THE ECONOMIC JOURNAL

JANUARY 2001

The Economic Journal, **111** (*January*), 1–28. Published by Blackwell Publishers, 108 Cowley Road, Oxford OX4 1JF, UK and 350 Main Street, Malden, MA 02148, USA.

THE EFFECT OF ATTENDING A SMALL CLASS IN THE EARLY GRADES ON COLLEGE-TEST TAKING AND MIDDLE SCHOOL TEST RESULTS: EVIDENCE FROM PROJECT STAR*

Alan B. Krueger and Diane M. Whitmore

This paper provides a long-term follow-up analysis of students who participated in the Tennessee STAR experiment. In this experiment, students and their teachers were randomly assigned to small, regular-size, or regular-size classes with a teacher aide in the first four years of school. We analyse the effect of past attendance in small classes on student test scores and whether they took the ACT or SAT college entrance exam. Attending a small class in the early grades is associated with an increased likelihood of taking a college-entrance exam, especially among minority students, and somewhat higher test scores.

Project STAR was an experiment in which an eventual 11,600 students in their first four years of school (from kindergarten until 3rd grade) were randomly assigned to a small class (target of 13–17 students), regular-size class (target of 22–25 students), or regular-size class with a teacher aide within 79 Tennessee public schools.[1] Teachers were also randomly assigned to class types. The experiment began with the wave of students who entered kindergarten in the 1985–86 school year. Students who entered a participating school while this cohort was in first, second, or third grades were added to the experiment and randomly assigned to a class type. After four years, all students were returned to regular-size classes. Students were supposed to stay in their original class-assignment type for four years, although students were randomly re-assigned between regular and regular/aide classes in first grade.[2] Students who moved along on pace graduated from high school in the Spring of 1998. Mosteller (1995) described Project STAR as 'a controlled experiment which is one of

* We thank Helen Pate-Bain and Jayne Boyd-Zaharias for providing data and answering many questions, James Maxey and Amy Schmidt for assistance merging the STAR database to the ACT and SAT databases, and Joshua Angrist, Anders Björklund, and David Card, two referees and seminar participants at Hebrew University, American Institutes for Research, Berkeley and Princeton for helpful comments. We are also grateful to Stacy Dale, Cecilia Rouse, and Mark Votruba for help with data. Naturally, we bear sole responsibility for all views expressed in the paper.

[1] The experiment is described in extensive detail in Word *et al.* (1990), Folger and Breda (1989), Finn and Achilles (1990), Krueger (1999) and Achilles (1999).

[2] In addition, about 10% of students switched between class types for other reasons. Krueger (1999) examines the impact of these transitions on the experiment, and finds that they have relatively little effect on the main results.

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.' Given the scarcity of large-scale educational experiments like Project STAR, it is important to follow up on the long-term outcomes of the subjects of the experiment.

Another reason to continue tracking the progress of the STAR participants is that some educational innovations have produced short-term gains in terms of test scores without producing lasting academic or nonacademic benefits (e.g., STEP; see Grossman and Sipe (1992)), while others have produced ephemeral gains on standardised tests but nonetheless had significant long-term benefits in terms of economic and social outcomes (e.g., Perry and many other pre-school programmes; see Barnett (1992)). The real test of educational interventions like reducing class size is whether the intervention imparts lasting economic and social benefits for society, such as increased educational attainment, enhanced earnings power and employability, reduced welfare utilisation, and reduced crime. Here we provide a first step toward evaluating the long-term impact of being assigned to a small class by examining college-entrance exam data.

This paper is organised chronologically, in terms of students' progression through school. In the next Section we present population characteristics comparing Project STAR students to students in the state and nation. Section 2 evaluates evidence on random assignment. Section 3 analyses students' scores on standardised tests taken each year from kindergarten to 8^{th} grade (i.e., grades K-8). Section 4 provides an analysis of the effect of attending a small class in the early grades on students' propensity to take the ACT or SAT college-admissions tests by the senior year of high school. Section 5 provides an analysis of the effect of class size on students' ACT and SAT scores, for the subset of students who took one of the exams. This Section presents several alternative estimators to account for sample selection bias that could arise because test scores are only available for test takers.

We regard the analysis of college test taking behaviour as the main contribution of this paper. To analyse ACT and SAT data, we worked with ACT, Inc. and the College Board and Educational Testing Service (ETS) to link information on high school seniors in the class of 1998 who took the ACT or SAT exam to records on the 11,600 students from Project STAR, regardless of where the students resided in 1998. The resulting database contains information on whether Project STAR students wrote either the ACT or SAT exam, their test scores, and information from the background questionnaire students fill out when they take the ACT or SAT exam. The ACT exam is the more prevalent college aptitude test taken by Tennessee students: some 40% of Tennessee high school seniors in our sample wrote the ACT exam while fewer than 6% wrote the SAT. This is the first database that permits a long-term examination of the behaviour and post-high school aspirations of Project STAR participants.

Our main finding is that students who were assigned to a small class are more likely to take the ACT and SAT exams. For the sample of high school seniors in 1998, 43.7% of students initially assigned to a small class took either the ACT or SAT exam, whereas 40.0% of those assigned to a regular class took one of the exams. The increase in the college-entrance-exam-taking rate due to attending a small class was substantially greater for black students than for white students. Assignment to a small class as opposed to a regular-size class appears to have raised the likelihood that black students take the ACT or SAT exam by a quarter, from 31.7 to 40.2%. As a consequence, the black-white gap in the college-test-taking rate was 54% smaller among students assigned to small classes than among students assigned to regular-size classes.

Lastly, we find insignificant differences between small- and regular-size-class students in the average SAT or ACT score among those who wrote an exam, although this comparison is clouded by selection problems since a wider pool of students assigned to small classes took one of the exams. When we adjust for selection effects, using either a parametric Heckman-selection-correction procedure or by linearly truncating the sample of test takers from small classes (based on the rank of their score) to correspond to the same proportion from regular-size classes, we find that students in small classes outperformed those in regular-size classes by about 0.1 standard deviation overall, and by about 0.2 standard deviation for black students. A nonparametric bound of the effect that attending a small class would have had for the average student who attended a regular class is between 0 and 0.5 standard deviations.

1. Sample and Population Characteristics

Schools were selected to participate in the STAR experiment if they met certain requirements (e.g., sufficient enrollment and geographic criteria), and volunteered to participate. As a consequence, the 79 participating elementary schools were not a random sample of Tennessee elementary schools. To be eligible for the experiment, a school had to be large enough to have at least three classes per grade so students could be assigned to a small, regular, or regular with teacher's aide class within each school. Furthermore, the state legislature mandated that the sample consist of a specified fraction of schools from inner-city, suburban, urban and rural areas, which led to participation of a higher proportion of inner-city schools than the overall state proportion. To assess how Project STAR schools compare to all schools in Tennessee and in the United States, we present selected characteristics of schools in Table 1.

Project STAR schools have a larger minority population than do schools in Tennessee overall, but have a proportion similar to the national average. But most minority students in the STAR experiment are black – only a small fraction of students are Hispanic, Asian, or other races – so the proportion of black students in the participating schools is nearly twice the national average. STAR schools are also located in areas with somewhat higher child poverty rates, and teachers are slightly less likely to have completed more than a bachelor's degree. Average student performance as measured by ACT scores is slightly worse for STAR students than for all Tennessee students, and Tennessee performs worse than the nation as a whole.

Table 1

Selected Population characteristics

	STAR (1)	Tennessee (2)	United States (3)
Percent minority students	33.1	23.5	31.0
Percent black students	31.7	22.6	16.1
Percent of children below poverty level	24.4	20.7	18.0
Percent of teachers with master's degree or higher	43.4	48.0	47.3
Average ACT score	19.2	19.8	21.0
Average 3rd grade enrollment across schools	89.1	69.5	67.1
Average current expenditures per student across schools	$3,423	$3,425	$4,477

Notes: With the following exceptions, data are from the 1990 Common Core of Data (CCD) from the Department of Education. For comparability, the Project STAR characteristics were calculated from the CCD. (Nevertheless, the characteristics were very similar when calculated directly from Project STAR data.) Teacher education data are for 3rd grade teachers from Project STAR data, and for 1993–4 public elementary and secondary school teachers from the Digest of Education Statistics. Race and poverty statistics for the United States are from the Census Bureau. ACT scores for Tennessee and United States are from ACT, Inc.

Since schools in the experiment were required to have at least three classes per grade, the STAR schools are larger than the average school. Average 3rd grade enrollment in Tennessee schools is about 70, whereas STAR schools had almost 90 students per grade – equal to the 72nd percentile of 3rd grade enrollment statewide. Average current expenditures per student in 1990 were virtually identical in the STAR and Tennessee sample at about $3,425. Per-pupil spending levels in Tennessee were only about three-quarters of the national average.

Most schools in the STAR experiment consisted of students in kindergarten (the typical first year of school) through sixth grade. The average kindergarten student in the experiment was 5.4 years old at the beginning of his or her first school year. Kindergarten attendance was not mandatory in Tennessee when the STAR experiment began, so some students started school in first grade. In addition, some students repeat a year of school (e.g., they are retained in the same grade level), especially in the early years, so additional students joined the wave of students going through the experiment in first, second, and third grade. New students in participating schools were randomly assigned to a class each year. After attending elementary school, students typically attend middle school and then high school. Students graduate from high school after successfully completing 12 years of school beyond the kindergarten level. Most students are 17–18 years old by the time they finish high school. In their last or penultimate year of high school, students who intend to enroll in college take the ACT or SAT exam. These are privately administered exams that are required by most colleges for admission.

