NBER WORKING PAPER SERIES

EXPERIMENTAL ESTIMATES OF EDUCATION PRODUCTION FUNCTIONS

Alan B. Krueger

Working Paper 6051

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
June 1997

I thank Helen Bain, a founder and principal director of Project STAR, for providing me with the data used in this study, Jayne Zaharias and DeWayne Fulton for answering several questions regarding the data, and Jessica Baraka and Aaron Saiger for providing outstanding research assistance. Helpful comments were provided by Jessica Baraka, Jeremy Finn, John Folger, Cecilia Rouse and seminar participants at MIT, NBER, Princeton, Vanderbilt, and the Society of Labor Economists. Financial support was provided by the National Institute of Childhood Health and Development. This paper is part of NBER's research programs in Children and Labor Studies. Any opinions expressed are those of the author and not those of the National Bureau of Economic Research.

Experimental Estimates of Education
Production Functions
Alan B. Krueger
NBER Working Paper No. 6051
June 1997
JEL Nos. I21, J24
Children and Labor Studies

ABSTRACT

This paper analyzes data from Project STAR, an experiment in which 11,600 Tennessee kindergarten students and teachers were randomly assigned to one of three types of classes beginning in the 1985-86 school year: small classes (13-17 students), regular-size classes (22-25 students), and regular-size classes with a teacher's aide. According to the original design, students were to remain in their initial class type through the third grade. In practice, however, students in regular-size classes were randomly re-assigned at the end of kindergarten, and about 10 percent of students moved between class types in second and third grade. Attrition was also common. Several statistical methods are used to investigate the impact of these limitations. The main conclusions are: (1) on average, performance on standardized tests increases by about 4 percentile points the first year students are assigned to a small class, irrespective of the grade in which the student first attends a small class; (2) after initial assignment to a small class, student performance increases by about one percentile point per year relative to those in regular-size classes; (3) teacher aides have little effect on student achievement; (4) class size has a larger effect on test scores for minority students and for those on free lunch; (5) the beneficial effect of smaller classes does not appear to result from *Hawthorne* effects.

Alan B. Krueger
Woodrow Wilson School
Princeton University
Princeton, NJ 08544
and NBER
akrueger@pucc.princeton.edu

I. Introduction

The large literature on the effect of school resources on student achievement generally finds ambiguous, conflicting, and weak results. Even quantitative summaries of the literature tend to reach conflicting conclusions. For example, based on the fact that most estimates of the effect of school inputs on student achievement are statistically insignificant, Hanushek (1986) concludes, "There appears to be no strong or systematic relationship between school expenditures and student performance." By contrast, Hedges, et al. (1994) conduct a meta-analysis of (a subset of) the studies enumerated by Hanushek and conclude, "the data are more consistent with a pattern that includes at least some positive relation between dollars spent on education and output, than with a pattern of no effects or negative effects."

Much of the uncertainty in the literature derives from the fact that the appropriate specification -- including the functional form, level of aggregation, relevant control variables, and identification -- of the "education production function" is uncertain.[1] Some specifications do consistently yield significant effects, however. Notably, estimates that use cross-state variation in school resources typically find positive effects of school resources, whereas studies that use within-state data are more likely to find insignificant or wrong-signed estimates (see Hanushek, 1996).[2] Many of these specification issues arise because of the possibility of omitted variables, either at the student, class, school, or state level. Moreover, functional form issues are driven in part by concern for omitted variables, as researchers often specify education production

---

[1]There is also debate over what should be the appropriate measure of school outputs (see Card and Krueger, 1996). Whereas education researchers tend to analyze standardized test scores, economists tend to focus on student's educational attainment and subsequent earnings.

[2]Hanushek attributes this difference to omitted state level variables that bias the multiple state studies, although it is possible that endogenous resource decisions within states (e.g., assignment of weaker students to smaller classes as required by compensatory education) bias the within state estimates and the interstate variability provides an unbiased estimate of resource effects.

functions in terms of test-score changes to difference out omitted characteristics that might be correlated with school resources (although such differencing could introduce greater problems if the omitted characteristics have a greater effect on the trajectory of student performance than on the level.) A classical experiment, in which students are randomly assigned to classes with different resources, would help overcome many of these issues and provide guidance for observational studies.

This paper provides an econometric analysis of the only large-scale randomized experiment on class size ever conducted in the United States, the Tennessee Student/Teacher Achievement Ratio experiment, known as Project STAR. Project STAR was a longitudinal study in which kindergarten students and their teachers were randomly assigned to one of three groups beginning in the 1985-86 school year: small classes (13-17 students per teacher), regular size classes (22-25 students), and regular/aide classes (22-25 students) which also included a full-time teacher's aide. After their initial assignment, the design called for students to remain in the same class type for four years. Some 6,000-7,000 students were involved in the project each year. Over all four years, the sample included 11,600 students from 80 schools. Each school was required to have at least one of each class-size type, and random assignment took place within schools. The students were given a battery of standardized tests at the end of each school year.

The STAR data have been examined extensively by an internal team of researchers. This analysis has found that students in small classes tended to perform better than students in larger classes, while students in classes with a teacher aide typically did not perform differently than students in regular-size classes without an aide (see Word, et al., 1990; Finn and Achilles, 1990; and Folger and Breda 1989). Past research primarily consists of comparisons of means between the assignment groups, and analysis of variance at the class level. In a review article, Mosteller (1995) described Project STAR as "a controlled experiment which is one of the most important educational investigations ever carried out and illustrates the kind and magnitude of research needed in the field of education to strengthen schools."

As in any experiment, there were deviations from the ideal experimental design in the actual implementation of Project STAR. First, students in regular-size classes were randomly assigned again between classes with and without full-time aides at the beginning of first grade, while students in small classes continued on in small classes, often with the same set of classmates.[3] Re-randomization was done to placate parents of children in regular classes who complained about their children's initial assignment. Because analysis of data for kindergartners did not indicate a significant effect of a teacher aide on achievement in regular-size classes, it was felt that this procedure would create few problems. But if the constancy of one's classmates influences achievement, then the experimental comparison after kindergarten is compromised by the re-randomization.

A second limitation of the experiment is that approximately 10 percent of students switched between small and regular classes between grades, primarily because of behavioral problems or parental complaints. These nonrandom transitions could also compromise the experimental results. Furthermore, because some students and their families naturally relocate during the school year, actual class size varied more than intended in small classes (11 to 20) and in regular classes (15 to 30). Finally, as in most longitudinal studies of schooling, sample attrition was common -- half of students who were present in kindergarten were missing in at

[3]If a school had more than one small class, students could be moved among the small classes.

least one subsequent year. These potential limitations of the experiment have not been addressed in previous work.

This paper has three related goals. First, to probe the sensitivity of the experimental estimates to flaws in the experimental design. For examples, a variable measuring the constancy of students' classmates is included as a control variable in the regressions, and initial random assignment is used as instrumental variable for actual class size. Second, to use the experiment to identify an appropriate specification of the education production function to estimate with nonexperimental data. And third, to use the experimental results to interpret estimates from the large literature based on observational data.

II. Background on Project STAR and Data

A. Design and Implementation

Project STAR was funded by the Tennessee legislature, at a total cost of approximately $12 million over four years.[4] The Tennessee legislature required that the study include students in inner-city, suburban, urban and rural schools.[5] The research was designed and carried out by a team of researchers at Tennessee State University, Memphis State University, the University of Tennessee, and Vanderbilt University. To be eligible to participate in the experiment, a public school was required to sign up for four years and be large enough to accommodate at least three classes per grade, so within each school students could be assigned to a small class (13-17),

[4]This section draws heavily from Word (1990) and Folger (1989).

[5]Inner city schools were defined as schools in metropolitan areas in which more than half of students received free lunch; suburban was defined as the balance of metropolitan area schools; urban was defined as towns with more than 2,500 inhabitants; and rural was defined as town with fewer than 2,500 inhabitants.

regular class (22-25 students), or regular plus a full-time aide class.[6] The statewide pupil-teacher ratio in kindergarten in 1985-86 was 22.3, so students assigned to regular classes fared about as well as the average student in the state (Word, et al., 1990). Schools with more than 67 students per grade had more than three classes. One limitation of the comparison between regular and regular/aide classes is that in grades 1-3 each regular class had the services of a part-time aide 25-33 percent of the time on average, so the variability in aide services was restricted.[7]

The cohort of students who entered kindergarten in the 1985-86 school year participated in the experiment through third grade. Any student who entered a participating school in a relevant grade was added to the experiment, and participating students who were retained a grade or left the school exited the sample. Entering students were randomly assigned (based on a list of random numbers) to one of the three types of classes: small, regular, or regular/aide. Students in regular classes and in regular/aide classes were randomly re-assigned between these two types of classes at the end of kindergarten, while students initially in small classes continued on in small classes. Notice, however, that results from the kindergarten year are uncontaminated by this feature of the experiment.

Because kindergarten attendance was not mandatory in Tennessee at the time of the study, many new students entered the program in first grade. Additionally, students were added to the sample over time because they repeated a grade or because their families moved to a school zone

---

[6]Participating schools had an average per-pupil expenditure in 1986-87 of $2,724, compared to the statewide average of $2,561.

[7]The reason that regular classes often had a teacher aide is that the ethic underlying the study was that students in the control group (i.e., regular classes) would not be prevented from receiving resources that they ordinarily would receive.

that included a participating school. In all, some 2,200 new students entered the project in first grade and were randomly assigned to the three types of classes. About 1,200 to 1,500 new students entered the experiment in the second or third grade. Newly entering students were randomly assigned to class types, although the uneven availability of slots in small and regular classes often led to an unbalanced allocation of new students across class types.

A total of 11,600 children were involved in the experiment over all four years. After third grade, the experiment ended and all students were assigned to regular-size classes. Although data have been collected on students through 9th grade, the present study only has access to data covering grades K-3.

