	.300	.495	.156	.594
.25	.90	.666	.024**	.029**	.013**	.713	.376	.095*	.048**
.50	.75	.774	.802	.057*	.007**	.191	.788	.119*	.737
.50	.90	.178	.140*	.118*	.025**	.713	.550	.100*	.195
.75	.90	.192	.137*	.732	.529	.684	.619	.359	.066**

5.2b. 1st grade									
		Math				Reading			
quantiles		1	2	3	4	5	6	7	8
.10	.25	.030**	.052*	.153	.937	.122*	.012**	.250	.396
.10	.50	.146*	.087*	.056*	.968	.036**	.011**	.187	.051*
.10	.75	.009**	.001**	.003**	.201	.009**	.000**	.007**	.004**
.10	.90	.007**	.017**	.001**	.201	.030**	.011**	.017**	.027**
.25	.50	.707	.708	.297	.973	.171	.228	.691	.151
.25	.75	.163	.011**	.014**	.196	.054*	.027**	.026**	.017**
.25	.90	.090*	.089*	.007**	.185	.102*	.235	.034**	.078*
.50	.75	.092*	.006**	.116	.121*	.227	.141*	.022**	.228
.50	.90	.051*	.088*	.080**	.149*	.257	.702	.035**	.384
.75	.90	.592	.890	.440	.581	.588	.569	.490	.992

5.2c. 1st grade controlling for grade K score									
		Math				Reading			
quantiles		1	2	3	4	5	6	7	8
.10	.25	.916	.029**	.479	.554	.818	.137	.674	.486
.10	.50	.601	.035**	.654	.661	.902	.153	.626	.856
.10	.75	.323	.021**	.674	.630	.322	.035**	.185	.193
.10	.90	.433	.066*	.712	.785	.371	.045**	.074*	.038**
.25	.50	.637	.725	.900	.994	.710	.496	.834	.408
.25	.75	.321	.397	.965	.929	.183	.117*	.067*	.350
.25	.90	.470	.674	.970	.465	.309	.135*	.033**	.070*
.50	.75	.386	.442	.952	.921	.184	.193	.017**	.054*
.50	.90	.563	.832	.960	.424	.319	.229	.021**	.013**
.75	.90	.871	.713	.999	.326	.776	.728	.349	.145*

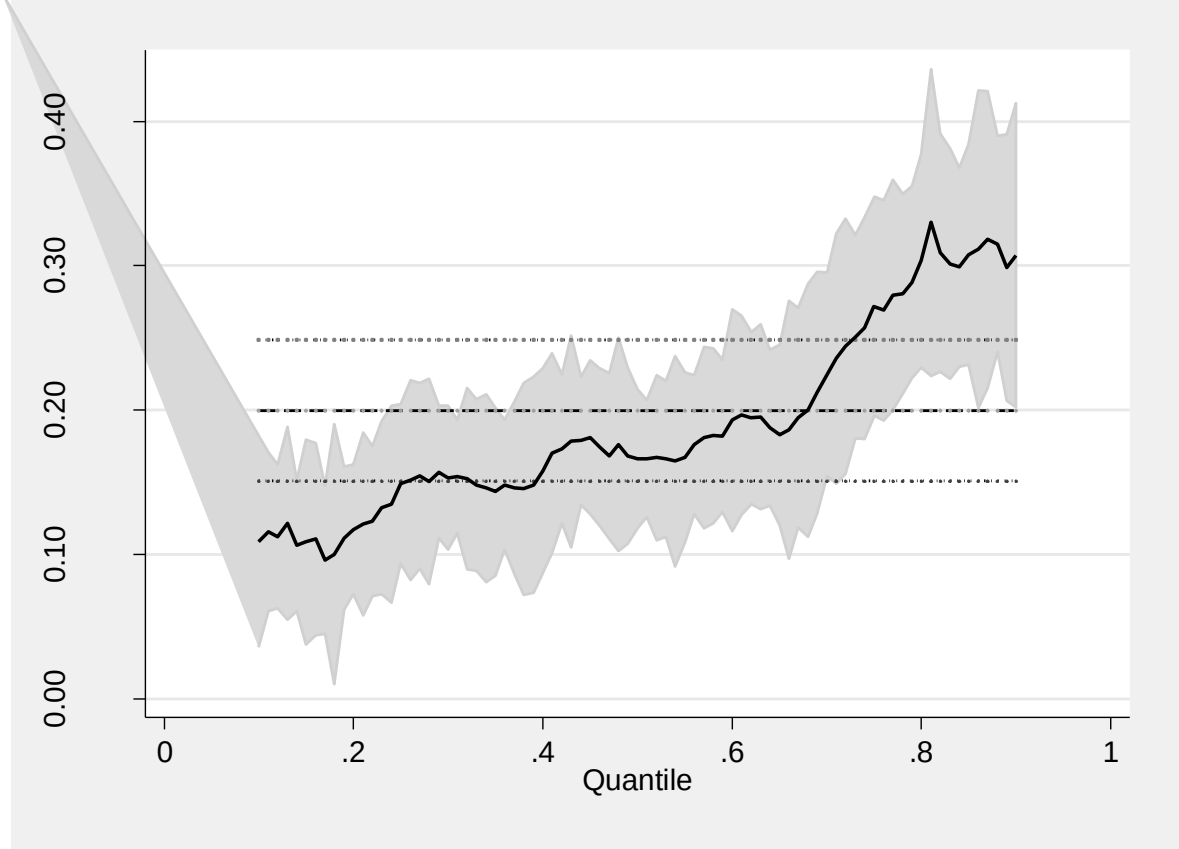
P-values are for tests of the null hypothesis that the coefficients on SMALL for the given quantiles in each row are equal. * indicates significance at the 15% level. ** indicates significance at the 5% level.

