EXPERIMENTAL ESTIMATES OF EDUCATION PRODUCTION FUNCTIONS*

ALAN B. KRUEGER

This paper analyzes data on 11,600 students and their teachers who were randomly assigned to different size classes from kindergarten through third grade. Statistical methods are used to adjust for nonrandom attrition and transitions between classes. The main conclusions are (1) on average, performance on standardized tests increases by four percentile points the first year students attend small classes; (2) the test score advantage of students in small classes expands by about one percentile point per year in subsequent years; (3) teacher aides and measured teacher characteristics have little effect; (4) class size has a larger effect for minority students and those on free lunch; (5) *Hawthorne* effects were unlikely.

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

* I thank Helen Bain, a founder and principal director of Project STAR, for providing me with the data used in this study, Jayne Zaharias, DeWayne Fulton, and Van Cain for answering several questions regarding the data, and Jessica Baraka, Aaron Saiger, and Diane Whitmore for providing outstanding research assistance. The STAR data have been collected and maintained by the Center of Excellence for Research in Basic Skills at Tennessee State University. The STAR data are available from www.telalink.net/~heros. Helpful comments on my research were provided by Charles Achilles, Jessica Baraka, Ronald Ehrenberg, William Evans, Jeremy Finn, John Folger, Victor Fuchs, Joseph Hotz, Lawrence Katz, Cecilia Rouse, James P. Smith, two referees, and seminar participants at the Milken Institute, Massachusetts Institute of Technology, National Bureau of Economic Research, Princeton University, Vanderbilt University, University of California at Los Angeles, the Kennedy School (Harvard University), the London School of Economics, Stockholm University, the Econometric Society, World Bank, and Society of Labor Economists. Financial support was provided by the National Institute of Childhood Health and Development.

The Quarterly Journal of Economics, May 1999

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 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 affect the trajectory of student performance). A classical experiment, in which students are randomly assigned to classes with different resources, would help overcome many of these specification 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 1985–1986 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 6000–7000 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

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 students' educational attainment and subsequent earnings.

2. Hanushek attributes this difference to omitted state-level variables that bias the state-level 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 micro-data estimates, and that the interstate estimates are unbiased.

were given a battery of standardized tests at the end of each school year. 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."

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 this research, little attention has been paid to potential threats to the validity of the experiment or to the longitudinal structure of the data.

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

3. If a school had more than one small class, students could be moved between small classes.

longitudinal studies of schooling, sample attrition was common—half of students who were present in kindergarten were missing in at least one subsequent year. And some students may have nonrandomly switched to another public school or enrolled in private school upon learning their class-type assignments. These limitations of the experiment have not been adequately 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. Second, to use the experiment to identify an appropriate specification of the education production function to estimate with nonexperimental data. Third, to use the experimental results to interpret estimates from the literature based on observational data. The conclusion makes a rough attempt to compare the benefits and costs of reducing class size from 22 to 15 students.

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. 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), regular class (22–25 students), or regular plus a full-time aide class.[5] The statewide pupil-teacher ratio in kindergarten in 1985–1986 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

4. This section draws heavily from Word et al. [1990] and Folger [1989].
5. Participating schools had an average per-pupil expenditure in 1986–1987 of $2724, compared with the statewide average of $2561.

percent of the time on average, so the variability in aide services was restricted.[6]

The cohort of students who entered kindergarten in the 1985–1986 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 repeated a grade, skipped a grade, or left the school exited the sample. Entering students were randomly assigned to one of the three types of classes (small, regular, or regular/aide) in the summer before they began kindergarten.[7] Students were typically notified of their initial class assignment very close to the beginning of the school year. Students in regular classes and in regular/aide classes were randomly reassigned 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 that included a participating school. In all, some 2200 new students entered the project in first grade and were randomly assigned to the three types of classes. Another 1600 and 1200 new students entered the experiment in the second and third grades, respectively. 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 ninth grade, the present study only has access to data covering grades K–3. Data were

6. 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.

7. The procedure for randomly assigning students was as follows. Each school prepared an alphabetized enrollment list. Algorithms were centrally prepared which assigned every kth student to a class type; the algorithm was tailored to the number of enrolled students. A random starting point was used by each school to apply the algorithm. The schools were audited to ensure that they followed procedures for random assignment.

collected on students each fall and spring during the experiment. Class type is based on the class attended in the fall. All students who attended a STAR class in either the fall or spring are included in the database.

Unfortunately, the STAR data set does not contain students' original class type assignments resulting from the randomization procedure; only the class types that students actually were enrolled in each year are available. It is possible that some students were switched from their randomly assigned class to another class before school started or early in the fall. To determine the frequency of such switches, we obtained and (double) entered data on the initial random assignments from the actual enrollment sheets that were compiled in the summer prior to the start of kindergarten for 1581 students from 18 participating STAR schools.[8] It turns out that only 0.3 percent of students in the experiment were *not* enrolled in the class type to which they were randomly assigned in kindergarten. Moreover, only one student in this sample who was assigned a regular or regular/aide class enrolled in a small class. Consequently, in the analysis below, we will refer to the class type in which students are enrolled during the first year they enter the experiment as their initial random assignment.