A limitation of the study is that baseline test score information on the students is not available, so one cannot examine whether the treatment and control groups "looked similar" on this measure before the experiment began. But even if baseline tests were available, the information might be of limited value because it is difficult to meaningfully test kindergarten students as they enter school. Nonetheless, if the groups of students were truly randomly assigned, one would expect those assigned to small- and regular-size classes to look similar along other measurable dimensions at base line. Tables 1 and 2 provide some evidence on the differences among students assigned to the three types of classes.

Table 1 disaggregates the data into waves, based upon the grade the students entered the program, because this was the first time the students were randomly assigned to a class type. Sample means by class type for several variables are presented. As one would expect, students assigned to small classes had fewer students in their class than those in regular classes, on average. There are small differences in the fraction of students on free lunch, the racial mix, and the average age of students in classes of different size, although some of these differences are statistically significant (see rows 1-4).[8] Because random assignment was only valid within schools, these differences suggest the importance of controlling for school effects.

Table 2 presents p-values for joint F-tests of the differences among small, regular, and regular/aide classes for the variables presented in Table 1. Unlike results reported in Table 1, these p-values are conditional on school effects. None of the three background variables displays a statistically significant association with class-type assignment at the 10 percent level, which suggests that random assignment produced relatively even groups in each class size, on average. As an overall test of random assignment, I regressed a dummy variable indicating assignment to a small class on the three background measures in rows 1-3 and school dummies. For each wave, the student characteristics had no more than a chance association with class-type assignment. Furthermore, if the same regression model is estimated for a sample that pools all four entering waves of students together, the three student characteristics are still insignificantly related to assignment to a small class (p-value=.58). Within schools, there is no apparent evidence that initial assignment to class types was correlated with student characteristics.

There was a high rate of attrition from the project. Only half of students who entered the project in kindergarten were present for all grades K-3. Among those who entered kindergarten, students in small classes were 3-4 percentage points more likely to stay in the sample than those in regular-size classes. This pattern was reversed among those who entered in first grade, however. Attrition could occur for several reasons, including students moving to another school (perhaps endogenously), students repeating a grade, and students being advanced

[8]To be precise, the fraction on free lunch actually measures the fraction who receive free or reduce-price lunch.

a grade. Although I lack data on retention rates for the early grades, Word, et al. (1990) report that over the four years of the project, 19.8 percent of students in small classes were retained while 27.4 percent of students in regular classes were retained. This is consistent with the lower attrition rate of students in small classes. Some of the analysis that follows makes a crude attempt to adjust for possible nonrandom attrition.

To check whether teacher assignment was independent of observed teacher characteristics, I regressed each of three teacher characteristics (experience, race, or education) on dummies indicating the class type the teachers were assigned to and school dummies, and then performed an F-test of the hypothesis that the class-type dummies jointly had no effect. These regressions were calculated for each of the four grade levels, so there were a total of 12 regressions. In each case, the p-value for the class-type dummies was less than .05.[9] These results are as one would expect with random assignment of teachers to the different class types.

It is virtually impossible to control the exact number of students in a class: Families move in and out of a school district during the course of a year; students become sick; and varying numbers of students are enrolled in schools. As a result, in some cases actual class size deviated from the intended ranges. Table 3 reports the frequency distribution of class size for first graders, by assignment to small, regular, or regular/aide classes. Although students assigned to small classes clearly were more likely to attend classes with fewer students, there was considerable variability in class size within each class-type assignment, and some overlap between the distributions.

---

[9]In two cases the p-value was less than .10. Third grade teachers assigned to small classes were less likely to have a masters degree or higher than were teachers assigned regular-size classes, and first grade teachers in small classes had 2 more years of experience than those in regular-size classes (although less experience than those in regular/aide classes).

It is also virtually impossible to prevent some students from switching between class types over time. Table 4 shows a transition matrix between class types for students who continued from K-1, 1-2, and 2-3 grades. If students remained in their same class type over time, all the off-diagonal elements would be zero. The re-randomization of students in regular classes in first grade is apparent in panel A. But in second and third grades, when students were supposed to remain in their same type of class, 9-11 percent of students switched class-size types. Students were moved between class types because of behavioral problems or, in some cases, parental complaints. Obviously, if the movement between class types was associated with student characteristics (e.g., students with stronger academic backgrounds more likely to move into small classes), these transitions would bias a simple comparison of outcomes across class types.

To address this potential problem, and the variability of class size for a given type of assignment, in some of the analysis that follows initial random assignment is used as an instrumental variable for actual class size.[10]

B. Data and Standardized Tests

Students were tested at the end of March or beginning of April of each year. The tests consisted of the Stanford Achievement Test (SAT), which measured achievement in reading, word recognition, and math in grades K-3, and the Tennessee Basic Skills First (BSF) test, which measured achievement in reading and math in grades 1-3. The tests were tailored to each grade level. Because there are no natural units for the test results, I scaled the test scores into

[10]Initial assignment is measured by the students' class assignment the first year the student is observed in the experiment. Students were typically notified of their initial class assignment very close to the beginning of the school year.

percentile ranks.[11] Specifically, in each grade level the regular and regular/aide students were pooled together, and students were assigned percentile scores based on their raw test scores, ranging from 0 (lowest score) to 100 (highest score). A separate percentile distribution was generated for each subject test (e.g., Math-SAT, Reading-SAT, Word-SAT, etc.). For each test we then determined where in the distribution of the regular-class students every student in the small classes would fall, and the students in the small classes were assigned these percentile score. Finally, to summarize overall achievement, the average of the three SAT percentile rankings was calculated.[12] If the performance of students in the small classes was distributed in the same way as performance of students in the regular classes, the average percentile score for students in the small classes would be 50.

Table 5 presents the correlations among the individual components of the SAT and BSF test for the subset of students who attended both first and second grade.[13] It is reassuring that the strongest correlations typically are between tests of the same subject matter; for example, in second grade the SAT and BSF reading tests have a correlation of .80. The closely related word and reading test also have a high correlation. Also notice that tests of the same subject tend to have a higher correlation from one grade to the next than tests of different subjects. The

[11]There is some precedent for using percentile scores on tests as an explanatory variable in wage regressions; see, for example, Griliches and Mason (1972).

[12]Formally, denote the cumulative distribution of scores on test j (denoted T^j) of students in the regular and regular/aide classes as $F^R(T^j) = \text{prob}[T^j_{iR} < T^j) = y^j$. For each student i in a small class, we then calculated $F^R(T^j_{iS}) = y^j_{iS}$. Naturally, the distribution of y^j for students in regular classes follows a uniform distribution. We then calculated the average of the three (or two for BSF) percentile rankings for each student. If one subtest score was missing, we took the average of the two percentiles that were available, and if two were missing we used the percentile score corresponding to the only available test.

[13]The same general pattern holds for other grades. These grades were presented to reduce the effect of attrition, and keep the data tractable.

correlation between the average SAT percentile and average BSF percentile (not shown in table) is .79 in first grade and .85 in second grade. For most of the subsequent analysis, the SAT exam is the primary focus of study because this test has been used on a national level for a long period of time.

The average of the three SAT exams by class type is presented in the last row of Table 1. Figure 1 displays the kernel density of the average test score distributions for students in small and regular classes at each grade level. In all grades, the average student in small classes performed better on this summary test measure than did those in regular or regular/aide classes. There does not seem to be a very strong or consistent effect of the teacher aide, however. The rest of the paper probes the robustness of these preliminary finding.

Observe also that the average test score of students in all class types tends to decline with the grade in which the student entered the experiment. This correlation is likely to reflect the fact that kindergarten was optional and higher-achieving students were more likely to attend kindergarten, as well as the tendency of lower-achieving students to be retained and disproportionately added to the sample at higher grade levels. Because of this feature of the data, it is desirable to control for the grade in which the student entered project STAR in some of the analysis that follow.

The appendix table presents means for several additional variables that are available in the data set.

III. Statistical Models

To see the advantage of a randomized experiment in estimating the effect of school resources on student achievement, consider the following general model:[14]

$$(1)\quad Y_{ij} = a\,S_{ij} + b\,F_{ij} + \epsilon_{ij},$$

where Y_{ij} is the achievement level of student i in school j, S_{ij} is a vector of school characteristics, F_{ij} is a vector representing the family background of the student, and ϵ_{ij} is a stochastic error component. In principle, S_{ij} and F_{ij} include information cumulated over the student's life; for example, classroom size and teacher qualifications for each year the student attended school. The entire history of family background variables and school resources may contribute to students' achievement in a given year. In addition, children's unobserved inherent ability may also contribute to their achievement. In any actual application we will generally lack data on some relevant school, family, or student characteristics. These omitted variables will then appear in the error term. If the omitted variables are correlated with the included variables, then the estimated parameters will be biased.

If a school characteristic such as class size is determined by random assignment, however, it will be independent of the omitted variables. Thus, with random assignment, a simple comparison of mean achievement between children in small and large classes provides an unbiased estimate of the effect of class size on achievement.

We begin analyzing the STAR data by estimating the following regression equation for students in each grade level:

[14]This general framework is essentially the same model as in Boardman and Murnane (1987) and Hanushek and Taylor (1990).

(2) $Y_{ics} = \beta_0 + \beta_1 \text{ SMALL}_{cs} + \beta_2 \text{ REG/A}_{cs} + \beta_3 X_{ics} + \alpha_s + \epsilon_{ics}$

where Y_{ics} is the average percentile score on the SAT test of student i in class c at school s, $SMALL_{cs}$ is a dummy variable indicating whether the student was assigned to a small class that year, REG/A_{cs} is a dummy variable indicating whether the student was assigned to a regular size class with an aide that year, and X_{ics} is a vector of observed student and teacher covariates (e.g., gender). The independence between class-size assignment and other variables is only valid within schools, because randomization was done with-in schools. Consequently, a separate dummy variable is included for each school to absorb the school effects, α_s.