2. Another Look at Random Assignment

A limitation of the design of the STAR experiment is that students were not systematically tested prior to entering a small class (see Krueger, 1999;

Hanushek, 1999). Random assignment would be expected to produce groups of students that did not differ on average among the three assignment groups, conditional on school and entry grade. If data were available, one could test for significant differences in mean student achievement scores across class types. Nonetheless, if random assignment was implemented correctly, observable characteristics of students and teachers should be similar across class types. This is examined in Panel A of Table 2, which presents a linear regression of student class-type assignment on demographic characteristics.[3] The dependent variable is a dummy variable that equals one if the student initially attended a small class, and zero if he or she initially attended a regular or regular/aide class.[4] Each student appears in the sample once, in the year he or she initially joined the experiment. Standard errors have been adjusted for heteroskedasticity that arises in the linear probability model using White standard errors. Column 1 only controls for three explanatory variables: race, sex, and free-lunch status. Column 2 additionally controls for 78 school fixed effects. Strictly speaking, class-type was randomly assigned within schools for each grade (or entry wave) that the students entered the experiment. Thus, in column 3 we control for 304 school-by-entry-wave dummy variables. When school fixed effects or school-by-entry-wave fixed effects are controlled for, none of the student characteristics predict small-class assignment for the STAR sample (see columns 2 and 3). This finding is consistent with the students being randomly assigned to class types.

An important feature of the STAR experiment is that classroom teachers were also randomly assigned to class types within each participating school. If random assignment of teachers was properly executed, one would not expect a teacher's characteristics to be related to whether or not she taught a small class. Panel B of Table 2 reports results from a linear regression of teachers' class assignments on their demographic characteristics, using the sample of 1,330 teachers pooled across all grade levels. The dependent variable equals one if the teacher was in a small class, and zero if she was in a regular or regular/aide class. The results indicate that teachers' education, experience, race and gender are essentially uncorrelated with the class type to which they were assigned. Moreover, this result holds irrespective of whether school effects or school-by-grade-level effects are held constant.

Table 2 highlights the importance of controlling for school fixed effects, since random assignment of teachers and students was performed within schools. Moreover, students were randomly assigned within schools *in the grade they initially entered Project STAR*, which suggests that it is desirable to control for

[3] Although one may object to the use of a linear probability model in this instance (e.g., as opposed to a logit), because the class-type variable is an independent variable in the models that follow, and we are simply interested in whether class-type and personal characteristics are related, the linear model provides appropriate estimates.

[4] Unfortunately, we do not know which class type students were initially assigned to, as opposed to the class type they initially attended. However, for a subsample of 18 STAR schools, Krueger (1999) finds that 99.7% of kindergarten students attended the class type they were randomly assigned to their first year in the experiment. Consequently, henceforth we treat initial assignment and the initial class the student attended interchangeably

Table 2

Examination of Random Assignment, Linear Probability Models

Dependent variable equals 1 for small classes

	A: Students				B: Teachers			
Explanatory variable	Means (SD)	(1)	(2)	(3)	Means (SD)	(4)	(5)	(6)
Intercept	—	0.255	0.311	0.278	—	0.461	0.446	0.463
		(0.020)	(0.014)	(0.014)		(0.131)	(0.151)	(0.172)
White/Asian (1 = yes)	0.631	0.025	−0.006	−0.011	0.814	0.006	−0.017	−0.032
	(0.483)	(0.010)	(0.016)	(0.016)	(0.389)	(0.035)	(0.043)	(0.053)
Female (1 = yes)	0.471	0.001	−0.003	0.000	0.988	−0.057	−0.015	−0.011
	(0.499)	(0.008)	(0.008)	(0.008)	(0.109)	(0.126)	(0.140)	(0.164)
Free lunch (1 = yes)	0.547	−0.018	−0.008	−0.016	—	—	—	—
	(0.498)	(0.009)	(0.010)	(0.010)				
Master's degree or higher (1 = yes)	—	—	—	—	0.376	−0.047	−0.059	−0.069
					(0.485)	(0.028)	(0.031)	(0.037)
Total experience	—	—	—	—	12.027	0.000	−0.000	−0.001
					(8.323)	(0.002)	(0.002)	(0.002)
Entry-grade fixed effects	—	No	Yes	No	—	No	Yes	No
School fixed effects	—	No	Yes	No	—	No	Yes	No
School-by-entry-wave fixed effects	—	No	No	Yes	—	No	No	Yes
R-squared	—	0.00	0.03	0.08	—	0.00	0.02	0.04
P-value of significance of explanatory variables	—	0.000	0.837	0.450	—	0.560	0.392	0.380

Note: White standard errors in parentheses. The free lunch variable measures whether a student was on free or reduced-price lunch during his or her entry year. For columns 1–3, the mean dependent variable is 0.26 and sample size is 11,294. For columns 4–6, the mean dependent variable is 0.39 and sample size is 1,330. For teachers, entry-grade and entry-wave are the grade level they taught. Entry-grade fixed effects are three dummy variables indicating the grade the student first entered the programme.

school-by-entry-grade effects as in column 3. Most previous analyses of the STAR data have estimated treatment effects controlling for school fixed effects, but not school-by-entry-wave fixed effects. In most of what follows, we control for dummy variables indicating the school students initially attended interacted with dummy variables indicating the grade they entered the experiment (i.e., entry wave).

3. Grades K-8

One difficulty in conducting a long-term follow-up of test score results is that the STAR students were given different tests in different grades. In grades K-3, students took the Stanford Achievement Test, and in grades 4–8 they took the Comprehensive Test of Basic Skills (CTBS). Both are multiple-choice standardised tests that measure reading and math achievement, taken by students at the end of the school year. Panel A of Table 3 presents a correlation matrix between the percentile scores on the Stanford Achievement Test (grade K-3), the CTBS (grade 4–8) and the ACT or SAT percentile rank (generally taken in grade 11 or 12). For each of the exams, the percentile ranks are based on the distribution of scores among students assigned to regular and regular/aide classes.[5] The samples used to calculate the correlations vary from year to year; Panel B reports the sample sizes. The correlations along the diagonal of Table 3 correspond to correlations of percentile ranks in adjacent years for the sample of students who have available data in those two years.

A critical juncture occurred between third and fourth grade, when all students returned to regular size classes. Unfortunately, this also coincides with the switch to the CTBS exam. A further problem is that the fourth grade sample is a subset of the overall sample because only one-third of the Memphis schools administered the CTBS that year; all Memphis administered the CTBS in later years. Nonetheless, the correlation matrix does not display a discrete jump between third and fourth grade, which suggests that the sensitivity of the CTBS and Stanford Achievement Test may be similar. We similarly find that the correlations are of roughly the same magnitude if we restrict the sample to a common set of students with available scores in grades 2–5. These results suggest that the percentile ranks can be compared across the CTBS and Stanford Achievement Test, although we recognise that data from a consistent exam would be desirable.

[5] The Stanford Achievement Test percentiles were derived by using the distribution of raw scores for students in regular and regular/aide classes, as described in Krueger (1999). We use the average percentile score of the math and reading exams. The CTBS scores were converted to percentile ranks similarly. The distribution of raw scores for students in regular and regular/aide classes were used to generate percentile ranks for those students, and for students in small classes. The average of the math and reading percentile ranks was used in the analysis. If a student repeated a grade, we used his or her first test score for that grade level. The ACT and SAT data are described in more detail below, but briefly: if a student took the ACT, we used his or her ACT score. If a student took the SAT and not the ACT, we converted the SAT score to an ACT-equivalent score. We then used the distribution of ACT scores among regular and regular/aide students to calculate percentile ranks. The standard deviation of the average percentile ranks across students for all the exams was typically 26 to 27.