A limitation of the experiment 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. Nonetheless, if the students were successfully randomly assigned between class types, one would expect those assigned to small- and regular-size classes to look similar along other measurable dimensions at base line. Tables I and II provide some evidence on the differences among students assigned to the three class types.

Table I 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. The sample consists of all students who were enrolled in a STAR class when the fall or spring data were collected. 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

8. I thank Jayne Zaharias for providing the enrollment sheets. The sample I analyze excludes twins; schools were allowed to assign twins to the same class if that was the school's ordinary practice.

TABLE I
COMPARISON OF MEAN CHARACTERISTICS OF TREATMENTS AND CONTROLS: UNADJUSTED DATA

Variable	Small	Regular	Regular/Aide	Joint *P*-Value[a]
A. Students who entered STAR in kindergarten[b]				
1. Free lunch[c]	.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[d]	.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				
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
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

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 II
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

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.

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).[9] Because random assignment was only valid within schools, these differences suggest the importance of controlling for school effects as is done in Table II.

Table II presents *p*-values for joint *F*-tests of the differences among small, regular, and regular/aide classes for the variables presented in Table I. Unlike results reported in Table I, 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 similar 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

9. To be precise, the fraction on free lunch actually measures the fraction who receive free or reduced-price lunch.

is no apparent evidence that initial assignment to class types was correlated with student characteristics.

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 was a total of twelve regressions. In each case, the p-value for the class-type dummies exceeded .05.[10] These results are as one would expect with random assignment of teachers to the different class types.

There was a high rate of attrition from the project. Only half the students who entered the project in kindergarten were present for all grades K–3. For the kindergarten cohort, students in small classes were three–four 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, students repeating a grade, and students being advanced 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.

It is virtually impossible to prescribe 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 III 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.

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

TABLE III
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

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

It is also virtually impossible to prevent some students from switching between class types over time. Table IV 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.

TABLE IV
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

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.

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 percentile ranks. 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 I 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 scores. Finally, to summarize overall achievement, the *average* of the three SAT percentile rankings was calculated.[11] 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.

An examination of the correlations among the tests indicates 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. Tests of the same subjects tend to have a higher correlation from one grade to the next than tests of different subjects. The SAT and BSF tests are also highly correlated with each other: the correlation between the average SAT percentile and average BSF percentile 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 main findings are similar for the BSF test, however.

The average of the three SAT exams by class type is presented in the last row of Table I. Figure I displays the kernel density of the average test score distributions for students in small and regular classes at each grade level.[12] 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 findings.

11. Formally, denote the cumulative distribution of scores on test j (denoted T^j) of students in the regular and regular/aide classes as $F^R(T^j) = \text{prob}\,[T^j_{Ri} < T^j] = y^j$. For each student i in a small class, we then calculated $F^R(T^j_{Si}) = y^j_{Si}$. 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.

12. Note that because we have averaged over three percentile scores, the distributions are not uniform for students assigned to regular classes.

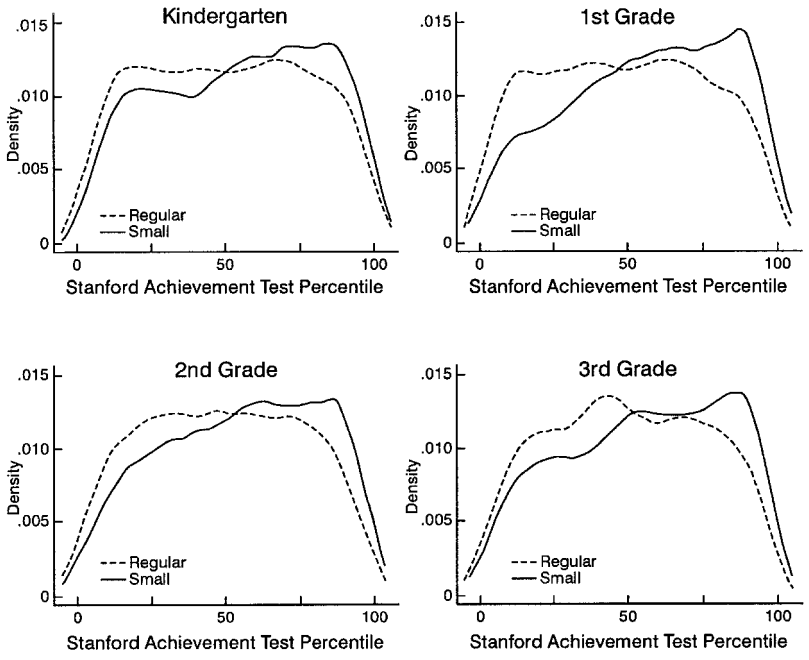

FIGURE I
Distribution of Test Percentile Scores by Class Size and Grade

Observe also that the average test score of students in all class types tends to be lower for those who entered the experiment in higher grades. This pattern 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, I control for the grade in which the student entered Project STAR in some of the analysis below.

The Appendix presents means for several 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:

$$Y_{ij} = aS_{ij} + bF_{ij} + \varepsilon_{ij}, \tag{1}$$

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 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:

$$(2) \quad Y_{ics} = \beta_0 + \beta_1 SMALL_{cs} + \beta_2 REG/A_{cs} + \beta_3 X_{ics} + \alpha_s + \varepsilon_{ics},$$

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 within 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 modeled in a components-of-variance framework. Specifically, ε_{ics} is assumed to consist of two components: $\varepsilon_{ics} = \mu_{cs} + \varepsilon'_{ics}$, 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. 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 reassigned 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 labeled "reduced-form" models, because one can think of initial assignment as an excluded variable that is correlated with actual class size. The initial assignment and actual assignment variables are identical in kindergarten, so the OLS and reduced-form estimates are identical for kindergarten students.