The equation is estimated by Ordinary Least Squares (OLS). In calculating the standard errors, however, the error term, ϵ_{ics}, is modelled in a components of variance framework. Specifically, ϵ_{ics} is assumed to consist of two components: $\epsilon_{ics} = \mu_{cs} + \epsilon'_{ics}$, where μ_{cs} is a class-specific random component that is common to all members of the same class, and ϵ'_{ics} is an idiosyncratic error term.[15] The class-specific component, μ_{cs}, may exist because of unobserved teacher characteristics, or because some students may exert a common influence over others in the class.

Because several students were re-assigned to different classes after their initial random assignment, in part based on their performance, equation (1) was also estimated using dummies indicating students' initial assignment the first year they entered the program, rather than their actual assignment each year. Models including initial assignment are labelled "reduced form" models, because one can think of initial assignment as an excluded variable that is correlated with

[15] The adjusted-standard errors are about two-thirds larger than the OLS standard errors.

actual class size. (Because initial assignment and actual assignment were identical in kindergarten, these models are identical for kindergarten.)

Regression results for these models are presented in Table 6. Columns 1-4 use actual assignment, and columns 5-8 use initial class assignment. Columns 1 and 5 omit the school dummies. As earlier analyses of the data have found, students in small classes tend to perform better than those in regular and regular/aide classes. Here, the gap in average performance is about 5 percentile points in kindergarten, 8.6 points in first grade, and 5-6 points in second and third grade. Columns 2 and 6 add unrestricted school dummies to the model. In three of four grades, including the school dummies leads to a slight increase in the effect of being assigned to a small class.

If class size were truly randomly assigned, including additional exogenous variables would not significantly alter the coefficient on the class-size dummies. In fact, including covariates seems to have a very modest effect on the class-size coefficients conditional on school effects. The student characteristics in columns 3 and 5 add considerable explanatory power. White and Asian students tend to score 8 percentile points higher than black students in kindergarten, and this gap is about 6 points in third grade.[16] Students on free lunch score 13 percentile points less than those not on free lunch, and girls score 3-4 points higher than boys in each grade level.

The teacher characteristics have notably weak explanatory power. Teacher education -- as proxied by a dummy indicating whether the teacher has a master's degree -- does not have a systematic effect. Hardly any of the teachers are male, so the gender results are not very

[16] Ninety-nine percent of the students are white or black. The small number of Asian students are included with white students in the analysis. The small number of hispanic students and others are included with the black students.

meaningful. Teacher experience has a small, positive effect. Experimentation with a quadratic in experience indicated that the experience profile tends to peak at about 20 years of experience, and students in classes where the teacher has 20 years of experience tend to score about 3 percentile points higher than those in classes where the teacher has zero experience, all else being equal. As a whole, however, consistent with much of the previous literature, the STAR data suggest that teacher characteristics explain relatively little of student achievement as measured by standardized tests.

Estimates of the effect of being in a small class which use initial assignment (columns 5-8) are only slightly smaller than the estimates which use the actual class assignment (columns 1-4), and are always statistically significant. This finding suggests that possible non-random movement of students between small and regular classes was not a major limitation of the experiment.

To summarize these results, based on column 4 it appears that students in small classes score about 5-7 percentage points higher than those assigned to regular size classes. Students assigned to a regular/aide class perform slightly better (1 or 2 percentile points, on average) than students assigned to a regular class without a full-time aide, but the gap is only statistically significant in one grade level. Thus, it is possible that a teacher aide has only a trivial effect on student achievement, or that the availability of part-time aides in regular classes confounds the true effect of an aide.

Is the impact of attending a small class big or small? Unfortunately, it is unclear how percentile scores on these tests map into to tangible outcomes. Nevertheless, a couple of comparisons are informative. First, relative to the standard deviation of the average percentile score, the effect sizes are: .20 in kindergarten, .28 in first grade, .22 in second grade, and .19

in third grade (based on the model in column 4). Second, one could compare the estimated class-size effects to the effects of other student characteristics. For example, in kindergarten the impact of being assigned to a small class is about 64 percent as large as the white-black test score gap, and in third grade it is 82 percent as large. By both metrics, the magnitudes are sizable.

A. Effects of Attrition

Table 7 provides some simple evidence on the impact of sample attrition. As is common in longitudinal studies of education, attrition was very high from Project STAR classes. If the students originally assigned to regular classes who left the sample had higher test scores, on average, than students assigned to small classes who also left the sample, then the small class effects will be biased upwards. One reason why this pattern of attrition might occur is that high-income parents of children in larger classes might have been more likely to enroll their children in private schools over time than similar parents of children in small classes. At heart, adjusting for possible nonrandom attrition is a matter of imputing test scores for students who exited the sample. With longitudinal data, this can be done crudely by assigning the student's most recent test percentile to that student in years when the student was absent from the sample.[17]

The sample used in column 1 of Table 7 includes the largest number of students with non-missing data available each grade. These results correspond closely to the model and sample used in column 7 of Table 6, except the free lunch variable is omitted because it changes over

[17]In the case of a student who left the sample but later returned, the average test score in the years surrounding the student's absence was used. Test scores were also imputed for students who had a missing test score but did not exit the sample (e.g., because they were absent when the test was conducted).

time.[18] For simplicity, only the coefficient on the dummy variable indicating initial assignment to a small class is reported in the table. The sample used in column 2 is larger than the sample in column 1 because it includes the column 1 sample plus any student who entered the program in an earlier grade and exited the sample by the current grade, assigning imputed test percentiles to students who exited the sample. (Because kindergarten students could not have previously exited the sample, the sample size is the same in the first row.) Estimates using imputed test percentiles for missing observations are qualitatively quite similar to the estimates using the subsample of observations who were present in each particular grade.[19] Thus, nonrandom attrition does not appear to bias the estimated class size effects in Table 6, and the remainder of the paper utilizes only those observations with non-missing data.

B. Two-Stage Least Squares Models

As noted, students in the Project STAR experiment who were assigned to small classes had a varying number of students in their classes because of student mobility and enrollment differences across schools. Similarly, students in the regular-size classes had variable class sizes. A more appropriate model of achievement would take actual class size into account. A natural model for this situation is a triangular model of student achievement in which the actual number of students in the class is included on the right-hand side, and initial assignment to a class type is used as an instrumental variable for actual class size. Specifically, we estimate the following

---

[18]The estimated model uses initial class assignment so as to avoid imputing actual class size for missing observations.

[19]The coefficient on the regular/aide initial assignment dummy is also quite similar if the model is estimated with or without the imputed data.

model by Two-Stage Least Squares (2SLS):

$$(3)\quad CS_{ics} = \pi_0 + \pi_1 S_{ics} + \pi_2 R_{ics} + \pi_3 X_{ics} + \delta_s + \tau_{ics}$$

$$(4)\quad Y_{ics} = \beta_0 + \beta_1 CS_{ics} + \beta_2 X_{ics} + \alpha_s + \epsilon_{ics}$$

where CS_{ics} is the actual number of students in the class, S_{ics} is a dummy variable indicating assignment to a small class the first year the student is observed in the experiment, R_{ics} is a dummy variable indicating assignment to a regular class the first year the student is observed in the experiment, and all other variables are defined as before.[20] Again, the error term (ϵ_{ics}) is treated as consisting of a common class effect and an idiosyncratic individual effect, and the standard errors are adjusted for correlation in the residuals among students in the same class.

In this setup, only variation in class size due to <u>initial</u> random assignment to a regular or small class is used to provide variation in actual class size in the test score equation. Due to the random assignment of initial class type, one would expect that this excluded instrumental variable is uncorrelated with ϵ_{ics}, as required for 2SLS to be asymptotically unbiased. If attending a small class has a beneficial effect on students' test scores, we would expect β_1 to be negative.

OLS and 2SLS estimates are presented in Table 8. The 2SLS estimates tend to be a little larger in absolute value, especially in third grade. According to the 2SLS estimates, a reduction of 10 students is associated with a 7 to 9 point increase in the average percentile ranking of students, depending on the grade. There is no obvious trend over grade levels in the effect of class size in these data.

[20]Because the teacher aide was found to have a small effect in Table 6, we do not hold constant the availability of an aide in equation (4). One could, however, add a dummy indicating the presence of a full-time aide to equation (4).

Table 9 presents several additional 2SLS estimates of the effect of actual class size on achievement, disaggregating the sample by the grade the student entered Project STAR and current grade. The model and identification strategy are the same as in Table 8, column 2. Figure 2 also displays the estimated effects of attending a small, regular, and regular/aide class by entry wave and grade using the OLS specification in Table 6, column 4. Both sets of results indicate that for each cohort of students, those attending smaller class tend to score higher on the standardized test by the end of the first year they entered the experiment. If assignment to small or regular classes was somehow nonrandom, then the initial assignment would have to have been skewed in the direction of producing higher test scores in the small classes for each wave of students who entered the program -- an unlikely event. Interestingly, for the wave of students who entered in kindergarten, the beneficial effect of attending a small class does not appear to increase as students spend more time in their class assignment. For students entering the experiment in first or second grade, however, the test score gap between those in small- and regular-size classes grows as students progress to higher grades. The effect of time spent in a small class is explored further by pooling students in all grades together below.