Table 3

Correlations of Percentile Scores, Various Tests

Grade	K	1	2	3	4	5	6	7	8
					A. Correlations Grade				
1	0.65								
2	0.58	0.80							
3	0.51	0.71	0.80						
4	0.56	0.67	0.75	0.80					
5	0.52	0.65	0.72	0.76	0.83				
6	0.51	0.64	0.70	0.75	0.82	0.84			
7	0.53	0.67	0.73	0.75	0.81	0.84	0.86		
8	0.52	0.66	0.73	0.74	0.80	0.82	0.84	0.88	
ACT/SAT	0.42	0.57	0.65	0.68	0.71	0.74	0.75	0.79	0.81
					B. Sample sizes Grade				
Grade	K	1	2	3	4	5	6	7	8
1	4,177								
2	3,287	4,687							
3	2,904	3,988	4,724						
4	3,810	4,540	4,232	4,386					
5	4,352	5,092	4,862	5,028	6,531				
6	4,239	4,951	4,766	4,924	6,330	7,447			
7	4,178	4,854	4,642	4,762	6,216	7,308	7,174		
8	4,221	4,882	4,624	4,711	6,023	7,060	7,024	7,066	
ACT/SAT	2,351	2,720	2,666	2,723	2,905	3,335	3,314	3,227	3,319

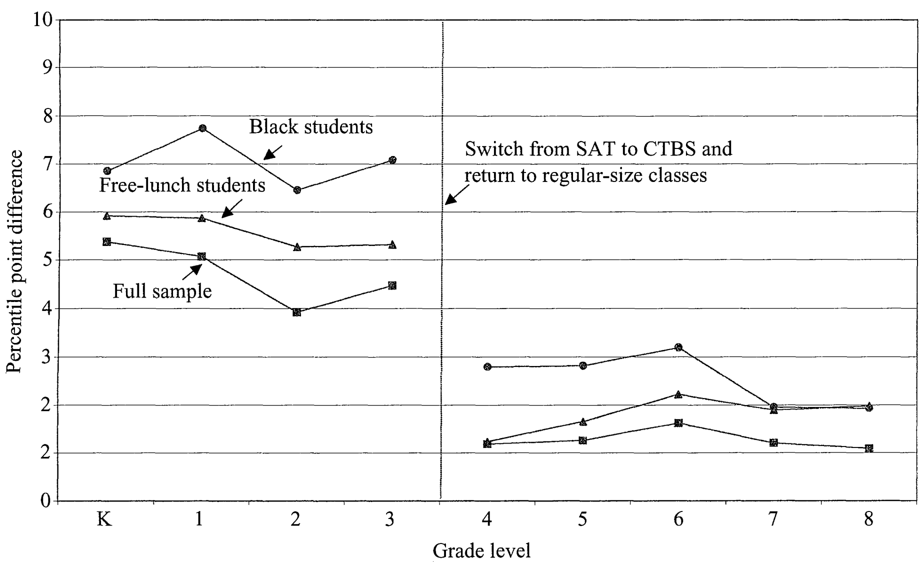
Note: Tests are Stanford Achievement Test (K-3), CTBS (4–8) and ACT or SAT normalised to ACT percentile ranks (see text). All correlations are statistically significant at the 0.05% level.

To summarise the effect of being assigned to a small class on test scores, for each grade we estimated the following regression

$$Y_{isg} = \beta_{0g} + \beta_{1g} SMALL_{is} + \beta_{2g}\mathbf{X}_{is} + \alpha_{sw} + \varepsilon_{isg}, \tag{1}$$

where Y_{isg} represents the test score percentile rank for student i in grade $g(g = K, \ldots, 8)$ who initially attended school $s(s = 1, \ldots, 79)$, $SMALL_{is}$ is a dummy variable that equals one if student i initially was assigned to a small class and zero if he or she was assigned to a regular or regular/aide class, $\mathbf{X}_{is}$ is a vector of covariates reflecting the students' sex and race, and whether the student ever received free or reduced-price lunch in grades K-3, and α_{sw} is a set of school-by-entry-wave fixed effects (based on initial school attended). The base group for the small-class-size effect consists of students who were assigned to either regular or regular/aide classes.[6] It is important to stress that class-type is based on the class the student attended the initial year of the experiment, and does not vary over time. As a consequence, the coefficient estimates are not subject to bias because of possible non-random transitions after the initial assignment.

Equation (1) was estimated separately for the full sample, for students on free or reduced-price lunch, and for the subset of black students. Fig. 1



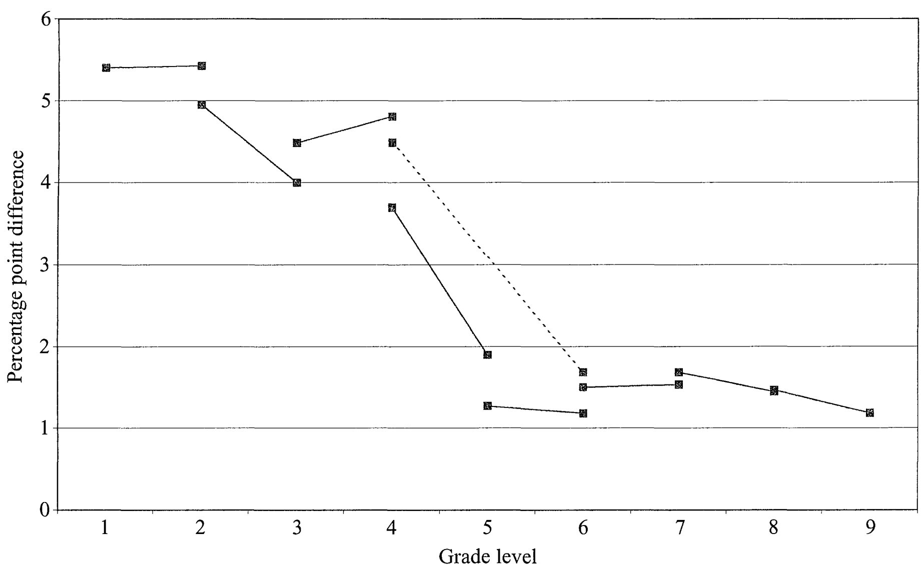
Note: Effect of class size after controlling for student's race, gender, free-lunch status and initial school-by-entry-wave fixed effects.

Fig. 1. *Small-Class Effect, K-8*

[6] For students who were present in grades K and 1, we tested this specification against a less restrictive one that differentiated the base group among those who were consistently in regular classes, those who were consistently in regular/aide classes, and those who switched between regular and regular/aide classes. This less restrictive specification typically performed no better than the one reported in the text.

summarises the coefficients on the *SMALL* dummy variable, using the largest sample of observations available for each group in each year. Because our interest is in comparing the treatment effect over time, we also calculated the small-class effects for the subset of students with available data in each adjacent pair of years. Figs. 2 and 3 summarise these results for all students and for the black students, where each segment in the figures consists of students with available data in two adjacent years. Thus, in Figs. 2 and 3, the year-over-year comparisons are always between the same set of students on each segment of the graph. The results are similar, however, if we include the largest number of students each year.

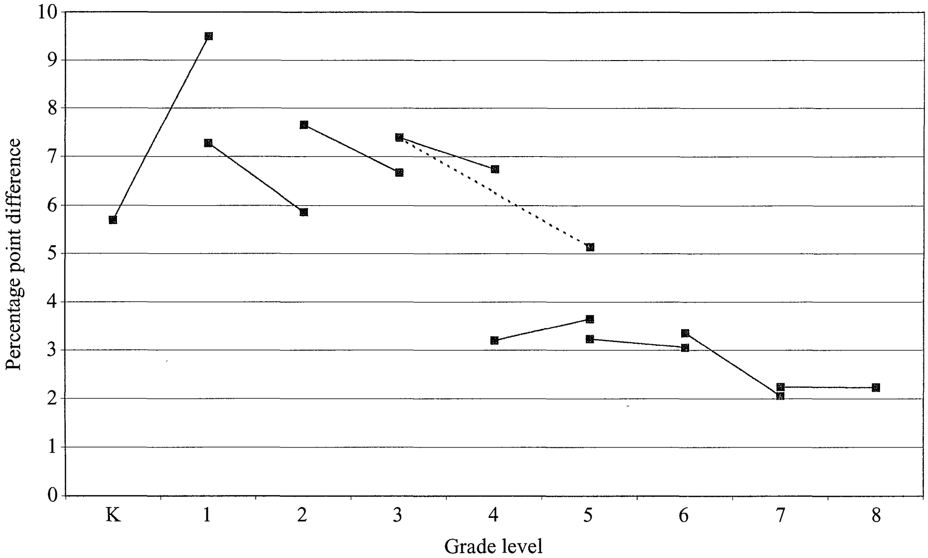
Fig. 1 summarises many of the findings of the earlier work on STAR. A 5 percentile-point gap opened up between students in small and regular-size classes by the end of kindergarten, and the gap stayed roughly constant in subsequent grades during the course of the experiment.[7] The small-class advantage was larger for the minority children and those on free lunch. Several studies have found that minority and disadvantaged students benefit more than other students from attending small classes (Summers and Wolfe, 1977; Hanushek *et al.*, 1998). We also examined how the small-class effect varies across the distribution of scores by running quantile regressions at every decile for 3rd and 8th grade test scores. These results were suggestive that the largest test score gains occur just above the middle of the distribution, since the



Note: Effect of class size after controlling for student's race, gender, free-lunch status and initial school-by-entry-wave fixed effects.

Fig. 2. *Small-Class Effect on All Students*

[7] Previous work tends to find that the small class advantage expanded between kindergarten and first grade, but that appears to result from the omission of controls for school-by-entry-wave effects.



Note: Effect of class size after controlling for student's gender, free-lunch status and initial school-by-entry-wave fixed effects.

Fig. 3. *Small-Class Effect on Black Students*

coefficient on small class peaked between deciles 5 and 7. However, we could not statistically reject coefficient equality at all deciles.

In fourth grade, when the experiment ended and students returned to regular size classes, the effect size in terms of mean percentile ranks was reduced approximately to half to one quarter of its previous magnitude. From teacher reports, we have data on the actual class size for a subset of 520 fourth grade students. Interestingly, the average fourth grade class size for students who were initially assigned to regular size classes was about 0.36 ($t = 2.4$) student smaller than it was for students initially assigned to small classes, conditional on initial school fixed effects. It is possible that, to some extent, school principals attempted to compensate for the earlier effects of the experiment, which may partially account for the relative improvement of students who were previously in larger classes. In addition, peer effects could have raised the performance of students from regular classes relative to those from small classes after the experiment ended.