Regression results are presented in Table V.[13] 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 eight percentile points higher than black students in kindergarten, and this gap is about six points in third grade.[14] Students on free lunch score thirteen percentile points less than those not on free lunch, and girls score three–four points higher than boys in each grade level.

The teacher characteristics have notably weak explanatory power. Teacher education—as proxied by a dummy indicating

13. The robust standard errors are about two-thirds larger than the OLS standard errors. The estimated standard deviation of the class effects (μ_{cs}) is about 8 in the models in column 4.

14. 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.

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 meaningful. Teacher experience has a small, positive effect. Experimentation with a quadratic in experience indicated that the experience profile tends to peak at about

TABLE V
OLS AND REDUCED-FORM ESTIMATES OF EFFECT OF CLASS-SIZE ASSIGNMENT ON AVERAGE PERCENTILE OF STANFORD ACHIEVEMENT TEST

Explanatory variable	OLS: actual class size				Reduced form: initial class size			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
A. Kindergarten								
Small class	4.82	5.37	5.36	5.37	4.82	5.37	5.36	5.37
	(2.19)	(1.26)	(1.21)	(1.19)	(2.19)	(1.25)	(1.21)	(1.19)
Regular/aide class	.12	.29	.53	.31	.12	.29	.53	.31
	(2.23)	(1.13)	(1.09)	(1.07)	(2.23)	(1.13)	(1.09)	(1.07)
White/Asian (1 = yes	—	—	8.35	8.44	—	—	8.35	8.44
			(1.35)	(1.36)			(1.35)	(1.36)
Girl (1 = yes)	—	—	4.48	4.39	—	—	4.48	4.39
			(.63)	(.63)			(.63)	(.63)
Free lunch (1 =	—	—	−13.15	−13.07	—	—	−13.15	−13.07
yes)			(.77)	(.77)			(.77)	(.77)
White teacher	—	—	—	−.57	—	—	—	−.57
				(2.10)				(2.10)
Teacher experience	—	—	—	.26	—	—	—	.26
				(.10)				(.10)
Master's degree	—	—	—	−.51	—	—	—	−.51
				(1.06)				(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	8.43	7.91	7.40	7.54	7.17	6.79	6.37
	(1.97)	(1.21)	(1.17)	(1.18)	(1.76)	(1.14)	(1.10)	(1.11)
Regular/aide class	3.44	2.22	2.23	1.78	1.92	1.69	1.64	1.48
	(2.05)	(1.00)	(0.98)	(0.98)	(1.12)	(0.80)	(0.76)	(0.76)
White/Asian (1 =	—	—	6.97	6.97	—	—	6.86	6.85
yes)			(1.18)	(1.19)			(1.18)	(1.18)
Girl (1 = yes)	—	—	3.80	3.85	—	—	3.76	3.82
			(.56)	(.56)			(.56)	(.56)
Free lunch (1 =	—	—	−13.49	−13.61	—	—	−13.65	−13.77
yes)			(.87)	(.87)			(.88)	(.87)
White teacher	—	—	—	−4.28	—	—	—	−4.40
				(1.96)				(1.97)
Male teacher	—	—	—	11.82	—	—	—	13.06
				(3.33)				(3.38)
Teacher experience	—	—	—	.05	—	—	—	.06
				(0.06)				(.06)
Master's degree	—	—	—	.48	—	—	—	.63
				(1.07)				(1.09)
School fixed effects	No	Yes	Yes	Yes	No	Yes	Yes	Yes
R^2	.02	.24	.30	.30	.01	.23	.29	.30

TABLE V
(CONTINUED)