C. Models with Pooled Data

To explore the cumulative effects of having been in a small or regular class, several models were estimated with the data pooled over students and grades. The general model was of the form:

$$(5)\quad Y_{igcs} = \beta_0 + \beta_1\, SMALL_{igcs} + \beta_2\, REG/A_{igcs} + \beta_3 \sum_{0}^{g-1} SMALL_{igcs} + \beta_4 \sum_{0}^{g-1} REG/A_{igcs} + \beta_5\, X_{igcs} + \alpha_g + \alpha_f + \alpha_s + \epsilon_{igcs}$$

where g indicates grade level (K,1,2 or 3), SMALL and REG/A are dummy variables indicating the class type, Σ SMALL is the cumulative number of past years the student was in a small class, Σ REG/A is the cumulative number of past years the student was in a regular/aide class, α_g is a set of three current grade dummies, α_f is a set of three dummies indicating the first year the student entered the STAR sample, and α_s is a set of school fixed effects. Estimation is done by OLS and 2SLS, but robust standard errors which allow for a random individual component in the error term are reported. Because some students were switched between class types after their first year in the experiment, the 2SLS estimates use initial class assignment and potential cumulative years in the class type if the student had stayed in the initially assigned class type each grade as instruments for class type and cumulative years in each class type.

The first three columns of Table 10 present OLS estimates, and the second three columns present 2SLS estimates. Estimates shown in column 1 exclude student, teacher and classmate characteristics. In column 2, regressors for student and teacher characteristics are included. Both of these models indicate that achievement of students in small classes jumps up by about 4 percentile points if the student attends a small class, and improves by a little less than one percentile point for each additional year thereafter the student spends in a small class. The initial effect of being in a small class is highly significant (t=8), while the cumulative effect is just on the margin of statistical significance. The corresponding 2SLS estimates in columns 4 and 5 show a slightly larger discrete increase from being assigned to a small class.

Column 3 adds four variables reflecting the composition of student's classmates. Students in small classes were more likely to remain with their classmates in first grade because students in regular classes were randomly re-assigned between regular classes with and without full-time aides. Two variables are included to control for the impact of the constancy of one's classmates. First, the fraction of each student's classmates who were in that student's class the preceding year is included. If a student is new to the school in a particular grade, this variable will have a value of 0; and if a student attends a class that consists only of students who were in that student's class the preceding year, the variable will have a value of 1. As a second measure of the environment in the class, we take the average of this variable over all the other students in the class. This variable might influence achievement because the extent to which other students in a class know each other could influence one's adjustment to the class.

In addition to these two "class constancy" variables, the regression includes the fraction of students in a class who receive free lunch and the fraction of students in the class who ever attended kindergarten (based on the STAR data). Because students on free lunch score lower on standardized tests than other students, a higher proportion of classmates on free lunch in a class may lower overall performance. The fraction of a class that attended kindergarten could affect achievement because kindergarten attendance is likely to make the class more socialized for school, which should enable the teacher to convey more material. Due to the random assignment of students, these variables should be uncorrelated with any omitted variables within schools.

The results of adding these class-level variables are quite interesting. Most importantly, including the four variables leads the discrete jump in test score associated with attending a small class to decline to 3.6 percentile points in column (3), although it is still highly statistically significant.[21] Including the variables also causes the cumulative effect of years in a small class to decline slightly, and to slip just below the level of statistical significance.

[21] Students in small classes are about 13 percentage points more likely to have classmates who attended kindergarten than students in regular classes.

Also notice that attendance in classes with a higher proportion of classmates who attended kindergarten has a large, positive effect on one's own achievement. A two standard deviation change in the fraction of one's classmates who attended kindergarten is associated with about a 3 percentile point change in test scores. Test scores are not significantly related to the variables measuring the constancy of one's classmates. However, these variables are set to zero in kindergarten as all kindergarten students are new to the class. If the model in column (3) is estimated using the subsample from first grade on, students who are new to classes that include many students who were together the previous grade tend to score significantly lower on the SAT exam ($t = -3.2$). Thus, if a student is new to a class, he or she does better if most of the other students are new to the class as well. A higher fraction of classmates on free lunch has a negative, marginally statistically significant effect on achievement in the post-kindergarten sample.

The pooled models in Table 10 allow for a one-time, discrete improvement in test scores from attending a small class, and for a constant increase for each additional year the student spends in a small class. One could estimate a more general model. Most obviously, the initial effect of being in a small class could vary by grade level (i.e., interact grade dummies and SMALL), and the linear trend of past cumulative time in a small class could be relaxed by including a set of unrestricted dummies indicating the number of past years spent in a small class. In results not presented here, such a less restrictive model was estimated. The estimates in Table 10 are nested in this model, so they can be tested against it. An F-test rejects the parsimonious specification in Table 10 at the .01 level. However, inspection of the coefficients suggest that the main reason for the rejection is that the effect of being in a small class varies somewhat from grade to grade; the linear trend appears to be a plausible representation of the cumulative effect of time spent in a small class. Despite this rejection, the parsimonious model is a convenient way to summarize the effect of attending a small class in the early grades.

The relationship between the pooled models and the "value-added" specification that is commonly estimated in the education production function literature should be emphasized. The value-added model is only identified by the cumulative effect of time spent in a small class; the initial effect is differenced out for students who spend more than one year in a small class. Had the estimates in Table 10 indicated that the effect of the initial year spent in a small class was no different than the effect of additional years, the value-added specification would capture the only parameter of interest. But the pooled estimates and the pattern in Figure 2 indicate that the most important benefit of attending a small class occurs the first year a student is placed in a small class. This benefit is missed in the value added specification.

This point is illustrated by estimating the following value-added specification:

$$(6)\quad Y_{ics,g} - Y_{ics,g-1} = \beta_0 + \beta_1\, SMALL_{ics,g} + \beta_2\, X_{ics,g} + \alpha_g + \alpha_s + \epsilon_{ics,g}$$

where the dependent variable is the change in students' percentile test scores between grade g and g-1. The coefficient β_1 essentially corresponds to the coefficient on cumulative time spent in a small class in equation (5). When this specification is estimated, the estimate of β_1 is 1.2, with a t-ratio of 3.1.[22] The coefficient is slightly larger than, but in the same ballpark as, the coefficient on the cumulative years in a small class variable in Table 10. Thus, although the estimated value-added specification indicates that students gain from attending small classes, the

[22]The other covariates in this regression are the same as in column 3 of Table 10.

benefit is substantially less than the full effect estimated from a comparison of levels.

D. Heterogeneous Treatment Effects

The effect of being in a small class may vary for students with different backgrounds. Table 11 presents OLS estimates of the pooled model (equation 5) for several subsamples of students. The pooled model was selected to summarize the class size effects over all grade levels, although a less restrictive model would probably fit the data better.

Smaller classes tend to have about an equal initial effect, but a greater cumulative effect, for boys as compared to girls, and for students on free lunch as compared to those not on free lunch. Black students tend to have a somewhat greater initial test score effect of attending a small class and a greater increase over time, but the difference between the cumulative effect for blacks and whites is not statistically significant. Finally, inner-city students (defined as those attending schools in metropolitan areas with greater than half of students on free lunch) tend to have a more beneficial effect of attending a small class in the first year they attend a small class than students from other areas, and roughly the same cumulative benefit over time. Word, et al. (1990) similarly found smaller classes had a more beneficial effect for black students, students on free lunch, and inner-city students, but did not examine whether these differences were due to the initial effect or cumulative effect of time spent in a small class. In general, the pattern of effects reported in Table 11 suggests that the lower achieving students benefit the most from attending smaller classes. Summers and Wolfe (1977) also find that attending a small class is more beneficial for low achieving students than high achieving students.

The effect of attending a small class can also be estimated for each of the 80 schools in

Project STAR. To estimate school-level small-class effects, I pooled the data for students across grades, and for each school regressed the percentile score on dummies indicating attendance in small and regular/aide classes, current grade dummies, and dummies indicating the grade the student entered project STAR. A parsimonious model was estimated for simplicity and to preserve degrees of freedom. A kernel density for the coefficients on the small-class dummy is shown in Figure 3. Two-thirds of the school-specific small-class effects are positive, while one-third are negative. Furthermore, 2.5 percent of the 80 coefficients had t-ratios less than -2, while 30 percent had t-ratios exceeding +2. The mean coefficient estimate is 4.6. The standard deviation of the coefficients (after adjusting for sampling variability) is 6.9 percentage points.[23] Thus, some schools are more adept at translating smaller classes into student achievement than are other schools.

E. *Hawthorne* and *John Henry* Effects

It has been suggested by some that the effectiveness of small classes found in the STAR experiment may have been due to "Hawthorne effects," in which teachers in small classes responded to the fact that they were part of an experiment that was expected to show that small classes benefit students, rather than a true causal effect of small classes themselves. Others have suggested that the effect sizes might actually be larger than measured by the STAR experiment because teachers in regular classes provided greater than normal effort to demonstrate that they could overcome the bad luck of being assigned more students: a "John Henry" effect. Either set of responses could limit the external validity of the results of the STAR experiment.

[23] To adjust for sampling variability in the coefficient estimates, the average squared standard error was subtracted from the variance of the estimated coefficients.

As a partial check on these potential "reactive" effects, I examined the relationship between class size and student achievement just among students assigned to regular-size classes. Recall that there is considerable variability in class size even in the regular-size classes (see Table 3).[24] Obviously, Hawthorne and John Henry effects do not apply to a sample in which all teachers were randomly assigned to the control group. On the other hand, variability in class size is likely to be due primarily to idiosyncratic factors in this sample, such as integer effects in assigning classes and student mobility during the school year. Moreover, there is limited variability in class sizes within schools because many schools had only one control class per grade. Also note that the effect of class size in this sample will be diminished if there are threshold effects in achievement around 15 students in a class, because the variability in class size is around a higher mean for students in regular-size classes.