Table 5.3
Results from quantile regressions in Column 4 of Table 5.1a
Dependent variable: kindergarten math score

	quantile				
	.10	.25	.50	.75	.90
Small class	**.109**	**.149**	**.166**	**.272**	**.307**
	(.044)	*(.042)*	*(.039)*	*(.042)*	*(.062)*
Female	**.080**	**.142**	**.163**	**.163**	**.111**
	(.036)	*(.031)*	*(.036)*	*(.044)*	*(.054)*
Non-white	**-.266**	**-.287**	**-.339**	**-.436**	**-.494**
	(.088)	*(.077)*	*(.062)*	*(.076)*	*(.130)*
Age	**.222**	**.396**	**.484**	**.441**	**.399**
	(.078)	*(.071)*	*(.060)*	*(.074)*	*(.107)*
Free lunch recipient	**-.279**	**-.422**	**-.409**	**-.433**	**-.412**
	(.039)	*(.044)*	*(.043)*	*(.044)*	*(.075)*
Repeat	**.011**	**.287**	**.379**	**.368**	**.342**
	(.124)	*(.105)*	*(.111)*	*(.124)*	*(.194)*
Special Ed	**-.333**	**-.436**	**-.235**	**-.382**	**-.262**
	(.150)	*(.119)*	*(.113)*	*(.142)*	*(.208)*
Special instruction	**-.670**	**-.500**	**-.503**	**-.395**	**-.312**
	(.137)	*(.088)*	*(.113)*	*(.121)*	*(.203)*
Days present	**.003**	**.003**	**.003**	**.004**	**.004**
	(.001)	*(.001)*	*(.001)*	*(.001)*	*(.001)*
Teacher new	**.042**	**-.086**	**-.117**	**-.188**	**-.218**
	(.075)	*(.074)*	*(.072)*	*(.068)*	*(.095)*
Teacher adv. degree	**.014**	**-.027**	**-.046**	**-.078**	**-.119**
	(.067)	*(.057)*	*(.054)*	*(.066)*	*(.084)*
Teacher black	**.059**	**.064**	**.126**	**-.436**	**.183**
	(.093)	*(.088)*	*(.078)*	*(.076)*	*(.133)*
Inner city school	**-.428**	**-.272**	**-.143**	**-.248**	**.420**
	(.353)	*(.258)*	*(.319)*	*(.378)*	*(.635)*
Suburban school	**-.585**	**-.380**	**-.073**	**.068**	**.644**
	(.559)	*(.257)*	*(.366)*	*(.352)*	*(.720)*
Urban school	**.018**	**-.075**	**.458**	**.093**	**.526**
	(.363)	*(.246)*	*(.382)*	*(.390)*	*(.568)*
Pseudo R^2	.180	.180	.188	.205	.214

Bootstrapped standard errors in parentheses. These regressions included constant terms

Figure 5.1
Coefficients on SMALL from quantile regressions of kindergarten math SAT score on SMALL and controls (including school effects)



The black line is the point estimate and the grey band is the 90% confidence interval. The dashed line is the OLS estimate and the dotted lines are the bounds of the OLS 90% confidence interval.