Figs. 2 and 3, which use the consistent subset of students available in each pair of adjoining years, show a similar pattern. Moreover, when we use the subsample of students with scores available in both 3rd and 5th grade to avoid possible problems created by the omission of many Memphis students in the 3rd to 4th grade comparison, the results still show a sharp decline in test scores at the conclusion of the experiment when all students returned to normal-size classes. Nye *et al.* (1994) find a similar pattern with CTBS data through the seventh grade.

One important qualification should be kept in mind while considering changes in the magnitude of the small-class effect in Figs. 1–3: the tests are scaled by percentile ranks. Test score percentile ranks are not a cardinal measure. It is possible, perhaps likely, that a given percentile gap implies a larger educational difference in the higher grades than in the lower grades. Indeed, Finn *et al.* (1999) present evidence that, when the Stanford Achievement Test and CTBS scores are scaled in terms of grade equivalents, the gap between students in small and regular-size classes expands from grade K to 3, and from grade 4 to 8.

4. Effect of Class Size on College Entrance Exam Taking and ACT/SAT Scores

4.1. *Genesis of STAR-ACT-SAT Sample*

The ACT is approximately a 3-hour test, with 215 multiple choice questions covering reading, math, English and science. Similarly, the SAT is a 3.5-hour multiple choice test limited to math and verbal sections. Most students in Tennessee who aspire to attend college take the ACT exam. Nonetheless, it is important to know whether students took the SAT exam as well, because the SAT is required by many highly selective colleges, and because some students moved to states where the SAT is the predominant test. To create a longitudinal database with ACT and SAT information, in the summer of 1998 HEROS, Inc. provided the ACT and ETS organisations identical computer files which contained several variables from the STAR database, including demographic data, class assignment, and elementary school test scores. The Project STAR students' ACT and SAT data were merged to these records on the basis of the students' names, dates of birth and Social Security numbers. If a STAR record was missing information on one of these three identifiers, the remaining identifiers were used to complete the merger. The data were merged by searching over ACT and SAT records for the *entire* United States, so any student who had moved away from Tennessee should still be included in the sample. In fact, about 9% of the STAR students who were identified by the search algorithm took the ACT or SAT exam outside Tennessee. Once the data were merged, the students' names, dates of birth, and Social Security numbers were concealed to preserve confidentiality.

Several checks indicated that the data were linked properly for students who were matched. For example, the correlation between the students' ACT score percentile rank and their 8th grade CTBS percentile rank was 0.81, which is about the same as the correlation between other percentile scores of tests given four years apart (see Table 3). Additionally, the sex of the students based on their STAR records matched their sex in the ACT records in 98.7% of cases. These checks suggest that STAR students were correctly linked to their ACT and SAT records.

The ACT and SAT databases are organised by graduating high school classes. Thus, only members of the High School Class of 1998 were included in the ACT and SAT records that formed the basis of their search. As a consequence, STAR students who either repeated a grade or for some other reason were not high school seniors in 1998 could not be matched to their ACT and SAT records, even if they had taken one of the exams. Because students who were not seniors in 1998 could not be matched to their records, they were classified as not having taken the ACT or SAT exam, even though they may actually have taken it in their junior year or they may take it their senior year. This creates classification errors in our dependent variable.

Unlike the case for a continuous outcome variable, random classifications errors in a dichotomous outcome variable cause inconsistent regression coefficient estimates and inconsistent mean differences between groups (see Hausman *et al.*, 1998). The intuition for this result is that, with a dichotomous variable, errors are negatively related to the true outcome values: a one can only be misclassified as a zero, and *vice versa.* In the present case, students who fell behind a grade cannot be classified as having taken the ACT or SAT given the way the data are maintained by the ACT and ETS organisations. Randomly misclassifying some students who took the ACT or SAT exam as not having taken an exam will tend to attenuate the effect of class size on test-taking rates. Because of this feature of the data, for most of our analysis we restrict the sample to the subset of 9,397 students (81% of the full sample) who were not behind normal grade-level through eighth grade, based on information that we have on students who wrote the CTBS.[8] Measurement error in whether the student took the ACT or SAT is a much less serious problem for this subsample.

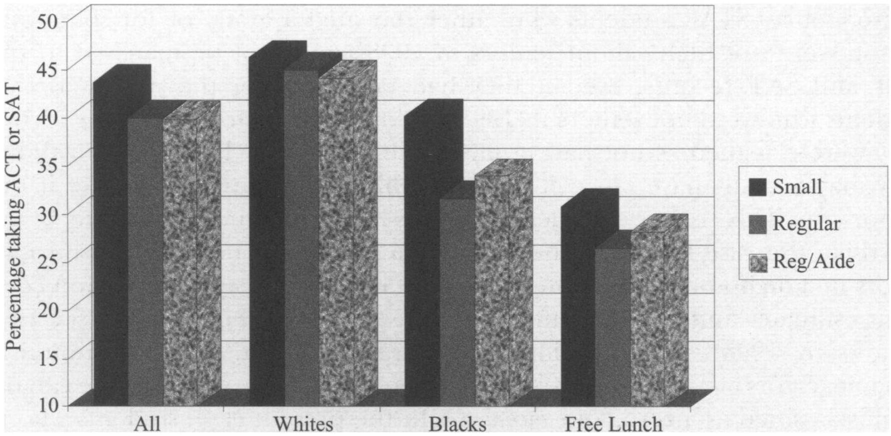
Restricting the sample to those who are on grade level, however, could introduce sample selection bias if being assigned to a small class affects the likelihood that students are behind grade level. Because we do not find a significant difference in the probability of being behind a grade by initial class assignment, this sample selection restriction is unlikely to bias our results.[9] Nonetheless, we also present logit results for the full sample for comparison. In the future, we hope to obtain additional ACT and SAT data for the Class of 1999 to augment the sample to include students who did not graduate on schedule.

4.2. *Test Taking Results*

Improving school quality can increase educational attainment by increasing the return to investment in schooling, by raising aspiration levels, and by raising skill levels (see Card and Krueger, 1996). Our main findings are illustrated in Fig. 4. This figure reports the percent of students who took *either* the ACT or the SAT exam by the type of class they attended during their initial year in Project STAR. The results are reported for all students combined, for white and black students separately, and for students who received free or reduced-price lunch

[8] That is, if the student's last available CTBS indicated that the student fell behind a grade, we excluded the student. If the CTBS information was missing, then the student was included.

[9] Pate-Bain *et al.* (1999) present preliminary evidence suggesting that students initially assigned to small classes were more likely to graduate on schedule (small: 72%; regular: 66%; regular/aide: 65%). If more (marginal) students from small classes were seniors in 1998, then restricting the sample to those who are on grade level will attenuate differences in test-taking rates between small and regular-size classes.



Notes: Fig. shows percentage of students who took either the ACT or the SAT exam, by their initial class-size assignment. Sample consists of 9,397 STAR students who were on grade level. Free lunch group includes students who ever received free or reduced-price lunch in grades K-3.

Fig. 4. *Percentage of Students Who Took the ACT or SAT College Entrance Exam by Initial Class Type*

in at least one year in grades K-3. The figure is based on the subset of students who were on grade level as of eighth grade. For all students, Fig. 4 indicates that 43.7% of students who were assigned to a small class took either the ACT or SAT exam, whereas 40.0% of those assigned to a regular-size class took one of the exams, and 39.9% of those assigned to a regular-size class with an aide took one of the exams. The 3.7 percentage-point differential between students assigned to small classes and those assigned to regular-size classes is statistically significant at the 0.05 level. The fact that regular and regular/aide students have essentially the same test-taking rates is not surprising because many of the students initially in regular classes were subsequently randomly re-assigned to a regular/aide class, and many of those initially in regular/aide classes were subsequently assigned to a regular class without an aide.

The raw data in Fig. 4 also indicate that attending a small class was particularly effective in raising the proportion of black students who wrote one of the college entrance exams. Only 31.7% of black students in regular-size classes wrote the ACT or SAT exam, whereas 40.2% of black students in small classes wrote the college entrance exam. To gain some perspective on the magnitude of this effect, note that the black-white gap in taking a college entrance exam was 13.3 percentage points for students in regular-size classes, and 6.1 percentage points for students in small classes. Thus, attending a small class reduced the black-white gap in the college-entrance-test-taking rate by 54%. Nationwide, 65.8% of white and 55.3% of black young high school graduates enrolled in college within 12 months of graduating from high school in 1996 (*Statistical Abstract of the United States*, 1998, Table 301). The 10.5 percentage point black-white gap in college enrollment for the nation as a whole is close in magnitude to the racial gap in college-entrance-exam taking rates in regular-size classes in Tennessee.

Recall that Fig. 1 showed that minority students and students on free lunch exhibited the greatest gains in test scores through middle school as a consequence of attending a small class during Project STAR. The findings in Fig. 4 complement a result that has been found consistently throughout Project STAR: minority students benefited most from attending a small class.[10] Small classes were able to narrow considerably, though not eliminate, the gap in educational performance between black and white students.