To estimate the effect of class size on achievement for the control sample, I pooled the sample of students in regular-size classes across all grade levels, and regressed the average SAT test score on the number of students in the class, grade level dummies, and student and teacher characteristics. Results are reported in Table 12. The coefficient on class size in this regression is -.55, with a t-ratio of -4.3. When school dummies are added to this model in column 2, the coefficient on class size falls to -.39, but remains statistically significant (t=-3.1). Based on these coefficients, an 8 student reduction in class size is associated with a 3 to 4 percentile increase in test scores. These regressions do not provide much evidence of Hawthorne or John Henry effects. And given that much of the variability in class size in the control group may be

[24]The standard deviation of class size in the sample of students assigned to regular classes is 2.3, as compared to 4.1 among all students in the experiment.

due to measurement errors (e.g., students moving in and out of class during the school year), it is notable that these regressions find significant evidence of class size effects at all.

F. Separate Subject Test Results

Table 13 presents estimates of the pooled data model corresponding to column 3 of Table 10 for each of the main subsections of the SAT test, as well as for the subsections of the BSF test and the average of the math and reading percentile scores on the BSF test. These results indicate relatively minor differences between the effect of attending a small class on the math, reading and word recognition tests. Furthermore, the BSF test shows the same basic pattern as the SAT test -- a substantial discrete increase in performance for attending a small class, with a small (statistically insignificant) increase thereafter. On the whole, little seems to have been lost by focusing on the average of the SAT tests in the mainstay of the analysis.

IV. Conclusion

One well designed experiment should trump a phalanx of poorly controlled, imprecise observational studies based on uncertain statistical specifications. The implementation of the STAR experiment was not flawless, but my re-analysis suggests that the flaws in the experiment did not jeopardize its main results. Adjustments for school effects, attrition, re-randomization after kindergarten, nonrandom transitions, and variability in actual class size do not overturn the main findings of Word, et al. (1990) and Finn and Achilles (1990): Students in small classes scored higher on standardized tests than students in regular-size classes. The results also indicate that the provision of a full-time teacher aide has, at best, a modest effect on student achievement, although this effect may be attenuated because of the frequent availability of part-time aides in regular classes.

Interestingly, at least for the early grades, my analysis suggests that the main benefit of attending a small class seems to arise by the end of the initial year a student attends a small class. After the first year, cumulative time spent in a small class has a relatively minor, positive impact on test scores. One possible explanation for this pattern is that attending a small class in the lower grades may confer a one-time, "school socialization effect" which permanently raises the level of student achievement without greatly affecting the trajectory.

Because much of the previous literature estimates class-size effects using a "value-added" specification that uses student test score gains as the dependent variable and current class size as the main explanatory variable for a sample of students after their initial exposure to small or large classes, much of the past research may miss the main benefit of smaller classes. More research is needed to develop an appropriate model of student learning. But for now, one should be concerned that the value-added specification may miss much of the value that is added from attending a smaller class. Moreover, studies that identify class size effects by comparing differences in the level of test scores between students who were subject to different class sizes for exogenous reasons, such as Angrist and Lavy's (1997) clever use of Maimonides law, may stand a better chance of uncovering the total effect of class size than estimates based on the value-added specification.

No single study, even an experimental one, could be definitive. The STAR results suggest that the magnitude of the achievement gains from attending smaller classes varies across schools and student characteristics. It is possible (though probably unlikely) that Tennessee has a much higher concentration of students or schools that benefit from smaller classes than other states. It is also possible that reducing class size does not have a beneficial effect for students after the third grade. Obviously, more experimentation would help examine these issues. It would also be helpful to compare the STAR findings to the rest of the literature. Before concluding that the weight of the literature suggests that attending a small class does not matter for the average student, it would be useful to know how many of the studies enumerated in Hanushek's (1986, 1996) surveys have sufficient power to reject either the level effect (for level specifications) or cumulative effect (for value added specifications) of attending a small class that is implied by the Project STAR data.

Finally, experiments of the scale and quality of Project STAR are disappointingly rare in the education field. When these experiments are conducted, they should be analyzed and followed-up to the fullest extent possible. The students who participated in Project STAR were returned to regular classes after third grade, and have been followed-up through the ninth grade. Nye, et al. (1994) find that students who were placed in small classes have lasting achievement gains through at least the seventh grade, although it is difficult to compare the magnitude of the benefits to those at earlier grades because of changes in the tests that were administered. The students studied in Project STAR are currently in high school. To learn more about the long-term effects of attending smaller classes, it would be useful to continue studying the academic -- and just as importantly, nonacademic -- outcomes of the STAR participants as they enter early adulthood.

References

Angrist, Joshua and Victor Lavy. (1996.) "Using Maimonides' Rule to Estimate the Effect of Class Size on Children's Academic Achievement." Mimeo. MIT, Cambridge, MA.

Boardman, Anthony E. and Richard J. Murnane. (1979). "Using Panel Data to Improve Estimates of the Determinants of Educational Achievement." Sociology of Education 52 (No. 2): 113-121.

Card, David and Alan B. Krueger. (1996). "Labor Market Effects of School Quality: Theory and Evidence." In Gary Burtless, editor, Does Money Matter? The Effect of School Resources on Student Achievement and Adult Success Washington D.C.: Brookings Institution, 1996, pp. 97-140.

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

Folger, John. (1989). "Editor's Introduction: Project STAR and Class Size Policy." Peabody Journal of Education 67, no. 1, Fall, pp. 1-16.

Folger, John and Carolyn Breda. (1989). "Evidence From Project STAR About Class Size and Student Achievement." Peabody Journal of Education 67, no. 1, Fall, pp. 17-33."

Griliches, Zvi and William Mason. (1972). "Education Income and Ability." Journal of Political Economy 80: S37-S64.

Hanushek, Eric A. (1986). "The Economics of Schooling: Production and Efficiency in Public Schools." Journal of Economic Literature 24 (September): 1141-1177.

Hanushek, Eric A. and Lori L. Taylor. (1990). "Alternative Assessments of the Performance of Schools: Measurement of State Variations in Achievement." Journal of Human Resources 25 (Spring): 179-201.

Hanushek, Eric. (1996). "A More Complete Picture of School Resource Policies." Review of Educational Research 66, no. 3, (Fall): 397-409.

Hedges, Larry V., Richard Laine and Rob Greenwald (1994). "Does Money Matter? A Meta-Analysis of Studies of the Effects of Differential School Inputs on Student Outcomes." Education Researcher 23, no. 3 (April): 5-14.

Mosteller, Frederick. (1995). "The Tennessee Study of Class Size in the Early School Grades." The Future of Children: Critical Issues for Children and Youths 5, (Summer/Fall): 113-27.

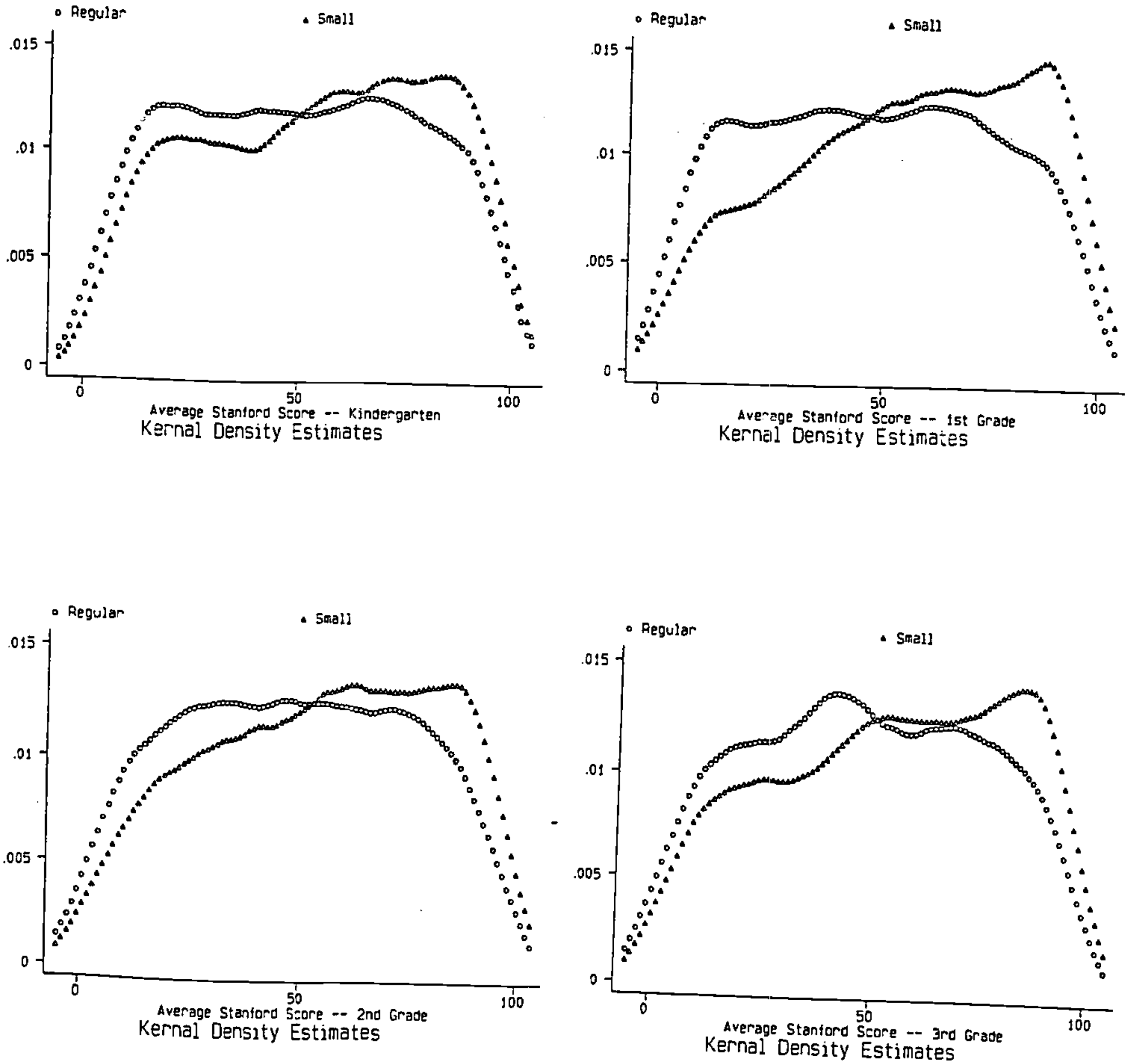
Nye, Barbara, Jayne Zaharias, B.D. Fulton, et al. (1994). "The Lasting Benefits Study: A continuing analysis of the effect of small class size in kindergarten through third grade on student achievement test scores in subsequent grade levels." Seventh grade technical report. Nashville: Center of Excellence for Research in Basic Skills, Tennessee State University, 1994.