Table 6.1
Significant predictors of post-K attrition

Math Score	**-.060**
	(.011)
Reading Score	**-.034**
	(.012)
Small Class	**-.038**
	(.015)
Non-white	**-.107**
	(.023)
Free Lunch Recipient	**.045**
	(.017)
Days Present	**-.002**
	(.000)
Inner City School	**.258**
	(.036)
Suburban School	**.263**
	(.024)
Urban School	**.125**
	(.032)
N	**3728**

The right column shows coefficient estimates from a probit regression where the dependent variable is an indicator for leaving STAR after grade K. Marginal effects are reported. Standard errors in parentheses. Regression included the same variables as those in Table 5.3

Table 6.2
Coefficients on SMALL for weighted first grade quantile regressions

	Math				Reading			
quantile	1	2	3	4	5	6	7	8
.10	**.158**	**.165**	**.083**	**.075**	**.115**	**.117**	**.059**	**.058**
	(.052)	*(.019)*	*(.034)*	*(.046)*	*(.036)*	*(.013)*	*(.031)*	*(.015)*
.25	**.162**	**.149**	**.061**	**.077**	**.149**	**.162**	**.085**	**.075**
	(.036)	*(.026)*	*(.044)*	*(.022)*	*(.050)*	*(.012)*	*(.024)*	*(.028)*
.50	**.160**	**.167**	**.061**	**.073**	**.217**	**.211**	**.051**	**.062**
	(.041)	*(.029)*	*(.030)*	*(.020)*	*(.041)*	*(.038)*	*(.025)*	*(.018)*
.75	**.232**	**.232**	**.057**	**.069**	**.279**	**.273**	**.133**	**.140**
	(.037)	*(.037)*	*(.051)*	*(.026)*	*(.061)*	*(.044)*	*(.066)*	*(.033)*
.90	**.265**	**.288**	**.101**	**.093**	**.280**	**.310**	**.207**	**.201**
	(.050)	*(.038)*	*(.034)*	*(.040)*	*(.046)*	*(.076)*	*(.054)*	*(.048)*
w/ Prev Year	no	no	yes	yes	no	no	yes	yes
weighted	no	yes	no	yes	no	yes	no	yes
N	3001	2780	2817	2780	2934	2714	2716	2714

Dependent variable: SAT score for subject in first row. "w/ Prev Year" regressions control for grade K test scores. Standard errors in parentheses. Note that the standard errors should be estmiated with bootstrap, at least in the second stage and ideally in a way to account for uncertainty in the first stage probit estimates. Due to time constraints, I do not do this.

Table 7.1

Quantiles of the grade K regular class SAT distributions by subgroup

	Math			**Reading**		
q	Non-free lunch	free lunch	difference	Non-free lunch	free lunch	difference
.10	-1.17	-1.63	0.45	-1.06	-1.47	0.41
.25	-0.66	-1.17	0.52	-0.65	-1.09	0.44
.50	-0.10	-0.55	0.45	-0.07	-0.55	0.48
.75	0.57	0.27	0.30	0.59	-0.07	0.67
.90	1.42	0.92	0.50	1.33	0.63	0.70
	white	non-white	difference	white	non-white	difference
.10	-1.28	-1.76	0.47	-1.15	-1.54	0.38
.25	-0.74	-1.17	0.43	-0.74	-1.15	0.41
.50	-0.20	-0.55	0.35	-0.23	-0.55	0.32
.75	0.42	0.27	0.15	0.44	0.02	0.41
.90	1.16	0.92	0.24	1.10	0.63	0.48