Explanatory variable	OLS: actual class size				Reduced form: initial class size			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
C. Second grade								
Small class	5.93	6.33	5.83	5.79	5.31	5.52	5.27	5.26
	(1.97)	(1.29)	(1.23)	(1.23)	(1.70)	(1.16)	(1.10)	(1.10)
Regular/aide class	1.97	1.88	1.64	1.58	.47	1.44	1.16	1.18
	(2.05)	(1.10)	(1.07)	(1.06)	(1.23)	(0.87)	(0.81)	(0.81)
White/Asian (1 = yes)	—	—	6.35	6.36	—	—	6.27	6.29
			(1.20)	(1.19)			(1.21)	(1.20)
Girl (1 = yes)	—	—	3.48	3.45	—	—	3.48	3.44
			(.60)	(.60)			(.60)	(.60)
Free lunch (1 = yes)	—	—	−13.61	−13.61	—	—	−13.75	−13.77
			(.72)	(.72)			(.73)	(.73)
White teacher	—	—	—	.39	—	—	—	.43
				(1.75)				(1.76)
Male teacher	—	—	—	1.32	—	—	—	.82
				(3.96)				(4.23)
Teacher experience	—	—	—	.10	—	—	—	.10
				(.06)				(.07)
Master's degree	—	—	—	−1.06	—	—	—	−1.16
				(1.06)				(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	5.58	5.01	5.00	5.51	5.42	5.30	5.24
	(1.91)	(1.22)	(1.19)	(1.19)	(1.46)	(1.08)	(1.03)	(1.04)
Regular/aide class	−.22	−.16	−.33	−.75	−.30	.12	.13	−.10
	(1.95)	(1.12)	(1.11)	(1.07)	(1.17)	(0.85)	(0.81)	(0.78)
White/Asian (1 = yes)	—	—	6.12	6.11	—	—	5.97	5.96
			(1.45)	(1.44)			(1.44)	(1.43)
Girl (1 = yes)	—	—	4.16	4.16	—	—	4.17	4.18
			(.66)	(.65)			(.66)	(.66)
Free lunch (1 = yes)	—	—	−13.02	−12.96	—	—	−13.21	−13.16
			(.81)	(.81)			(.82)	(.81)
White teacher	—	—	—	.64	—	—	—	.19
				(1.75)				(1.75)
Male teacher	—	—	—	−7.42	—	—	—	−6.83
				(2.80)				(2.76)
Teacher experience	—	—	—	.04	—	—	—	.03
				(.06)				(.06)
Master's degree	—	—	—	1.10	—	—	—	.88
				(1.15)				(1.15)
School fixed effects	No	Yes	Yes	Yes	No	Yes	Yes	Yes
R^2	.01	.17	.22	.23	.01	.16	.22	.22

All models include constants. Robust standard errors that allow for correlated residuals among students in the same class are in parentheses. Sample size is 5861 for kindergarten, 6452 for first grade, 5950 for second grade, and 6109 for third grade.

twenty years of experience, and students in classes where the teacher has twenty years of experience tend to score about three 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 measured teacher characteristics explain relatively little of student achievement on 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 nonrandom 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 five–seven percentage points higher than those assigned to regular-size classes. Students assigned to a regular/aide class perform slightly better (one or two 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 tangible outcomes. Nevertheless, a couple of comparisons are informative. First, relative to the standard deviation (S.D.) 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 with 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 VI 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

TABLE VI
EXPLORATION OF EFFECT OF ATTRITION DEPENDENT VARIABLE: AVERAGE PERCENTILE SCORE ON SAT

	Actual test data		Actual and imputed test data	
Grade	Coefficient on small class dum.	Sample size	Coefficient on small class dum.	Sample size
K	5.32	5900	5.32	5900
	(.76)		(.76)	
1	6.95	6632	6.30	8328
	(.74)		(.68)	
2	5.59	6282	5.64	9773
	(.76)		(.65)	
3	5.58	6339	5.49	10919
	(.79)		(.63)	

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.

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 subsequently 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.[15]

The sample used in the first panel of Table VI includes the largest number of students with nonmissing data available each grade. These results correspond to the model estimated in column 7 of Table V, except the free lunch variable is omitted because it

15. 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). This technique is closely related to the "last-observation-carry-forward" method that has been used in clinical studies.

changes over time.[16] 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 the second panel 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. To be included in the sample, it is necessary to have test data in at least some year. (Because kindergarten students could not have previously exited the sample, the sample size is the same in the first row.) The 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.[17] Thus, nonrandom attrition does not appear to bias the estimated class size effects in Table V.

The sample used in column 2 of Table VI excludes students who were listed on the enrollment logs for kindergarten but withdrew from school prior to the start of school. For example, if a parent chose to withdraw a child from a STAR public school and enroll him or her in a private school immediately upon learning that the child was assigned to a regular-size class, the student is excluded from the sample. This type of behavior appears to have been rare (based on our inspection of notes on a sampling of enrollment sheets), but 12 percent of students who were listed on the enrollment logs and assigned to a class prior to the start of school were not actually enrolled in the school the following fall. These students moved to another school zone, enrolled in private school, or were withdrawn from kindergarten over the summer for some other reason. Using data for eighteen of the participating schools for which we were able to obtain initial kindergarten enrollment sheets, we calculate that 10.4 percent of students who were listed on the enrollment sheets and assigned to small kindergarten classes were missing from our sample by the start of kindergarten; the corresponding figures for regular/aide and regular classes are 12.2 percent and 14.3 percent, respectively. The differential withdrawal rate between the regular and small classes is statistically significant ($t = 1.86$), while the difference

16. The estimated model uses initial class assignment to avoid imputing actual class size for missing observations. The sample in column 1 is a little larger than that in column 7 of Table V because Table V uses a balanced sample, and some observations were excluded due to missing data on free lunch status and teacher characteristics.

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

between the regular/aide and small class is not ($t = 0.86$). These findings suggest that 2 to 4 percent of students may have been withdrawn from the STAR schools because they were not assigned to a small class.

An upper bound estimate of the impact on test scores of the higher kindergarten withdrawal rate for students in regular and regular/aide classes can be calculated. Suppose that the 2–4 percent extra students who withdrew from regular and regular/aide classes all would have scored in the one-hundredth percentile of the SAT exams. With this intentionally extreme assumption, the average score of students in the regular and regular/aide classes would only have increased by one–two percentile points if the extra students had not withdrawn from kindergarten. At the opposite extreme, if the higher withdrawal rate is due to the lowest achieving students leaving regular-size classes, the regular-size class students would have scored one–two points lower, on average, if they had remained. The actual impact is probably even smaller, however, because the extra withdrawals probably would have scored closer to the average student if they remained in the STAR schools. But even the upper and lower bounds estimates suggest that the higher withdrawal rate from regular-size classes does not have much impact on the results.