Summers, Anita A., and Barbara L. Wolfe. (1977). "Do Schools Make a Difference?" The American Economic Review, 67, no. 4 (September): 639-652.

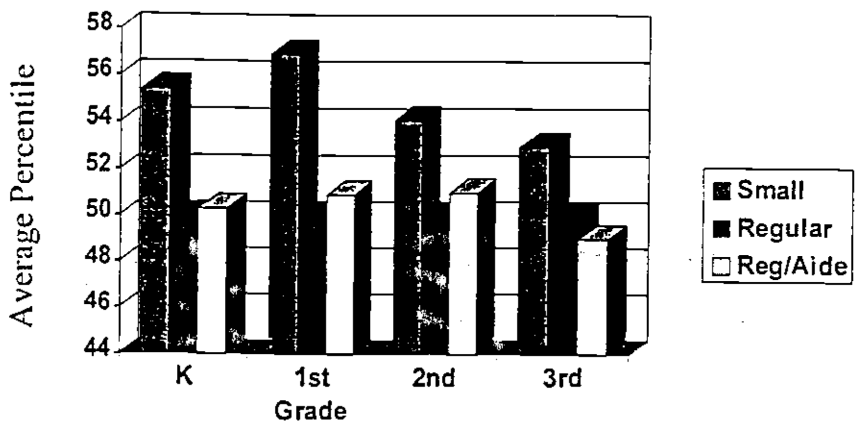
Word, Elizabeth, J. Johnston, Helen Bain, et al. (1994). "The State of Tennessee's Student/Teacher Achievement Ratio (STAR) Project: Technical Report 1985-1990." Nashville: Tennessee State Department of Education.

Figure 1

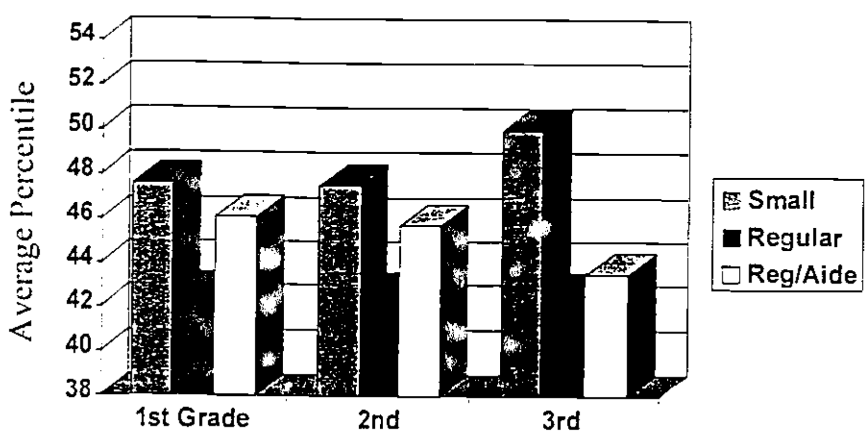
Distribution of Test Percentile Scores by Class-Size and Grade



Sample Starting in Kindergarten



Sample Starting in 1st Grade



Sample Starting in 2nd Grade

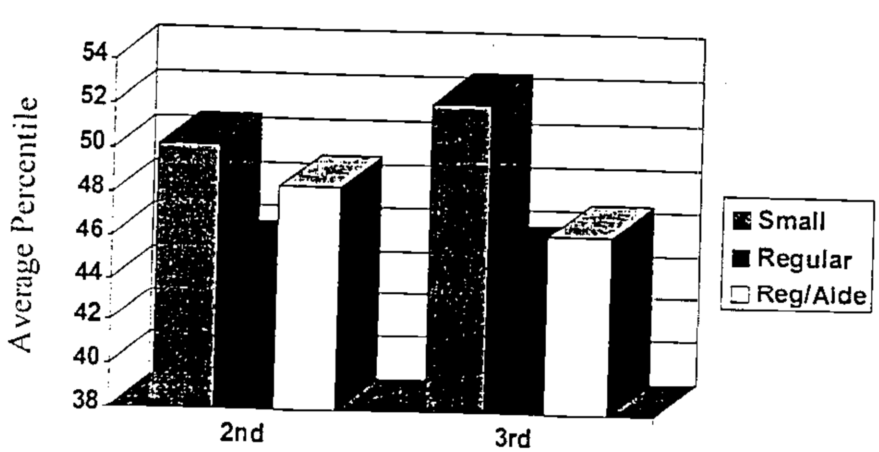
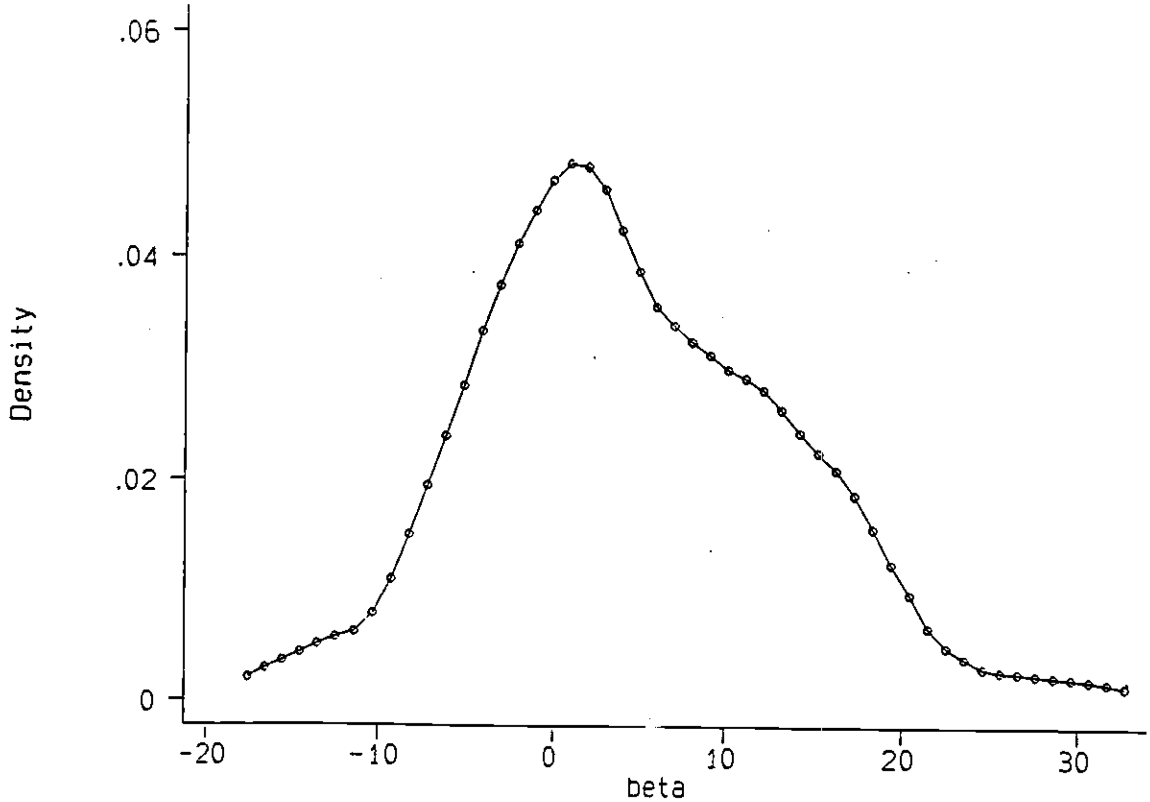


Table 6, column 4 model.

Figure 3



Kernel Density Estimate

Table 1: Comparison of Mean Characteristics of Treatments and Controls; Unadjusted Data

A. Students Who Entered STAR in Kindergarten

Variable	Small	Regular	Regular/Aide	Joint P-Value[a]
1. Free Lunch	.47	.48	.50	.09
2. White/Asian	.68	.67	.66	.26
3. Age in 1985	5.44	5.43	5.42	.32
4. Attrition Rate	.49	.52	.53	.02
5. Class Size in Kindergarten	15.1	22.4	22.8	.00
6. Percentile Score in Kindergarten	54.7	49.9	50.0	.00

B. Students Who Entered STAR in First Grade

Variable	Small	Regular	Regular/Aide	Joint P-Value[a]
1. Free Lunch	.59	.62	.61	.52
2. White/Asian	.62	.56	.64	.00
3. Age in 1985	5.78	5.86	5.88	.03
4. Attrition Rate	.53	.51	.47	.07
5. Class Size in First Grade	15.9	22.7	23.5	.00
6. Percentile Score in First Grade	49.2	42.6	47.7	.00

-- Continued --

Table 1 -- Continued

C. Students Who Entered STAR in Second Grade				
1. Free Lunch	.66	.63	.66	.60
2. White/Asian	.53	.54	.44	.00
3. Age in 1985	5.94	6.00	6.03	.66
4. Attrition Rate	.37	.34	.35	.58
5. Class Size in Third Grade	15.5	23.7	23.6	.01
6. Percentile Score in Second Grade	46.4	45.3	41.7	.01
D. Students Who Entered STAR in Third Grade				
1. Free Lunch	.60	.64	.69	.04
2. White/Asian	.66	.57	.55	.00
3. Age in 1985	5.95	5.92	5.99	.39
4. Attrition Rate	NA	NA	NA	NA
5. Class Size in Third Grade	16.0	24.1	24.4	.01
6. Percentile Score in Third Grade	47.6	44.2	41.3	.01

Notes:

a. p-value is for F-test of equality of all three groups.

b. Sample size in panel A ranges from 6299 to 6324, in panel B ranges from 2240 to 2314, in panel C ranges from 1585 to 1679, and in panel D ranges from 1202 to 1283.

c. Free lunch pertains to the fraction receiving a free lunch in the first year they are observed in the sample (i.e., in kindergarten for panel A; in first grade in panel B; etc.) Percentile score pertains to the average percentile score on the three Stanford Achievement Tests the students took in the first year they are observed in the sample.

d. Attrition rate is the fraction that ever exits the sample prior to completing third grade, even if they return to the sample in a subsequent year. Attrition rate is unavailable in third grade.