Comparing raw test-taking numbers does not take advantage of the within-school randomised design. Since the initial random assignment was done at the school level, student characteristics, on average, would not be expected to vary by class type within school. We can therefore employ a simple, balanced-within-school estimator that compares test-taking rates within schools. This estimator allows the treatment effect to vary by school, and calculates the weighted average of the school-level treatment effects. Our balanced-sample estimator takes the weighted average of the school-level difference in test-taking rates between small and regular classes for each of the 79 elementary schools, where the weights are the number of students in regular-size classes in the school.

Let D_i^S and D_i^R represent binary variables equal to one if a student in a small or regular class, respectively, took a college entrance exam, and N_j^S and N_j^R equal the number of students in small and regular classes at school j. Then the treatment effect for each school is calculated as:

$$\delta_j = \sum_i \frac{D_{ij}^S}{N_j^S} - \sum_i \frac{D_{ij}^R}{N_j^R}. \tag{2}$$

The overall small-class effect, δ, is calculated as the average of the school-level effects, weighted by enrollment in regular-size classes:

$$\delta = \frac{\sum_j \delta_j N_j^R}{\sum_j N_j^R}. \tag{3}$$

The balanced-sample estimator yields a statistically significant small-class effect of 4.4 percentage points in the overall sample, and 8.2 percentage points for black students. (The standard errors for these estimates are 1.4 and 2.3 points, respectively.) These estimates are quite similar to the effects found in the raw data, and are similar if we weight the school-level-treatment effects by the total number of observations from each school, instead of by the number of regular-class students.

Table 4 provides further evidence on the effect of class size on the percentage of students who took the college entrance exam. The first three columns

[10] This pattern is not explained by a more-intensive treatment, as small classes were not relatively smaller for black students. The reduction in class size between regular and small classes is about one third of a student larger for white students than for black students.

Table 4

Effect of Class Size on the Probability of Taking the ACT or SAT for Students on Grade Level, Logit Models

Dependent variable equals 1 if student took either SAT or ACT, and 0 otherwise

	All students				Black students			
Explanatory variable	Means (SD)	(1)	(2)	(3)	Means (SD)	(4)	(5)	(6)
Intercept	—	−0.374	−0.035	0.812	—	−0.759	0.099	−0.091
		(0.035)	(0.288)	(1.476)		(0.062)	(1.423)	(1.427)
		[−0.093]	[−0.007]	[0.151]		[−0.184]	[0.020]	[−0.018]
Small class	0.263	0.149	0.166	0.133	0.243	0.368	0.331	0.295
	(0.440)	(0.054)	(0.059)	(0.062)	(0.429)	(0.097)	(0.104)	(0.112)
		[0.036]	[0.035]	[0.027]		[0.085]	[0.068]	[0.059]
Regular/Aide class	0.364	0.003	0.041	0.042	0.378	0.107	0.150	0.119
	(0.481)	(0.050)	(0.054)	(0.057)	(0.485)	(0.087)	(0.094)	(0.100)
		[0.001]	[0.008]	[0.008]		[0.024]	[0.030]	[0.023]
White/Asian (1 = yes)	0.651	—	−0.242	−0.285	—	—	—	—
	(0.477)		(0.087)	(0.091)				
			[−0.050]	[−0.057]				
Female (1 = yes)	0.496	—	0.678	0.664	0.504	—	0.693	0.672
	(0.500)		(0.046)	(0.048)	(0.500)		(0.081)	(0.085)
			[0.241]	[0.135]			[0.210]	[0.135]
Free lunch (1 = yes)	0.568	—	−1.289	−1.229	0.843	—	−0.868	−0.876
	(0.495)		(0.055)	(0.058)	(0.364)		(0.115)	(0.123)
			[−0.291]	[−0.265]			[−0.193]	[−0.188]
School fixed effects	—	No	Yes	No	—	No	Yes	No
School-by-entry-wave fixed effects	—	No	No	Yes	—	No	No	Yes
Pseudo R-squared	—	0.00	0.11	0.14	—	0.00	0.08	0.11
Log likelihood	—	−6189.9	−5543.2	−5310.4	—	−2017.6	−1853.8	−1751.8
P-value for small class	—	0.01	0.00	0.03	—	0.00	0.00	0.01

Note: Standard errors in parentheses. Marginal effects in brackets. Sample consists of students on grade level. The mean of the dependent variable in columns (1)–(3) is 0.42 and the sample size is 9,117. The mean of the dependent variable in columns (4)–(6) is 0.35 and the sample size is 3,133. There are 78 school fixed effects in column (2), and 56 in column (5). There are 293 school-by-entry-wave fixed effects in column (3), and 140 in column (6).

of Table 4 contain logit models for all students who have not fallen behind grade level. The last three columns contain logit models for the subsample of black students who have not fallen behind grade level. The dependent variable in these models equals one if the student took the ACT or SAT, and zero if not. The logit coefficients were converted to changes in marginal probabilities, which are reported in brackets beneath the coefficients and standard errors.[11] Conditional on school-by-wave fixed effects and student race, sex, and earlier free-lunch status, we still find that students in small classes are more likely to take the ACT or SAT exam. For the combined sample, students who initially attended a small class are 2.7 percentage points more likely to take the ACT or SAT (see column 3), and black students who attended a small class are 5.9 percentage points more likely to take one of the college tests than are black students who attended regular classes (see column 6). For both samples, the gap in test taking between those in regular/aide and regular classes is statistically insignificant.

The results in columns 2 and 3 indicate that, conditional on the other regressors, black students and females are more likely to take the ACT or SAT exam than are white students and males, while students who received free lunch are substantially less likely to take the ACT or SAT exam. As mentioned earlier, black students are 6 to 13 percentage points less likely to take the ACT or SAT exam depending on class-type assignment when we do not condition on the covariates. These results are consistent with Griliches *et al.* (1978) and Lang and Ruud (1986), who find that, on average, African-Americans have lower educational attainment than whites, although African-Americans have greater average educational attainment than whites *conditional* on family background variables. The school effects and free-lunch variable probably pick up much of the family background variation controlled for in these earlier studies.

Table 5 presents corresponding results for the entire sample, regardless of whether students have fallen behind grade level. These results show a somewhat smaller effect of class size on the probability of taking a college-entrance exam, but the patterns are qualitatively similar. For these samples, attending a small class is associated with a 2 percentage-point increase in the test-taking rate for the full sample, and a 4 point increase for the sample of black students. The smaller class-size effects found in Table 5 are probably a result of greater classification errors in the test taking data in the wider sample resulting from the fact that only members of the Class of 1998 are included in the ACT and SAT databases, so all others are automatically assigned a zero for the value of the dependent variable even though they may have (or still might) taken the

[11] Since all the independent variables in the logit models are dummy variables, the marginal effects in brackets were calculated from the logit coefficients by comparing the average of the logistic distribution function evaluated at the values of the sample points, setting the independent variable (X) of interest to a value of one, and then to zero. That is, if we define the coefficient on the dummy variable of interest as δ, and let $\mathbf{Z}$ represent a vector of all the other independent variables and $\boldsymbol{\beta}$ represent a vector of their logit coefficients, the marginal effects were calculated as:

$$\Delta p/\Delta X = (1/n)\Sigma\{[\exp(\mathbf{Z}'\boldsymbol{\beta}+\delta)]/[1-\exp(\mathbf{Z}'\boldsymbol{\beta}+\delta)]\} - (1/n)\Sigma\{[\exp(\mathbf{Z}'\boldsymbol{\beta})]/[1-\exp(\mathbf{Z}'\boldsymbol{\beta})]\}.$$

Table 5

Effect of Class Size on the Probability of Taking the ACT or SAT for All Students, Logit Models

Dependent variable equals 1 if student took either SAT or ACT

	All students				Black students			
Explanatory variable	Means (SD)	(1)	(2)	(3)	Means (SD)	(4)	(5)	(6)
Intercept	—	−0.692	−0.797	0.863	—	−1.128	0.483	−0.117
		(0.033)	(0.348)	(1.089)		(0.059)	(1.437)	(1.417)
		[−0.167]	[−0.167]	[0.142]		[−0.262]	[0.076]	[−0.020]
Small class	0.262	0.122	0.117	0.093	0.240	0.340	0.293	0.238
	(0.440)	(0.050)	(0.055)	(0.058)	(0.427)	(0.090)	(0.098)	(0.104)
		[0.028]	[0.022]	[0.017]		[0.068]	[0.052]	[0.041]
Regular/Aide class	0.365	0.003	0.036	0.036	0.376	0.126	0.162	0.117
	(0.481)	(0.046)	(0.051)	(0.053)	(0.485)	(0.082)	(0.089)	(0.093)
		[0.001]	[0.007]	[0.007]		[0.024]	[0.028]	[0.020]
White/Asian (1 = yes)	0.631	—	−0.228	−0.269	—	—	—	—
	(0.483)		(0.083)	(0.086)				
			[−0.044]	[−0.050]				
Female (1 = yes)	0.471	—	0.765	0.772	0.474	—	0.788	0.783
	(0.499)		(0.044)	(0.045)	(0.499)		(0.076)	(0.079)
			[0.149]	[0.147]			[0.141]	[0.137]
Free lunch (1 = yes)	0.605	—	−1.354	−1.317	0.865	—	−0.958	−0.982
	(0.489)		(0.052)	(0.054)	(0.341)		(0.107)	(0.114)
			[−0.286]	[−0.269]			[−0.194]	[−0.194]
School fixed effects	—	No	Yes	No	—	No	Yes	No
School-by-entry-wave fixed effects	—	No	No	Yes	—	No	No	Yes
Pseudo R-squared	—	0.00	0.12	0.14	—	0.00	0.09	0.12
Log likelihood	—	−7243.5	−6404.6	−6210.1	—	−2393.6	−2172.8	−2074.3
P-value for class size	—	0.02	0.03	0.11	—	0.00	0.00	0.02

Note: Standard errors in parentheses. Marginal effects in brackets. Sample consists of all students. The mean of the dependent variable in columns (1)–(3) is 0.34 and the sample size is 11,294. The mean of the dependent variable in columns (4)–(6) is 0.27 and the sample size is 4,117. There are 78 school fixed effects in column (2) and 56 in column (5). There are 292 school-by-entry-wave fixed effects in column (3) and 143 in column (6).