B. Two-Stage Least Squares (2SLS) Models

As noted, students in the Project STAR experiment who were assigned to small classes had varying numbers 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 model by 2SLS:

$$CS_{ics} = \pi_0 + \pi_1 S_{ios} + \pi_2 R_{ios} + \pi_3 X_{ics} + \delta_s + \tau_{ics} \tag{3}$$

$$Y_{ics} = \beta_0 + \beta_1 CS_{ics} + \beta_2 X_{ics} + \alpha_s + \varepsilon_{ics}, \tag{4}$$

where CS_{ics} is the actual number of students in the class, S_{ios} is a dummy variable indicating assignment to a small class the first year the student is observed in the experiment, R_{ios} is a dummy variable indicating assignment to a regular class the first year the

TABLE VII
OLS AND 2SLS ESTIMATES OF EFFECT OF CLASS SIZE ON ACHIEVEMENT
DEPENDENT VARIABLE: AVERAGE PERCENTILE SCORE ON SAT

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

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.

student is observed in the experiment, and all other variables are defined as before. 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.[18]

In this setup, only variation in class size due to *initial* 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 consistent.[19] If attending a small class has a beneficial effect on students' test scores, β_1 would be negative.

Table VII presents OLS and 2SLS estimates of equation (4). The 2SLS estimates tend to be a little larger in absolute value, especially in third grade. According to the 2SLS estimates, a reduction of ten students is associated with a seven-to-nine 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.

18. Because the teacher aide was found to have a small effect in Table V, 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).

19. To interpret this model as yielding the causal effect of current class size on achievement, it is necessary to assume that initial class assignment only affects current test scores by affecting current class size. If previous class sizes affect current performance, initial assignment will be correlated with the error term in equation (4). Of course, in the kindergarten year this assumption is not controversial, but it may not hold in later grades.

Table VIII presents 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 VII, column 2. The results indicate that for each cohort of students, those attending smaller classes 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

TABLE VIII
2SLS Estimates of Effect of Class Size on Achievement, by Entry Grade and Current Grade
Dependent Variable: Average Score on Stanford Achievement Test

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

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 5901 and ends at 3124; sample size in column 2 begins at 2190 and ends at 1110; sample size in column 3 begins at 1492 and ends at 1010; sample size in column 4 is 1110.

over students and grades. The general model was of the form,

$$(5)\quad Y_{ig} = \beta_0 + \beta_1 S_{io} + \beta_2 REG/A_{io} + \beta_3 N_{ig}^S + \beta_4 N_{ig}^A + \beta_5 X_{ig} + \alpha_g + \alpha_f + \alpha_s + \varepsilon_{ig},$$

where g indicates grade level (K, 1, 2, or 3) and i indicates students, Y_{ig} is the test score, S_{io} and REG/A_{io} are dummy variables indicating a student's class type in the first year he or she participated in the program, N_{ig}^S and N_{ig}^A are the cumulative number of years (including the current grade) the student has spent in a small or regular/aide class, X_{ig} is a vector of student, teacher, and class characteristics, 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 robust standard errors that allow for a random individual component in the error term are reported.

Results including various sets of explanatory variables are reported in Table IX. Estimates shown in column 1 exclude student, teacher, and classmate characteristics. In column 2, regressors for measured student and teacher characteristics are included. Both of these models indicate that achievement of students in small classes jumps up by about four percentile points the first year a student attends a small class ($\beta_1 + \beta_3$), and increases by about one percentile point for each additional year the student spends in a small class thereafter. Both the initial effect of being in a small class and the cumulative effect are statistically significant in these models.

Column 3 adds four variables reflecting the composition of a 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 reassigned 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.

TABLE IX
ESTIMATES OF POOLED MODELS
DEPENDENT VARIABLE: AVERAGE PERCENTILE RANKING ON SAT TEST
COEFFICIENT ESTIMATES WITH ROBUST STANDARD ERRORS IN PARENTHESES

Variable	(1)	(2)	(3)
Initial class small (1 = yes)	2.87	3.16	2.99
	(.83)	(.80)	(.80)
Initial class regular/aide (1 = yes)	.29	.49	.58
	(.69)	(.67)	(.67)
Cumulative years in small class	1.19	1.05	.65
	(.39)	(.38)	(.39)
Cumulative years in reg/aide class	.37	.25	.14
	(.39)	(.37)	(.37)
Fraction of classmates in class previous year	—	—	.60
			(1.03)
Average fraction of classmates together previous year	—	—	−.46
			(1.52)
Fraction of classmates on free lunch	—	—	−2.73
			(1.62)
Fraction of classmates who attended kindergarten	—	—	6.85
			(1.67)
Student and teacher characteristics	No	Yes	Yes
3 current grade dummies; 3 dummies indicating first grade appeared in sample; school effects	Yes	Yes	Yes
R^2	.18	.23	.23
Sample size	25,249	24,350	24,349

Student and teacher characteristics are as follows: student race, gender, and free lunch status; and teacher race, gender, experience, and master's degree or higher status. OLS estimates are reported, with robust standard errors that adjust for a possible correlation of residuals for the same student over time in parentheses.