Table 2: P-values for Tests of Within School Differences Among Small, Regular, and Regular/Aide Classes

Variable	Grade Entered STAR Program			
	K	1	2	3
1. Free Lunch	.46	.29	.58	.18
2. White/Asian	.66	.28	.15	.21
3. Age	.38	.12	.48	.40
4. Attrition Rate	.01	.07	.58	NA
5. Actual Class Size	.00	.00	.00	.00
6. Percentile Score	.00	.00	.46	.00

Notes:

Each p-value is for an F-test of the null hypothesis that assignment to a small, regular, or regular/aide class has no effect on the outcome variable in that grade, conditional on school of attendance.

All rows except 4 pertain to the first grade in which the student entered the STAR program. The attrition rate in row 4 measures whether the student ever left the sample after initially being observed.

Table 3: Distribution of Children Across Actual Class Sizes by Random Assignment Group in First Grade

Actual Class Size in First Grade	Assignment Group in First Grade		
	Small	Regular	Aide
12	24	0	0
13	182	0	0
14	252	0	0
15	465	0	0
16	256	16	0
17	561	17	0
18	108	36	0
19	57	76	57
20	20	200	120
21	0	378	378
22	0	594	329
23	0	437	460
24	0	384	264
25	0	175	225
26	0	130	234
27	0	54	108
28	0	28	56
29	0	29	58
30	0	30	30
Average Class Size	15.7	22.7	23.4

Note: Actual class was determined by counting the number of students in the data set with the same class identification.

Table 4: Transitions Between Class-Size in Adjacent Grades

Number of Students in Each Type of Class

A. Kindergarten to First Grade

	First Grade			
Kindergarten	Small	Regular	Reg/Aide	All
Small	1292	60	48	1400
Regular	126	737	663	1526
Aide	122	761	706	1589
All	1540	1558	1417	4515

B. First Grade to Second Grade

	Second Grade			
First Grade	Small	Regular	Reg/Aide	All
Small	1435	23	24	1482
Regular	152	1498	202	1852
Aide	40	115	1560	1715
All	1627	1636	1786	5049

C. Second Grade to Third Grade

	Third Grade			
Second Grade	Small	Regular	Reg/Aide	All
Small	1564	37	35	1636
Regular	167	1485	152	1804
Aide	40	76	1857	1973
All	1771	1598	2044	5413

Table 5: Correlation Between Subject Tests, First and Second Graders

	First Grade					Second Grade				
	Math-S	Read-S	Word-S	Math-B	Read-B	Math-S	Read-S	Word-S	Math-B	Read-B
First Grade										
Math-S	1.00									
Reading-S	0.69	1.00								
Word-S	0.62	0.92	1.00							
Math-B	0.71	0.56	0.50	1.00						
Reading-B	0.59	0.75	0.69	0.57	1.00					
Second Grade										
Math-S	0.71	0.62	0.55	0.59	0.52	1.00				
Reading-S	0.60	0.77	0.70	0.56	0.65	0.73	1.00			
Word-S	0.54	0.74	0.67	0.45	0.61	0.62	0.88	1.00		
Math-B	0.60	0.53	0.47	0.53	0.47	0.77	0.66	0.56	1.00	
Reading-B	0.56	0.66	0.59	0.47	0.59	0.68	0.80	0.71	0.65	1.00

Notes: A "-S" indicates the Stanford Achievement Test (SAT) and a "-B" indicates the Tennessee Basic Skills First (BSF) Test. Sample size is 3,999.

Table 6: OLS and Reduced Form Estimates of Effect of Class-Size Assignment on Avg. Percentile of Stanford Achievement Test

Explanatory Variable	OLS: Actual Class Size (1)	(2)	(3)	(4)	Reduced Form: Initial Class Size (5)	(6)	(7)	(8)
A. Kindergarten								
Small Class	4.82 (2.19)	5.37 (1.26)	5.36 (1.21)	5.37 (1.19)	4.82 (2.19)	5.37 (1.25)	5.36 (1.21)	5.37 (1.19)
Regular/Aide Class	.12 (2.23)	.29 (1.13)	.53 (1.09)	.31 (1.07)	.12 (2.23)	.29 (1.13)	.53 (1.09)	.31 (1.07)
White/Asian (1=yes)	....	....	8.35 (1.35)	8.44 (1.36)	....	....	8.35 (1.35)	8.44 (1.36)
Girl (1=yes)	....	....	4.48 (.63)	4.39 (.63)	....	....	4.48 (.63)	4.39 (.63)
Free Lunch (1=yes)	....	....	-13.15 (.77)	-13.07 (.77)	....	....	-13.15 (.77)	-13.07 (.77)
White Teacher	....	....	....	-.57 (2.10)	....	....	....	-.57 (2.10)
Teacher Experience	....	....	....	.26 (.10)	....	....	....	.26 (.10)
Masters Degree	....	....	....	-.51 (1.06)	....	....	....	-.51 (1.06)
School Fixed Effects	No	Yes	Yes	Yes	No	Yes	Yes	Yes
R^2	.01	.25	.31	.31	.01	.25	.31	.31
B. First Grade								
Small Class	8.57 (1.97)	8.43 (1.21)	7.91 (1.17)	7.40 (1.18)	7.54 (1.76)	7.17 (1.14)	6.79 (1.10)	6.37 (1.11)
Regular/Aide Class	3.44 (2.05)	2.22 (1.00)	2.23 (0.98)	1.78 (0.98)	1.92 (1.12)	1.69 (0.80)	1.64 (0.76)	1.48 (0.76)
White/Asian (1=yes)	....	....	6.97 (1.18)	6.97 (1.19)	....	....	6.86 (1.18)	6.85 (1.18)
Girl (1=yes)	....	....	3.80 (.56)	3.85 (.56)	....	....	3.76 (.56)	3.82 (.56)
Free Lunch (1=yes)	....	....	-13.49 (.87)	-13.61 (.87)	....	....	-13.65 (.88)	-13.77 (.87)
White Teacher	....	....	....	-4.28 (1.96)	....	....	....	-4.40 (1.97)
Male Teacher	....	....	....	11.82 (3.33)	....	....	....	13.06 (3.38)
Teacher Experience	....	....	....	.05 (0.06)	....	....	....	.06 (.06)
Masters Degree	....	....	....	.48 (1.07)	....	....	....	.63 (1.09)
School Fixed Effects	No	Yes	Yes	Yes	No	Yes	Yes	Yes
R^2	.02	.24	.30	.30	.01	.23	.29	.30

-- Continued --

Table 6 -- Continued

Explanatory Variable	OLS: Actual Class Size (1)	(2)	(3)	(4)	Reduced Form: Initial Class Size (5)	(6)	(7)	(8)
				C. Second Grade				
Small Class	5.93 (1.97)	6.33 (1.29)	5.83 (1.23)	5.79 (1.23)	5.31 (1.70)	5.52 (1.16)	5.27 (1.10)	5.26 (1.10)
Regular/Aide Class	1.97 (2.05)	1.88 (1.10)	1.64 (1.07)	1.58 (1.06)	.47 (1.23)	1.44 (0.87)	1.16 (0.81)	1.18 (0.81)
White/Asian (1=yes)	....	....	6.35 (1.20)	6.36 (1.19)	....	....	6.27 (1.21)	6.29 (1.20)
Girl (1=yes)	....	....	3.48 (.60)	3.45 (.60)	....	....	3.48 (.60)	3.44 (.60)
Free Lunch (1=yes)	....	....	-13.61 (.72)	-13.61 (.72)	....	....	-13.75 (.73)	-13.77 (.73)
White Teacher	....	....	....	.39 (1.75)	....	....	....	.43 (1.76)
Male Teacher	....	....	....	1.32 (3.96)	....	....	....	.82 (4.23)
Teacher Experience	....	....	....	.10 (.06)	....	....	....	.10 (.07)
Masters Degree	....	....	....	-1.06 (1.06)	....	....	....	-1.16 (1.05)
School Fixed Effects	No	Yes	Yes	Yes	No	Yes	Yes	Yes
R^2	.01	.22	.28	.28	.01	.21	.28	.28
				D. Third Grade				
Small Class	5.32 (1.91)	5.58 (1.22)	5.01 (1.19)	5.00 (1.19)	5.51 (1.46)	5.42 (1.08)	5.30 (1.03)	5.24 (1.04)
Regular/Aide Class	-.22 (1.95)	-.16 (1.12)	-.33 (1.11)	-.75 (1.07)	-.30 (1.17)	.12 (0.85)	.13 (0.81)	-.10 (0.78)
White/Asian (1=yes)	....	....	6.12 (1.45)	6.11 (1.44)	....	....	5.97 (1.44)	5.96 (1.43)
Girl (1=yes)	....	....	4.16 (.66)	4.16 (.65)	....	....	4.17 (.66)	4.18 (.66)
Free Lunch (1=yes)	....	....	-13.02 (.81)	-12.96 (.81)	....	....	-13.21 (.82)	-13.16 (.81)
White Teacher	....	....	....	.64 (1.75)	....	....	....	.19 (1.75)
Male Teacher	....	....	....	-7.42 (2.80)	....	....	....	-6.83 (2.76)
Teacher Experience	....	....	....	.04 (.06)	....	....	....	.03 (.06)
Masters Degree	....	....	....	1.10 (1.15)	....	....	....	.88 (1.15)
School Fixed Effects	No	Yes	Yes	Yes	No	Yes	Yes	Yes
R^2	.01	.17	.22	.23	.01	.16	.22	.22

Notes: All models include constants. Robust standard errors that allow for correlated residuals among students in the same class are in parentheses. Sample size is 5,861 for kindergarten, 6,452 for first grade, 5,950 for second grade, and 6,109 for third grade.