ACT or SAT exam. Nonetheless, even in this sample, past attendance in a small class is associated with a higher likelihood of taking the ACT or SAT exam.

As mentioned, a majority of college-bound students in Tennessee take the ACT exam: some 40% of on-grade-level STAR students wrote the ACT exam while fewer than 6% wrote the SAT exam. Table 6 presents results where the dependent variable in columns 1 and 3 is a dummy that equals one if the student took the ACT exam, and zero if not, and the dependent variable in columns 2 and 4 is a dummy that equals one if the student took the SAT exam, and zero if not. The disaggregated results in Table 6 indicate that, compared to students assigned to regular-size classes, students assigned to small classes were more likely to take the ACT exam, *and* were more likely to take the SAT exam.

Although the STAR experiment was designed to measure the effect of being assigned to one of two narrow class-size ranges, the actual number of students in the classes varied substantially – from 11 to 20 in small classes, and from 16 to 30 in regular-size classes, over all years of the experiment. We examined the impact on test-taking of the average number of students in a child's class in

Table 6

Effect of Class Size on the Probability of Taking the ACT or SAT, Logit Models

Dependent variable equals 1 if student took the test

	All students		Black students	
Explanatory variable	ACT (1)	SAT (2)	ACT (3)	SAT (4)
Intercept	−0.510	−0.446	−0.088	−0.230
	(1.260)	(1.159)	(1.418)	(1.313)
	[−0.105]	[−0.033]	[−0.018]	[−0.013]
Small class	0.100	0.303	0.272	0.464
	(0.062)	(0.127)	(0.112)	(0.258)
	[0.021]	[0.026]	[0.055]	[0.029]
Regular/Aide class	0.038	0.133	0.088	0.367
	(0.057)	(0.122)	(0.101)	(0.253)
	[0.008]	[0.011]	[0.018]	[0.022]
White/Asian (1 = yes)	−0.290	−0.327	—	—
	(0.092)	(0.194)		
	[−0.059]	[−0.029]		
Female (1 = yes)	0.642	0.446	0.644	0.953
	(0.048)	(0.101)	(0.085)	(0.220)
	[0.135]	[0.038]	[0.131]	[0.058]
Free lunch (1 = yes)	−1.221	−1.216	−0.839	−1.416
	(0.058)	(0.138)	(0.123)	(0.242)
	[−0.266]	[−0.104]	[−0.182]	[−0.122]
School-by-entry-wave fixed effects	Yes	Yes	Yes	Yes
Pseudo R-squared	0.13	0.16	0.10	0.13
Log likelihood	−5301.1	−1528.9	−1745.7	−404.5
P-value for class size	0.11	0.02	0.02	0.07

Note: Marginal effects in brackets. Sample consists of students on grade level. Columns (1) and (2) have 9,117 observations and the means of the dependent variables are 0.40 and 0.06, respectively. Columns (3) and (4) have 3,133 observations, and the means of the dependent variables are 0.34 and 0.04, respectively. The number of school by entry-wave fixed effects in columns (1)–(4) are 291, 168, 139 and 65, respectively.

grades K-3. That is, for a student who participated in the experiment for all four years, we calculated his or her average actual class size over the four years of the experiment. If a student was missing from the experiment in one or more years (e.g., because he or she moved to a school that was not participating in the experiment), we assigned the average class size of regular classes to the student for that year, and calculated the student's average class size over four years using the available data from the experiment for the other years. We then estimated the effect of average class size during these grades on the likelihood of taking the ACT or SAT exam by Two Stage Least Squares (2SLS), using a dummy variable for initial assignment to a small class as the exogenous instrument. The 2SLS estimates indicate that reducing average class size by one student resulted in a 0.7 ($t = 2.8$) percentage point increase in the probability of taking a college entrance exam.[12] Since the mean difference in 4-year average class size between regular- and small-classes was 4.4 students, this amounts to a 3.1 percentage point increase in test taking rates in small classes – very close to the logit results reported in Table 4.

We do not know how many students who took the ACT or SAT exam have actually enrolled in college, or how many years of higher education they will ultimately complete. But based on an analysis of the 1992 wave of the High School and Beyond database, high school students from the Class of 1982 who took the ACT or SAT exam completed an average of 1.7 more years of schooling than students who did not take one of the college entrance exams, conditional on race and sex.

5. ACT and SAT Scores, With and Without Selection Adjustment

Lastly, we examined the scores students achieved on the ACT and SAT exams. For students who took the SAT but not the ACT exam, we converted their SAT score to an ACT-equivalent score using a concordance developed jointly by ACT and the College Board.[13] For any student who wrote the ACT exam we used their ACT score, even if he or she also took the SAT exam. For students who took an exam more than once, we used their first score. Naturally, any analysis of ACT and SAT scores can only be performed for the subset of students who took one of the exams. This creates a potential sample selection problem. For example, because a higher proportion of students from small classes took the ACT and SAT exams, it is likely that the group of students from small classes contains a higher fraction of relatively weak students; that is, strong students are likely to take a college entrance exam regardless of their class assignment, but marginal students who are induced to take the exam because they attended a small class are likely to be relatively lower scoring

[12] The corresponding OLS estimate was 1.8 points higher test-taking probability for a one-student reduction. The fact that the OLS estimate was larger than the 2SLS estimate suggests that Hawthorne effects were not a factor in the experiment.

[13] See http://www.collegeboard.org/sat/html/counselors/stats/stat004.html. The concordance maps re-centered SAT I scores (verbal plus math) into ACT composite scores. For the 364 students in our sample who took both tests, the correlation between their SAT and ACT scores is 0.89.

students. Such a selection process would bias downward the effect of attending a small class on average test scores. We first present results for the selected sample of students who wrote an exam, and then provide two attempts to adjust for potential sample selection bias.

To simplify the analysis, we compare students who initially attended small classes to the combined sample of those who initially attended regular or regular/aide classes, and we control for school effects instead of school-by-wave effects. Also, because we later implement a Heckman (1976) selection correction, we use raw ACT scores instead of percentile ranks. The raw ACT scores in our sample range from 9 to 36, and are approximately normally distributed, although the left tail of the distribution is thinner than the right tail. Our basic results are summarised in Table 7. For the sample of test takers, the average ACT test scores were virtually identical for students who were assigned to small and regular-size classes. The average student in a small class scored 19.3 while the average student in a regular or regular/aide class scored 19.2. This 0.108 differential is statistically insignificant, and qualitatively small – only one-fiftieth as large as the standard deviation of raw scores for the full sample. When we control for school fixed effects in column 2, students from small classes still score a statistically insignificant 0.02 standard deviation higher on the exam.

Past studies of state-level data have found that average test scores tend to decline when more students take a college entrance exam, most likely because

Table 7

Effect of Class Size on ACT or SAT Score

Dependent variable equals ACT or ACT-equivalent score

	All students			Black students		
Explanatory variable	Means (SD)	(1)	(2)	Means (SD)	(3)	(4)
Intercept	—	19.215 (0.088)	17.957 (0.235)	—	16.520 (0.132)	17.640 (0.317)
Small class	0.298 (0.449)	0.108 (0.161)	0.142 (0.144)	0.281 (0.450)	0.179 (0.234)	0.232 (0.240)
White/Asian (1 = yes)	0.709 (0.454)	—	2.603 (0.262)	—	—	—
Female (1 = yes)	0.580 (0.494)	—	−0.058 (0.139)	0.614 (0.487)	—	0.271 (0.232)
Free lunch (1 = yes)	0.389 (0.488)	—	−1.446 (0.164)	0.739 (0.439)	—	−1.760 (0.296)
School fixed effects	—	No	Yes	—	No	Yes
R-squared	—	0.00	0.21	—	0.00	0.11
Effect size	—	0.02 (0.04)	0.03 (0.03)	—	0.04 (0.05)	0.05 (0.05)

Note: White standard errors are in parentheses. Sample consists of students on grade level. If a student only took the SAT, that score is converted to its comparable ACT score (see text for details). The mean (standard deviation) of the dependent variable in columns (1) and (2) is 19.2 (4.5) and the sample size is 3,792. The mean (standard deviation) of the dependent variable in columns (3) and (4) is 16.6 (3.6) and the sample size is 1,086. The effect size is the coefficient on small class divided by the standard deviation of test scores among the full sample of students (4.5).

the marginal test takers are weaker students than the average student (see Dynarski, 1987; Card and Payne, 1998). In the STAR experiment, there were two confounding effects: selection and treatment. One might expect the treatment to result in small-class students scoring slightly higher on the ACT, as they did on previous tests up to 8th grade. But students assigned to small classes were also more likely to take the exam, suggesting that additional, weaker students in small classes were induced to write the test. Unfortunately, as a result it is difficult to interpret the score results because scores are reported conditional on taking the exam, and the treatment appears to have affected the probability of taking the exam. Table 8 presents two types of estimation results that attempt to adjust for the sample selection problem. In column 1 for the full sample, and column 3 for black students, we present results of a standard Heckman-correction procedure. Identification in these models is based on the assumption of normal errors, as there is no exclusion restriction.