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 were present in the experiment during kindergarten. 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.

Including these four variables hardly changes the initial jump in test scores associated with attending a small class (see column 3), although the cumulative effect of time spent in a small

class is reduced by one-third when they are included, and is only on the margin of statistical significance ($t = 1.66$). Also notice that attendance in classes with a higher proportion of classmates who attended kindergarten has a large, positive effect on a student's own achievement. A two-standard-deviation change in the fraction of one's classmates who attended kindergarten is associated with about a three-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 this sample.

The pooled models in Table IX 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 effect of cumulative years in a small class could be generalized 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 IX are nested in this model, so they can be tested against it. An F-test rejects the parsimonious specification in Table IX at the .01 level. However, inspection of the coefficients suggests that the main reason for the rejection is that the initial effect of being in a small class is smaller in grades 2 and 3 than in kindergarten and first 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 dynamic effects of attending a small class in the early grades.

The relationship between the pooled model and the "value-added" specification, which Hanushek and Taylor [1990] suggest is superior to other specifications of the education production function, should be emphasized. The value-added model only identifies the cumulative effect of time spent in a small class; the

initial effect is differenced out. This can be seen by taking the first-difference of equation (5). If the estimates in Table IX had indicated that there was no effect of the initial year spent in a small class, the value-added specification would capture the only parameter of interest. But the pooled estimates and the results in Table VIII indicate that perhaps 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 by OLS:

(6)

$$Y_{ics,g} - Y_{ics,g-1} + \beta_0 + \beta_1 SMALL_{ics,g} + \beta_2 X_{ics,g} + \alpha_g + \alpha_s + \varepsilon_{ics,g},$$

where the dependent variable is the change in students' percentile test scores between the end of grade g and g-1, and $SMALL_{ics,g}$ is class size during grade g. 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.[20] This value-added effect is of similar magnitude to the coefficient on the cumulative years in a small class variable in the models in Table IX. Thus, although the estimated value-added specification indicates that students gain from attending small classes, the benefit is less than the full effect that accounts for the discrete gain that occurs the first year students are in a small class.

Prais [1996] and Hanushek [1998] interpret the STAR experiment as providing evidence that smaller classes did not improve performance because previously published cross-sectional results do not show the achievement test gap between students in small and regular classes expanding significantly over time. For example, Hanushek [1998] writes: "If smaller classes were valuable in each grade, the achievement gap would widen. It does not. In fact, the gap remains essentially unchanged through the sixth grade. . . . The inescapable conclusion is that the smaller classes at best matter in kindergarten." This conclusion strikes me as questionable for two reasons. First, the mix of students compared at various grade levels in the results cited by Hanushek changes over time; half of the students exit or enter the sample after kindergarten. When the same students are tracked over time, the value-added and pooled specifications show students in small

20. The other covariates in this regression are the same as in column 3 of Table IX.

classes gaining about one percentile rank per year relative to students in regular classes. Second, students appear to benefit particularly from attending a small class the first year they attend one, whether that is in kindergarten, first, second, or third grade (see Table VIII). The discrete jump in scores occurring the first year students attend a small class, combined with the entry of new waves of students over time, can distort the simple cross-sectional comparison of gains for the changing mix of students.

D. Heterogeneous Treatment Effects

The effect of being in a small class may vary for students with different backgrounds. Table X 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 fit the data better.

Smaller classes tend to have a larger initial effect, but a smaller cumulative effect, for boys as compared with girls. Students on free lunch and black students tend to have both a larger initial effect and larger cumulative effect than those not on free lunch and white students. Finally, inner-city students tend to have a more beneficial effect of attending a small class in the first year they attend one than students from other areas, and a sharper gain over time from remaining in a small class.[21] Word et al. [1990] similarly found that 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 X 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 for 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

21. 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; towns were defined as areas with more than 2500 inhabitants; and rural was defined as areas with fewer than 2500 inhabitants.

TABLE X
SEPARATE ESTIMATES FOR SELECT SAMPLES
DEPENDENT VARIABLE: AVERAGE PERCENTILE RANKING ON SAT TEST
COEFFICIENT ESTIMATES WITH ROBUST STANDARD ERRORS IN PARENTHESES

	Boys	Girls
Small	4.18	1.28
	(1.11)	(1.13)
Cumulative years in small class	.60	.92
	(.56)	(.54)
Sample size	12,576	11,773

	Free lunch	Not on free lunch
Small	3.14	2.85
	(1.10)	(1.12)
Cumulative years in small class	.94	.55
	(.59)	(.51)
Sample size	12,064	12,285

	Black	White
Small	3.84	2.58
	(1.29)	(1.02)
Cumulative years in small class	1.04	.66
	(.68)	(.48)
Sample size	8,150	16,069

	Inner city	Metropolitan	Towns	Rural
Small	3.74	2.92	3.09	2.58
	(1.68)	(1.55)	(2.83)	(1.23)
Cumulative years in small class	1.71	.57	−1.35	1.03
	(.90)	(.83)	(1.50)	(.56)
Sample size	5,154	5,906	1,872	11,417

Model and covariates are the same as column 3 of Table IX.