Table 7: Exploration of Effect of Attrition

Dependent Variable: Average Percentile Score on SAT

Grade	Actual Test Data: Coefficient on Small Class Dum.	Actual Test Data: Sample Size	Actual And Imputed Test Data: Coefficient on Small Class Dum.	Actual And Imputed Test Data: Sample Size
K	5.32 (.76)	5,900	5.32 (.76)	5,900
1	6.95 (.74)	6,632	6.30 (.68)	8,328
2	5.39 (.76)	6,282	5.64 (.65)	9,773
3	5.58 (.79)	6,339	5.49 (.63)	10,919

Notes: Estimates of reduced form models are presented. Each regression includes the following explanatory variables: a dummy variable indicating initial assignment to a small class; a dummy variable indicating initial assignment to a regular/aide class, unrestricted school effects, a dummy variable for student gender, and a dummy variable for student race. The reported coefficient on small class dummy is relative to regular classes. Standard errors are in parentheses.

Table 8: OLS and 2SLS Estimates of Effect of Class Size on Achievement

Dependent Variable: Average Percentile Score on SAT

Grade	OLS (1)	2SLS (2)	Sample Size (3)
K	-.62 (.14)	-.71 (.14)	5,861
1	-.85 (.13)	-.88 (.16)	6,452
2	-.59 (.12)	-.67 (.14)	5,950
3	-.61 (.13)	-.81 (.15)	6,109

Notes: The coefficient on the actual number of students in each class is reported. All models also control for school effects; student's race, gender, and free lunch status; teacher race, experience and education. Robust standard errors that allow for correlated errors among students in the same class are reported in parentheses.

Table 9: 2SLS Estimates of Effect of Class Size on Achievement, by Entry Grade and Current Grade

Dependent Variable: Average Score on Stanford Achievement Test

	Entering Grade			
Current Grade	K	1	2	3
K	-.71 (.15)			
1	-.89 (.17)	-.49 (.23)		
2	-.49 (.16)	-.70 (.29)	-.24 (.21)	
3	-.66 (.17)	-1.21 (.34)	-.71 (.28)	-.66 (.21)

Note: The coefficient on the actual number of students in each class is reported. All models also control for school effects; student's race, gender, and free lunch status; teacher race, experience and education. Robust standard errors that allow for correlation of residuals among students in the same class are reported in parentheses. Sample size in column 1 begins at 5,901 and ends at 3,124; sample size in column 2 begins at 2,190 and ends at 1,110; sample size in column 3 begins at 1,492 and ends at 1,010; sample size in column 4 is 1,110.

Table 10: Estimates of Pooled Models

Dependent Variable: Average Percentile Ranking on SAT Test

Coefficient Estimates with Robust Standard Errors in Parentheses

Variable	OLS (1)	OLS (2)	OLS (3)	2SLS (4)	2SLS (5)	2SLS (6)
Small	4.40 (.51)	4.28 (.50)	3.55 (.52)	4.91 (.63)	4.92 (.62)	4.30 (.65)
Regular/Aide	.85 (.42)	.75 (.42)	.67 (.42)	.73 (.59)	.66 (.58)	.60 (.58)
Cumulative Years in Small Class	.89 (.38)	.93 (.37)	.67 (.38)	.87 (.47)	.97 (.47)	.76 (.47)
Cumulative Years in Reg/Aide Class	.26 (.40)	.29 (.39)	.19 (.39)	.60 (.65)	.78 (.63)	.76 (.63)
Fraction of classmates in class previous year	---	---	.80 (1.03)	---	---	.64 (1.04)
Average fraction of classmates together previous year	---	---	-.36 (1.51)	---	---	-.27 (1.52)
Fraction of classmates on free lunch	---	---	-2.05 (1.61)	---	---	-2.00 (1.61)
Fraction of classmates who attended kindergarten	---	---	6.42 (1.68)	---	---	5.58 (1.73)
Student and teacher characteristics	No	Yes	Yes	No	Yes	Yes
3 current grade dummies; 3 dummies indicating first grade appeared in sample; school effects	Yes	Yes	Yes	Yes	Yes	Yes
R^2	.18	.23	.23	.18	.23	.23
Sample Size	25,249	24,350	24,350	25,249	24,350	24,350

Notes:

Student and teacher characteristics are: student race, gender, and free lunch status; and teacher race, gender, experience, and masters degree or higher status. Robust standard errors that adjust for a possible correlation of residuals for the same student over time are in parentheses. Initial assignment to class types and potential cumulative years in class types are used as excluded instruments in columns 4-6.

Table 11: Separate Estimates for Select Samples
Dependent Variable: Average Percentile Ranking on SAT Test
Coefficient Estimates with Robust Standard Errors in Parentheses

	Boys	Girls
Small	3.64	3.46
	(.76)	(.73)
Cumulative Years in Small Class	1.30	.12
	(.54)	(.52)
Sample Size	12,577	11,773

	Free Lunch	Not Free Lunch
Small	3.64	3.37
	(.73)	(.73)
Cumulative Years in Small Class	1.17	.59
	(.57)	(.49)
Sample Size	12,064	12,286

	Black	White
Small	5.07	2.87
	(.85)	(.66)
Cumulative Years in Small Class	.94	.86
	(.67)	(.46)
Sample Size	8,150	16,070

	Inner City	Metropolitan	Towns	Rural
Small	6.80	2.65	4.21	3.04
	(1.06)	(1.06)	(1.85)	(.81)
Cumulative Years in Small Class	.99	1.24	-3.48	1.34
	(.87)	(.82)	(1.48)	(.52)
Sample Size	5,154	5,907	1,872	11,417

Notes: Model and covariates are the same as column 3 of Table 10.

Table 12: Effect of Class Size for Students in Regular Classes

Dependent Variable: Average Percentile Score on SAT

Explanatory Variable	(1)	(2)
Class size	-.55 (.13)	-.39 (.13)
White	10.62 (.85)	7.58 (1.28)
Girl	4.23 (.70)	3.96 (.66)
Free Lunch	-14.41 (.77)	-13.72 (.77)
White Teacher	-.25 (.80)	-2.95 (.84)
Male Teacher	-6.59 (2.12)	-4.68 (2.29)
Teacher Experience	.07 (.03)	.01 (.03)
Masters Degree or Higher	-.05 (.58)	.42 (.58)
3 Grade Dummies	Yes	Yes
80 School Dummies	No	Yes
R-square	.17	.25

Notes: All models also include a constant.

Sample size is 8,311. Sample consists of students in regular-size classes without an aide.

Table 13: Estimates of Pooled Data Model by Subject Test

Dependent Variable: Average Percentile Ranking on SAT Test
Coefficient Estimates with Robust Standard Errors in Parentheses

	Stanford Achievement Test			Basic Skills First		
	Math	Reading	Word	Math	Reading	Avg.
Small	3.44 (.57)	3.67 (.58)	3.74 (.58)	3.91 (.77)	4.16 (.77)	4.01 (.69)
Cumulative Years in Small Class	.35 (.41)	.44 (.42)	.77 (.41)	.41 (.45)	.07 (.45)	.25 (.41)
Sample Size	23,794	23,461	23,631	18,175	18,011	18,251

Notes: Model and covariates are the same as in column 3 of Table 10.

Appendix Table: Summary Statistics
Means with Standard Deviations in Parentheses

Variable	Grade K	1	2	3	All
Class Size	20.3 (4.0)	21.0 (4.0)	21.1 (4.1)	21.3 (4.4)	20.9 (4.1)
Percentile Score Avg. SAT	51.4 (26.7)	51.5 (26.9)	51.2 (26.5)	51.0 (27.0)	51.3 (26.8)
Percentile Score Avg. BSF	NA	51.8 (26.1)	51.6 (26.2)	51.4 (26.1)	51.6 (26.1)
Free Lunch	.48	.52	.51	.50	.51
White	.67	.67	.65	.66	.66
Girl	.49	.48	.48	.48	.47
Age[a]	5.43 (0.35)	6.58 (0.49)	7.67 (0.56)	8.70 (0.59)	7.12 (1.31)
Exited Sample[b]	.29	.26	.21	NA	.43
Retained	NA	NA	NA	.04	NA
Percent of Teachers with MA+ degree	.35	.35	.37	.44	.38
Percent of Teachers who are White	.83	.82	.79	.79	.81
Percent of Teachers who are Male	.00	.00	.01	.03	.01
No. of Schools	79	76	75	75	80
No. of Students	6,323	6,828	6,839	6,801	11,599
No. of Small Classes	127	124	133	140	524
No. of Reg. Classes	99	115	100	89	403
No. of Reg/A Classes	99	100	107	107	413

a. Age as of September of the school year they are observed.

b. The fraction that exited the sample in the next year, for K-2; for All it is the fraction that ever exited the sample.

c. Teacher characteristics are weighted by the number of students in each teacher's class.