For comparison, in column 2 (and column 4 for black students) we present results of a different approach for adjusting for selection. In these columns, we have artificially truncated the sample of students from small classes so that the same proportion of students from small and regular-size classes is represented

Table 8

Effect of Class Size on ACT or SAT Score with Selection Correction

Dependent variable equals ACT or ACT-equivalent score

	All students		Black students	
Explanatory variable	Heckman correction (1)	Linear truncation (2)	Heckman correction (3)	Linear truncation (4)
Intercept	14.488	18.088	8.107	17.835
	(0.632)	(0.236)	(3.886)	(0.315)
Small class	0.574	0.557	0.893	1.157
	(0.188)	(0.145)	(0.311)	(0.234)
White/Asian (1 = yes)	1.718	2.486	—	—
	(0.328)	(0.262)		
Female (1 = yes)	1.757	−0.101	2.128	0.120
	(0.174)	(0.139)	(0.283)	(0.231)
Free lunch (1 = yes)	−4.602	−1.485	−3.468	−1.897
	(0.216)	(0.164)	(0.375)	(0.292)
School fixed effects	Yes	Yes	Yes	Yes
Number of observations	9,117	3,706	3,133	1,032
Effect size	0.13	0.12	0.20	0.26
	(0.04)	(0.03)	(0.07)	(0.05)

Note: White standard errors are reported in parentheses for the linear truncation model. Sample consists of students on grade level. If a student only took the SAT, that score is converted to its comparable ACT score (see text for details). The mean (standard deviation) of the dependent variable in column (1) is 19.2 (4.5) with sample size 3,792, in column (2) it is 19.4 (4.5) with sample size 3,706, in column (3) it is 16.6 (3.6) with sample size 1,086, and in column (4) it is 16.8 (3.6) with sample size 1,032. The effect size is the coefficient on small class divided by the standard deviation of test scores among the full sample of students (4.5).

in the test-taking sample. We accomplish this by dropping from the sample the bottom $X\%$ of students based on their test results, where X is determined so that the proportion of students from small classes who took the exam equals the proportion from regular-size classes. This approach is valid if all the additional small-class students induced to take the ACT exam are from the left tail of the distribution, and if attending a small class did not change the ranking of students in small classes. Although the first assumption is clearly an extreme one, the results should provide an upper bound on the possible impact of selection bias, and provide an interesting point of comparison for the Heckman-selection results. We refer to this approach as the 'linear-truncation' procedure.[14] To compare the results to those in Table 7, in each column we calculated the 'effect size' by dividing the coefficient on the small class dummy by the standard deviation of ACT scores among all students who took the exam (equal to 4.5).

In principle, the Heckman procedure provides an estimate of the effect of attending a small class on test scores for the entire population of students (including those who do not take the test), whereas the linear-truncation approach provides an estimate of the effect of attending a small class on scores for students from regular classes who otherwise would have taken the ACT or SAT. Of course, if there is a homogeneous treatment effect, these two parameters are equal.

Interestingly, the results from both selection-adjustment procedures yield similar results. For the full sample, the Heckman-selection-correction procedure indicates that students who were assigned to a small class scored 0.13 standard deviation higher than those assigned to a regular-size class, and the linear-truncation procedure yields a 0.12 standard deviation advantage. For black students, the Heckman procedure indicates that students in small classes scored 0.20 standard deviation higher than those in regular-size classes, and the linear truncation adjustment yields an effect size of 0.26 standard deviation. In view of the extreme (and different) assumptions underlying the linear truncation and Heckman-correction procedures, it is noteworthy that the two approaches yield quantitatively similar results. The similarity of the estimates in this case follows mainly from the fact that the estimated correlation between the unobservable error terms in the selection equation and the test score equation in the Heckman procedure, denoted ρ, is close to one. The estimate of ρ is 0.96 for all students and 0.98 for black students. The (scaled) coefficients from the selection equation are also similar to those from the test score equation.[15] These results imply that the same factors that determine whether students are more likely to take the ACT or SAT test also determine how well they do on the test, which is the key assumption of the linear truncation

[14] Note that the linear truncation approach does not require normality.

[15] Notice that a Tobit model *imposes* the assumptions that the errors in the selection equation and test score equation are equal, and that the coefficients in the selection and test equations are also equal. Thus, it is not surprising that a Tobit model, in which those who do not take the ACT or SAT are treated as censored observations at the lowest score achieved, yields similar results as the Heckit and linear truncation models in this case.

model. Consequently, the linear truncation procedure has approximately the same effect as the parametric-selection correction in this case.

As a check on the procedures we used to adjust for sample selection, we performed the following experiment using the sample of 8th grade students who were on grade level. We first estimated the small-class effect size for the full sample of 6,062 students who had available 8th grade CTBS scores. Specifically, for this sample we regressed students' raw 8th grade CTBS scores on an initial small-class dummy, free lunch, sex, race, and school fixed effects. Although this is a select sample because some students did not take the exam (e.g., because they moved out of Tennessee), we think of the regression on this sample as providing an unbiased estimate of the effect of class size on achievement in the population. We then restricted this sample to the 3,262 students who took either the ACT or SAT exam, and re-estimated the same regression model. One can think of this as providing an estimate for the conditional sample, akin to the results in Table 7. Finally, using the selected sample of 3,262 observations, we estimated a Heckman-selection model and a linear truncation model (where the lowest-scoring students on the CTBS were dropped until the proportion with test scores was equal in the two class types). The results provide some limited support for the selection corrections. In particular, the effect size is 0.10 s.d. for the full sample, 0.05 s.d. for the select sample, 0.06 s.d. for the Heckman-correction estimate, and 0.15 s.d. for the linearly truncated sample.[16] The Heckman procedure and the linear truncation approach bound the estimate for the full sample.[17] Moreover, neither estimate based on the selection correction is significantly different from 0.10, the estimate for the full sample. Although we would not want to push these results too far, they do suggest that the sample selection correction estimates in Table 8 provide plausible estimates of the effect of attending a small class in the early grades on ACT test scores. Furthermore, the estimated small-class effect sizes on the college entrance exams are fairly close to the estimated effect size on the eighth grade CTBS exam, which also raises the plausibility of the findings.

As mentioned, if there are heterogeneous treatment effects, the linear truncation procedure provides an estimate of the treatment effect for regular-class students who took the ACT exam, not for the full population. We can provide lower and upper bound estimates of the small-class effect for this subsample.[18] Formally, let D_S and Y_S denote the test-taking decision and test score, respectively, if a student attends a small class, and let D_R and Y_R denote the test-taking decision and test score score if the student attends a regular–size class. D_S and D_R are binary variables that equal one if students take the

[16] The effect size is about half as large if OLS is run on the full sample and school-by-wave effects are held constant instead of school effects.

[17] Because the estimated ρ in the Heckit model is only 0.09 in this sample, results of the Heckman procedure and the linear truncation procedure are further apart in this sample than they were in the ACT sample. A Tobit model, in which non-test takers are treated as censored observations at the lowest score achieved, yields a small-class effect of 0.17 s.d.

[18] We are grateful to a referee for bringing this to our attention.

ACT or SAT exam, and zero if they do not. The target parameter is $E(Y_S - Y_R | D_R = 1)$.

Two assumptions are required to derive bounds: (1) a student in a regular class is no more likely to take a college entrance test than if he were in a small class, i.e. $D_R \leqslant D_S$, where individual student subscripts are not shown for simplicity; (2) students induced to take the test because they attended a small class, on average, achieve lower scores than their classmates who would have taken the test had they been assigned to a regular-size class. That is:

$$E(Y_S | D_S \neq D_R) \leqslant E(Y_S | D_R = 1).$$

By the first assumption, $E(Y_S D_R) \leqslant E(Y_S D_S)$, so it follows that:

$$E(Y_S | D_R = 1) \leqslant E(Y_S | D_S = 1) \frac{\Pr(D_S = 1)}{\Pr(D_R = 1)}. \tag{4}$$

And by the second assumption,

$$E(Y_S | D_R = 1) \geqslant E(Y_S | D_S = 1). \tag{5}$$

Because average scores for small- and regular-class test takers, and test taking probabilities, are observable, (4) and (5) yield the following estimable bound for $E(Y_S - Y_R | D_R = 1)$:

$$\begin{aligned} E(Y_S | D_S = 1) - E(Y_R | D_R = 1) &\leqslant E(Y_S - Y_R | D_R = 1) \\ &\leqslant E(Y_S | D_S = 1) \frac{\Pr(D_S = 1)}{\Pr(D_R = 1)} - E(Y_R | D_R = 1). \end{aligned}$$

To take advantage of the within-school randomised assignment, we implement this bounding approach using a balanced-within-school estimator, like the one in (3). Specifically, we estimated lower and upper bounds of the treatment effect for each school using the sample analog of the above formula, and then calculated the weighted average of the bounds across schools, using as weights the number of regular-class students in each school. The resulting bound of the small-class test score gain is between 0.02 and 0.47 standard deviation for the full sample. For black students, the bound is from 0.07 to 1.08 standard deviations. These bounds comfortably contain the linear truncation point estimates in Table 8, and provide an alternative way to estimate the effect that attending a small class has on test scores for those who would have taken the test even had they been assigned to a regular-size class.