Table 7.2

Coefficients on SMALL from quantile regressions by subgroup

7.2a. Non-white

	Kindergarten				First grade			
	Math		Reading		Math		Reading	
quantile	white	non-white	white	non-white	white	non-white	white	non-white
.10	**.117**	**.118**	**.042**	**.241**	**.064**	**.117**	**.124**	**.098**
	(.057)	*(.089)*	*(.038)*	*(.049)*	*(.065)*	*(.068)*	*(.049)*	*(.066)*
.25	**.074**	**.075**	**.121**	**.190**	**.201**	**.137**	**.164**	**.181**
	(.045)	*(.064)*	*(.035)*	*(.057)*	*(.064)*	*(.083)*	*(.055)*	*(.070)*
.50	**.166**	**.094**	**.152**	**.146**	**.179**	**.236**	**.289**	**.252**
	(.043)	*(.057)*	*(.040)*	*(.056)*	*(.068)*	*(.073)*	*(.063)*	*(.069)*
.75	**.173**	**.171**	**.131**	**.206**	**.271**	**.334**	**.325**	**.308**
	(.065)	*(.107)*	*(.057)*	*(.093)*	*(.064)*	*(.093)*	*(.067)*	*(.085)*
.90	**.332**	**.263**	**.185**	**.293**	**.279**	**.384**	**.299**	**.357**
	(.079)	*(.150)*	*(.082)*	*(.132)*	*(.069)*	*(.129)*	*(.075)*	*(.113)*
OLS	**.173**	**.122**	**.141**	**.232**	**.170**	**.246**	**.160**	**.264**
	(.038)	*(.060)*	*(.040)*	*(.051)*	*(.045)*	*(.061)*	*(.046)*	*(.047)*
N	2564	1215	2536	1194	1910	870	1849	865

7.2b. Free lunch

	Kindergarten				First grade			
	Math		Reading		Math		Reading	
quantile	non-fl	free lunch	non-fl	free lunch	non-fl	free lunch	non-fl	free lunch
.10	**.055**	**.186**	**.048**	**.182**	**.017**	**.115**	**.123**	**.106**
	(.074)	*(.060)*	*(.041)*	*(.048)*	*(.071)*	*(.066)*	*(.063)*	*(.042)*
.25	**.083**	**.133**	**.150**	**.145**	**.102**	**.181**	**.198**	**.149**
	(.050)	*(.041)*	*(.045)*	*(.039)*	*(.061)*	*(.065)*	*(.065)*	*(.058)*
.50	**.196**	**.155**	**.169**	**.177**	**.122**	**.248**	**.250**	**.254**
	(.050)	*(.068)*	*(.038)*	*(.043)*	*(.064)*	*(.049)*	*(.074)*	*(.049)*
.75	**.219**	**.165**	**.190**	**.209**	**.209**	**.308**	**.372**	**.169**
	(.070)	*(.077)*	*(.064)*	*(.057)*	*(.079)*	*(.062)*	*(.076)*	*(.095)*
.90	**.433**	**.161**	**.234**	**.173**	**.222**	**.371**	**.292**	**.430**
	(.085)	*(.110)*	*(.105)*	*(.087)*	*(.098)*	*(.095)*	*(.073)*	*(.118)*
OLS	**.167**	**.158**	**.151**	**.206**	**.176**	**.244**	**.198**	**.206**
	(.045)	*(.048)*	*(.047)*	*(.041)*	*(.050)*	*(.053)*	*(.050)*	*(.007)*
N	1991	1788	1964	1766	1526	1254	1498	1216

Dependent variable: SAT score for subject in second row. Standard errors in parentheses. "non-fl" denotes non-free lunch recipients. All regressions include controls but not school effects.