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 II. 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

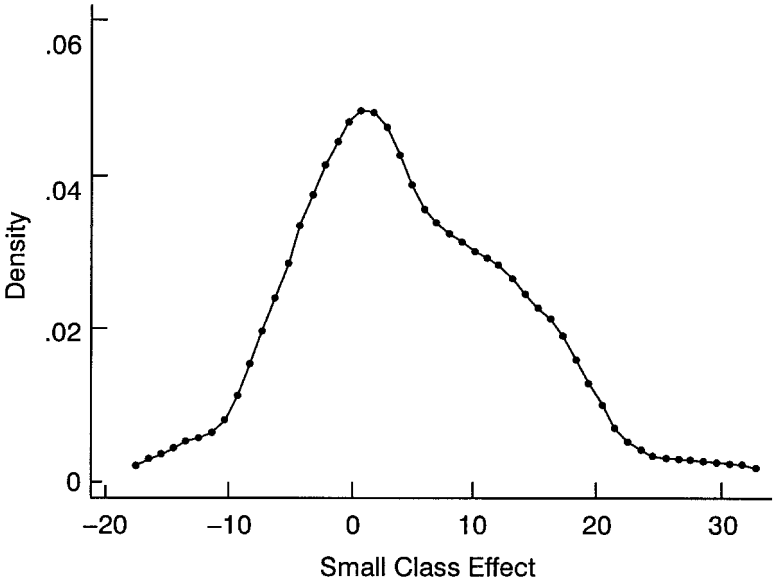

FIGURE II
Kernel Density of School-Level Small-Class Effects

exceeding +2. The mean coefficient estimate is 4.6. The standard deviation of the coefficients (after adjusting for sampling variability) 7.5 percentage points.[22] 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 resulted from "Hawthorne effects," in which teachers in small classes responded to the fact that they were part of an experiment, rather than a true causal effect of small classes themselves.[23] 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

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

23. For an interesting study that finds little evidence of Hawthorne effects in the original Hawthorne experiments, see Jones [1992]. One could argue in the current context that each individual teacher in small classes has an incentive to free ride rather than work extra hard.

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.

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 III).[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 regular-size class per grade.

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.[25] The coefficient on class size in this regression is $-.55$, with a t-ratio of -4.3. If school dummies are added to this model, the coefficient on class size falls to $-.39$, but remains statistically significant ($t = -3.1$). Based on these estimates, an eight-student reduction in class size is associated with a three-to-four-percentile increase in test scores, which is insignificantly different from estimates derived from the experimental variations in class size. These regression results do not provide much evidence of either Hawthorne or John Henry effects. And given that much of the variability in class size in the control group may be due to measurement errors (e.g., students moving in and out of class during the school year), it is noteworthy that these regressions find any evidence of class-size effects.

F. Separate Subject Test Results for SAT and BSF

Table XI presents estimates of the pooled data model corresponding to column 3 of Table IX 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 in the initial and cumulative effects of attending a small class on the

24. The standard deviation of class size in the sample of students assigned to regular classes is 2.3, as compared with 4.1 among all students in the experiment.
25. These regression results are reported in Table 12 of Krueger [1997].

TABLE XI
ESTIMATES OF POOLED DATA MODEL BY SUBJECT TEST
DEPENDENT VARIABLE: PERCENTILE SCORE ON SAT OR BSF TEST
COEFFICIENT ESTIMATES WITH ROBUST STANDARD ERRORS IN PARENTHESES

	Stanford Achievement Test			Basic Skills First		
	Math	Reading	Word	Math	Reading	Avg.
Small	2.83	3.52	2.97	1.09	3.04	2.02
	(.88)	(.88)	(.87)	(1.05)	(1.08)	(.96)
Cumulative years in small class	.45	.43	.80	1.23	.41	.83
	(.42)	(.43)	(.42)	(.44)	(.46)	(.41)
Sample size	23,794	23,461	23,630	18,174	18,010	18,250

Model and covariates are the same as in column 3 of Table IX.

math, reading and word recognition tests. Furthermore, the BSF test shows the same basic pattern as the SAT test—a 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 as 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 reanalysis 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], Folger and Breda [1989], 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 only 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, additional time spent in a small class has a positive but smaller effect 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, 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 [1999] 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 resolve these issues. It would also be helpful to compare the STAR findings with 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.

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

with 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 benefits 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.

In the meantime, we can perform the following rough benefit-cost analysis to gauge the likely order of magnitudes of the economic effects of reducing class size in the early grades. The STAR experiment reduced class size by seven or eight students, or about by one-third. Folger and Parker [1990] estimate that the cost of reducing class size in Tennessee (including capital costs) would be proportional to the total annual educational expenditures per student. In 1995–1996 total expenditures per enrolled public school student in the United States were $6459 [National Center for Educational Statistics 1996], so reducing class size by one-third would increase costs per student by about $2151 per year. We discount all benefits and costs to the present. Using a 3 percent real discount rate, the present value of the additional costs of reducing class size by one-third for the wave of entering kindergarten students for four years would be approximately $7400.