6. Conclusion

The benefit from being assigned to a small class in grades K-3 on test scores for participants in the Tennessee STAR experiment appears to have declined by at least half after students were returned to regular size classes in grade 4, although a persistent, positive effect still can be measured through the eighth grade. More importantly, attendance in a small class in grades K-3 appears to have raised the likelihood that students take either the ACT or SAT college-entrance exam by the end of high school. Since most colleges in the United States require students to take either the ACT or SAT exam to be admitted, these findings suggest that lowering class size in the elementary school grades raises the prospect that students will attend college. The beneficial effect of smaller classes on college aspirations appears to be particularly strong for minority students, and students on free or reduced-price lunch. Indeed, attendance in small classes appears to cut the black-white gap in the probability of taking a college-entrance exam in half. Students who attended small classes scored about as well on the ACT or SAT, on average, as students in regular-size classes. The latter finding may be affected by the wider pool of students from small classes who took the ACT or SAT exam, however. When we implement a parametric Heckman-selection-correction procedure or linearly truncate the sample of small class students to adjust for sample selection, we find that attending a small class in the early grades raises performance on the ACT exam by about 0.13 standard deviation overall, and by 0.20 to 0.26 standard deviation for black students.

The question remains as to whether these are economically worthwhile effects. We can estimate the internal rate of return from the test-score gain from lower class size based on the STAR experiment by solving for r in the following equation:

$$\sum_{t=1}^{4} C_t/(1+r)^t = \sum_{t=14}^{61} (E_t\beta\delta)/(1+r)^t, \tag{6}$$

where C_t is the cost of reducing class size in year t, E_t is annual earnings in year t, β is a parameter that converts a one standard deviation gain in test scores at the end of high school to a proportionate increase in earnings, δ is the gain in test scores from assignment to a small class, and r is the discount rate that equates the present value of the benefits and costs. We assume students start school at age 5, begin working at age 18, and retire at age 65. The left-hand side of (6) is the present value of the cost of reducing class size, and the right hand side is the present value of the benefits. We use national school cost and earnings data to illustrate these magnitudes.

To calculate the costs, note that in the STAR experiment classes were reduced from about 22 to about 15 students, so we assume that additional funds are required for $7/15 = 47\%$ more classes. It is probably reasonable to approximate the cost of creating and staffing additional classrooms in proportion to annual per pupil expenditures. Therefore, we assume the additional cost per pupil each year a pupil is in a small class is 47% of \$7,502, which was the total expenditures per student in the United States in 1997–98 (*Digest of Education Statistics*, 1998, Table 169). Although the STAR experiment lasted 4 years, the average student who was assigned to a small class spent only 2.3 years in a small class, because half the sample entered the experiment after the first year and other students exited from the experiment. We err on the side of overstating costs by assuming that additional costs are borne fully in the first and second year, and 30% in the third year.

To calculate benefits, we assume test scores are 0.13 standard deviation higher by the end of high school as a result of assignment to a small class, as we found for the ACT scores. A key issue is: by how much do future earnings increase as a consequence of improved test scores at the end of high school? One relevant estimate is from Neal and Johnson (1996), who use the National Longitudinal Survey of Youth to estimate the effect of students' scores on the Armed Forces Qualification Test (AFQT) taken at age 15–18 (adjusted for age when the test was taken) on their earnings at age 26–29. They find that a one standard deviation increase in scores is associated with about 20% higher earnings for both men and women 11 years later. Neal and Johnson do not condition on educational attainment, so their estimate should reflect the effect of increased test scores on educational attainment as well. Another issue concerns wage growth over time. To forecast future earnings, we calculated average earnings at each age between 18 and 65 in 1998 using the 1999 March Current Population Survey. These data display the usual concave cross-sectional age-earnings profile. We use this age-earnings profile to forecast E_t, assuming that real wages will grow by 1% per annum in the future, which is in line with the Social Security Trustees' forecast for the United States.

With these assumptions, the internal rate of return from the effect size found in the STAR experiment is estimated at 5.5%.[19] Because this calculation involves many important assumptions, such as pace of future wage growth, and ignores fringe benefits as well as possible social benefits from improved education, the estimated internal rate of return is best viewed as a rough approximation rather than a precise point estimate. Nonetheless, this back-of-the-envelope calculation suggests that there is a reasonable economic rate of return from reducing class sizes at the early grades.

Despite some encouraging signs, our findings on college test taking should be viewed as preliminary because students who fell behind a grade level are not included in the ACT or SAT files. Our ACT and SAT data only pertain to students who completed high school on schedule in the Class of 1998. When data for the Class of 1999 are available, they could be added to the sample. We also hope to continue to track Project STAR students by studying their economic and social outcomes in the future, including their employment, pay, arrest rates, and welfare utilisation rates.

Princeton University

Date of receipt of first submission: October 1999
Date of receipt of final typescript: June 2000

References

Achilles, C. (1999). *Let's Put Kids First, Finally: Getting Class Size Right.* Thousand Oaks, CA: Corwin Press.

Barnett, S. (1992). 'Benefits of compensatory preschool education.' *Journal of Human Resources*, vol. 27, pp. 279–312.

[19] If we assume a 40-year working career instead, the internal rate of return is 5.3%.

Card, D. and Krueger, A. B. (1996). 'Labor market effects of school quality: theory and evidence.' In *Does Money Matter? The Effect of School Resources on Student Achievement and Adult Success* (G. Burtless, ed.), Washington, D.C.: Brookings Institution, pp. 97–140.

Card, D. and Payne, A. A. (1998). 'School finance reform, the distribution of school spending, and the distribution of SAT scores.' U.C. Berkeley, Center for Labor Economics, Working Paper.

Dynarski, M. (1987). 'The Scholastic Aptitude Test: participation and performance.' *Economics of Education Review*, vol. 3, pp. 263–73.

Finn, J. D. and Achilles, C. M. (1990). 'Answers and questions about class size: a statewide experiment.' *American Educational Research Journal*, vol. 27, pp. 557–77.

Finn, J. D., Gerber, S., Achilles, C. M. and Boyd-Zaharias, J. (1999). 'Short- and long-term effects of small classes.' SUNY Buffalo, mimeo.

Folger, J. and Breda, C. (1989). 'Evidence from Project STAR about class size and student achievement.' *Peabody Journal of Education*, vol. 67, pp. 17–33.

Griliches, Z., Hall, B. and Hausman, J. (1978). 'Missing data and self-selection in large panels.' *Annales de L'Insee*, vol. 30–1, pp. 137–76.

Grossman, J. B. and Sipe, C. L. (1992). 'Summer Training and Education Program (STEP): report on long term impacts.' Public Private Ventures, Philadelphia, PA

Hanushek, E. (1999). 'Some findings from the Tennessee STAR experiment and other investigations of class size reductions.' University of Rochester, mimeo.

Hanushek, E., Kain, J. and Rivkin, S. (1998). 'Teachers, schools, and academic achievement.' NBER Working Paper No. 6691, Cambridge, MA.

Hausman, J., Abrevaya, J. and Scott-Morton, F. M. (1998). 'Misclassification of the dependent variable in a discrete-response setting.' *Journal of Econometrics*, vol. 97, pp. 239–69.

Heckman, J. (1976). 'The common structure of statistical models of truncation, sample selection, and limited dependent variables and a simple estimator for such models.' *Annals of Economic and Social Measurement*, vol. 5, pp. 475–92.

Krueger, A. B. (1999). 'Experimental estimates of education production functions.' *Quarterly Journal of Economics*, vol. 114, pp. 497–532.

Lang, K. and Ruud, P. A. (1986). 'Returns to schooling, implicit discount rates and black-white wage differentials.' *The Review of Economics and Statistics*, vol. 68, pp. 41–7.

Mosteller, F. (1995). 'The Tennessee study of class size in the early school grades.' *The Future of Children: Critical Issues for Children and Youths*, vol. 5, pp. 113–27.

Neal, D. and Johnson, W. (1996). 'The role of premarket factors in black-white wage differentials.' *Journal of Political Economy*, vol.104, October, pp. 869–95.

Nye, B., Zaharias, J., Fulton, B. D., *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.' Tennessee State University, Center of Excellence for Research in Basic Skills.

Pate-Bain, H., Fulton, B. D., and Boyd-Zaharias, J. (1999). 'Effects of class-size reduction in the early grades (K-3) on high school performance.' HEROS, Inc., mimeo.

Summers, A. A. and Wolfe, B. L. (1977). 'Do schools make a difference?' *American Economic Review*, vol. 67, pp. 639–52.

Word, E., Johnston, J., Bain, H., *et al.* (1990). 'The state of Tennessee's Student/Teacher Achievement Ratio (STAR) Project: technical report 1985–1990.' Tennessee State Department of Education.