Table 7.3

Coefficients on SMALL for teacher indicators from quantile regressions by techer experience

	Kindergarten				First grade			
	Math		Reading		Math		Reading	
quantile	old teacher	new teacher	old teacher	new teacher	old teacher	new teacher	old teacher	new teacher
.10	**.182**	**-.241**	**.113**	**.228**	**.077**	**.022**	**.111**	**.185**
	(.052)	*(.124)*	*(.038)*	*(.073)*	*(.055)*	*(.125)*	*(.038)*	*(.114)*
.25	**.177**	**-.238**	**.180**	**.007**	**.143**	**.249**	**.187**	**.269**
	(.039)	*(.101)*	*(.030)*	*(.090)*	*(.042)*	*(.116)*	*(.036)*	*(.102)*
.50	**.242**	**-.342**	**.199**	**.033**	**.159**	**.346**	**.247**	**.289**
	(.037)	*(.123)*	*(.033)*	*(.086)*	*(.050)*	*(.130)*	*(.042)*	*(.107)*
.75	**.228**	**-.311**	**.201**	**.032**	**.269**	**.323**	**.319**	**.352**
	(.059)	*(.147)*	*(.050)*	*(.087)*	*(.053)*	*(.154)*	*(.050)*	*(.129)*
.90	**.386**	**-.186**	**.216**	**.095**	**.301**	**.326**	**.312**	**.442**
	(.095)	*(.115)*	*(.073)*	*(.178)*	*(.075)*	*(.235)*	*(.065)*	*(.262)*
OLS	**.234**	**-.238**	**.190**	**.078**	**.219**	**.139**	**.187**	**.282**
	(.036)	*(.083)*	*(.035)*	*(.073)*	*(.040)*	*(.091)*	*(.039)*	*(.093)*
N	3210	569	3181	549	2401	379	2338	376

Dependent variable: SAT score for the subject in the second row. Standard errors in parentheses. "new teacher" denotes teachers with fewer than three years of experience and "old teacher" denotes teachers with more than two years of experience.

References

Akerhielm, Karen. 1995. "Does Class Size Matter?" *Economics of Education Review*, 14:3 (September): 229-241.

Betts, Julian; Shkolnik, Jamie. 1999. "The Behavioral Effects of Variations in Class Size: The Case of Math Teachers." *Educational Evaluation and Policy Analysis*, 21:2 (Summer): 193-213.

Djebbari, Habiba; Smith, Jeffrey. 2008. "Heterogeneous impacts in PROGRESA." *Journal of Econometrics*, 145 (June): 64-80.

Firpo, Sergio. 2007. "Efficient Semiparametric Estimation of Quantile Treatment Effects." *Econometrica*, 75:1 (January): 259-276.

Hanushek, Erik. 2003. "The Failure of Input-Based Schooling Policies." *The Economic Journal*, 113 (February): 64-98.

Hao, Lingxin; Naiman, Daniel. 2007. *Quantile Regression*. Thousand Oaks, CA : SAGE Publications, Inc.

Heckman, James; Smith, Jeffrey. 1997. "Programme Evaluation and Social Experiments: Accounting For Heterogeneity in Programme Impacts." *Review of Economic Studies*, 64: 487-535

Hoxby, Caroline. 2000. "The Effects of Class Size on Student Achievement: New Evidence from Population Variation." *The Quarterly Journal of Economics*, 115:4 (November): 1239-1285.

Koenker, Roger; Hallock, Kevin. 2001. "Quantile Regression." *Journal of Economic Perspectives*, 15:4 (Fall): 143-156.

Krueger, Alan. 1999. "Experimental Estimates of Education Production Functions." *The Quarterly Journal of Economics*, 114:2 (May): 497-532.

---------- 2003. "Economic Considerations and Class Size." *The Economic Journal*, 113 (February): 34-63.

Krueger, Alan ; Whitmore, Diane. 2001. "The Effect of Attending a Small Class in the Early Grades on College-Test Taking and Middle School Test Results: Evidence from Project STAR." *The Economic Journal*, 111 (January): 1-28.

Lazear, Edward. 2001. "Educational Production." *The Quarterly Journal of Economics*, 116:3 (August): 777-803.

Levin, Jesse. 2001. "For whom the reductions count: A quantile regression analysis of class size and peer effects on scholastic achievement." *Empirical Economics*, 26: 221-246.

Ma, Lingjie; Koenker, Roger. 2006. "Quantile regression methods for recursive structural equation models." *Journal of Econometrics*, 134 (September): 471-506.

Maitra, Pushkar; Vahid, Farshid. 2006. "The Effect of Household Characteristics on Living Standards in South Africa 1993-1998: A Quantile Regression Analysis with Sample Attrition." *Journal of Applied Econometrics*, 21: 999-1018.