The economic benefits of the STAR experiment are much more difficult to assess than the costs. The Table V results suggest that test scores for students in small classes rose by about 0.22 standard deviations. I am not aware of any study that links achievement on the Stanford Achievement Test to later economic outcomes. Furthermore, it is possible that the cognitive gains from attending smaller classes will dissipate or grow by the time the STAR students enter the labor market. As a rough assumption, suppose that the 0.22 S.D. gain persists. How does this translate into economic benefits? Estimates based on the High School and Beyond sample in Murnane, Willet, and Levy [1995] indicate that male high school seniors who score 0.22 S.D.'s higher on the basic math achievement test in 1980 earned 1.7 percent higher earnings six years later. The comparable figure for females was 2.4 percent. Average earnings for workers age 18 and older in the United States in 1996 were $34,705 for men and $20,570 for women [U. S. Census Bureau 1996]. If we assume that real earnings will be unchanged in the future and that Murnane, Willet, and Levy's estimates can be applied to the STAR experiment, then the present value of the earnings gain from raising

test scores .22 S.D.'s is $9603 for men and $7851 for women, assuming that students enter the workforce at age 20 and retire at age 65, and using a real discount rate of 3 percent.

Many assumptions underlying this cost-benefit calculation could turn out to be wrong, including the following: real earnings may grow or shrink; the effect of test scores on future earnings may be different than assumed; class size may influence other economic outcomes, such as crime and dependency; the cost of reducing class size may be different than assumed. There is no substitute for directly measuring the economic outcomes that may be affected by reducing class size. Nonetheless, these calculations suggest that the benefit of reducing class size in terms of future earnings is in the same ballpark as the costs.

APPENDIX: SUMMARY STATISTICS
MEANS WITH STANDARD DEVIATIONS IN PARENTHESES

Variable	Grade				
	K	1	2	3	All
Class size	20.3	21.0	21.1	21.3	20.9
	(4.0)	(4.0)	(4.1)	(4.4)	(4.1)
Percentile score avg. SAT	51.4	51.5	51.2	51.0	51.3
	(26.7)	(26.9)	(26.5)	(27.0)	(26.8)
Percentile score avg. BSF	NA	51.8	51.6	51.4	51.6
		(26.1)	(26.2)	(26.1)	(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	6.58	7.67	8.70	7.12
	(0.35)	(0.49)	(0.56)	(0.59)	(1.31)
Exited sample[b]	.29	.26	.21	NA	.43
Retained	NA	NA	NA	.04	NA
Percent of teachers[c] 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/aide 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.

PRINCETON UNIVERSITY AND THE NATIONAL BUREAU OF ECONOMIC RESEARCH

REFERENCES

Angrist, Joshua, and Victor Lavy, "Using Maimonides' Rule to Estimate the Effect of Class Size on Children's Academic Achievement," *Quarterly Journal of Economics,* CXIV (1999), 533–575.

Card, David, and Alan B. Krueger, "Labor Market Effects of School Quality: Theory and Evidence," in Gary Burtless, ed., *Does Money Matter? The Effect of School Resources on Student Achievement and Adult Success* (Washington, DC: Brookings Institution, 1996).

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

Folger, John, "Editor's Introduction: Project STAR and Class Size Policy," *Peabody Journal of Education,* LXVII (1989), 1–16.

Folger, John, and Carolyn Breda, "Evidence from Project STAR about Class Size and Student Achievement," *Peabody Journal of Education,* LXVII (1989), 17–33.

Folger, John, and Jim Parker, "The Cost-Effectiveness of Adding Aides or Reducing Class Size," Vanderbilt University, mimeo, 1990.

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

——, "A More Complete Picture of School Resource Policies," *Review of Educational Research,* LXVI (1996), 397–409.

——, "The Evidence on Class Size," Occasional Paper Number 98-1, W. Allen Wallis Institute of Political Economy, University of Rochester, Rochester, NY, February 1998.

Hanushek, Eric A., and Lori Taylor, "Alternative Assessments of the Performance of Schools," *Journal of Human Resources,* XXV (1990), 179–201.

Hedges, Larry V., Richard Laine, and Rob Greenwald, "Does Money Matter? A Meta-Analysis of Studies of the Effects of Differential School Inputs on Student Outcomes," *Education Researcher,* XXIII (1994), 5–14.

Jones, Stephen, "Was There a Hawthorne Effect?" *American Journal of Sociology,* III (1992), 451–468.

Krueger, Alan, "Experimental Estimates of Education Production Functions," Princeton University, Industrial Relations Section Working Paper No. 379, Princeton, NJ, 1997.

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

Murnane, Richard, John Willet, and Frank Levy, "The Growing Importance of Cognitive Skills in Wage Determination," *Review of Economics and Statistics,* LXXVII (1995), 251–266.

National Center for Education Statistics, *Digest of Education Statistics* (Washington, DC: 1996), Table 166.

Nye, Barbara, Jayne Zaharias, B. D. Fulton, C. M. Achilles, and Richard Hooper, "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.

Prais, S. J., "Class-Size and Learning: The Tennessee Experiment—What Follows?" *Oxford Review of Education,* XXII (1996), 399–414.

Summers, Anita A., and Barbara L. Wolfe, "Do Schools Make a Difference?" *American Economic Review,* LXVII (1977), 639–652.

U. S. Census Bureau, *Historical Income Tables* (Washington, DC: 1996), Table P25.

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