# *What Have Researchers Learned from Project STAR?*

DIANE WHITMORE SCHANZENBACH

Project STAR (Student/Teacher Achievement Ratio) was a large-scale randomized trial of reduced class sizes in kindergarten through the third grade. Because of the scope of the experiment, it has been used in many policy discussions. For example, the California statewide class-size-reduction policy was justified, in part, by the successes of Project STAR. Recent (failed) proposals in the Senate that sought federal assistance for class-size reductions were motivated by Project STAR research. Even the recent discussion of small schools often conflates the notion of small schools and smaller classrooms.

Because of the importance of Project STAR, it has been studied by many scholars looking at a wide variety of outcomes and even exploiting the randomization to understand variations in inputs and other aspects of the education production function that do not directly relate to class size. This paper provides an overview of the academic literature using the Project STAR experiment.

## What Was Project STAR?

Project STAR was a randomized experiment that assigned students to a small-size class (target of thirteen to seventeen students), a regular-size class (target of twenty-two to twenty-five students), or a regular-size class with a full-time teacher's aide. Teachers were also randomly assigned to class types. Randomization was done within school, so all analysis presented here looks at within-school differences by class size. The experiment took place in seventy-

The author thanks Alan Krueger, Dan Goldhaber, and Diane Ravitch for helpful comments and Rachel Henry Currans-Sheehan for comments and research assistance.

nine Tennessee public schools for a single cohort of students in kindergarten through third grade in 1985–89. Eventually, 11,600 students and 1,330 teachers took part in the experiment. The experiment was funded by the Tennessee legislature under Governor Lamar Alexander (later the secretary of education under President George H. W. Bush and currently a U.S. senator), at a total cost of approximately $12 million.[1]

In the ideal implementation of this experiment, students were to remain with the same randomly assigned class type from kindergarten through the end of the third grade. In practice, there were several major sources of deviation from this model. Students who entered a participating school while the cohort was in first, second, or third grades were added to the experiment and randomly assigned to a class type. There was a substantial number of new entrants: 45 percent of eventual participants entered after kindergarten. An especially large group of students entered in first grade—fully one-third of first-grade participants were new in first grade—in part because, at the time, kindergarten was not required in Tennessee. A relatively large fraction of students exited Project STAR schools (45 percent of overall participants), due to school moves, grade retention, or grade skipping, which also caused deviations from the original plan.[2] Students who were male, African American, or on free or reduced-price lunch were more likely both to exit and to enter Project STAR. In addition, in response to parental concerns about fairness to students, all students in regular and regular-aide classes were randomized again in the first grade.

Finally, a smaller number of students (about 10 percent of participants) were moved from one type of class to another in a nonrandom manner. Most of these moves reportedly were due to student misbehavior and not typically the result of parental requests to move their child to a small class.[3] This weakness of the experiment can be addressed through use of an "intent-to-treat" setup—that is, use of the variation caused by initial randomly assigned class instead of the actual (possibly nonrandom) class attended.

The experiment only manipulated class size and did not provide additional teacher training, new curriculum, or any other intervention. One exception is that teachers in fifteen schools were offered a three-day training seminar between years two and three of the experiment (that is, as the students were entering second grade) and again the following year. The training was given to all teachers in the Project STAR grades in those schools and occurred prior to assigning teachers to a class type. Investigation of the impacts of the training has shown that teachers who received the additional training performed no better (or worse) than teachers who were not offered the training.[4]

**Table 1. Characteristics of Students in STAR, Tennessee, and the United States**

| *Characteristic* | *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 third-grade enrollment across schools | 89.1 | 69.5 | 67.1 |
| Average current expenditures per student across schools (dollars) | 3,423 | 3,425 | 4,477 |

Source: 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 are calculated from the CCD. (Nevertheless, the characteristics are very similar when calculated directly from Project STAR data.) Teacher education data for third-grade teachers are from Project STAR data; for 1993–94 public elementary and secondary school teachers are from the *Digest of Education Statistics*. Race and poverty statistics for the United States are from the Census Bureau. ACT scores for Tennessee and the United States are from ACT, Inc.

A few aspects of the sample may limit the validity of generalizing the study to other settings. In order to be eligible to participate in the program, schools were required to have a minimum-size cohort of fifty-seven students, enough to sustain both a regular and a regular-aide classroom of twenty-two students and one small class of fifteen students. As a result, the schools that participated were about 25 percent larger, on average, than other Tennessee schools. The implications of this requirement are discussed in more detail below. Because of requirements imposed by the legislature for geographic diversity, schools in inner cities were overrepresented, and the students included were more economically disadvantaged and more likely to be African American than those in the state overall (see table 1). Even though the percentage non-white in STAR closely mirrors the percentage in the United States overall (33 versus 31 percent), this masks the fact that there were very few Hispanic and Asian students in Tennessee at the time compared to the rest of the nation. Finally, average school spending in Tennessee was about three-fourths of the nationwide average, and teachers were less likely to have a master's degree. If additional resources have greater impacts when the baseline levels are already low, this could mean that schools with higher levels of spending might experience a smaller impact from class-size reduction.

## Was Randomized Assignment Maintained?

Volumes of research have looked at the relationship between class size and student performance in nonexperimental settings, but Project STAR is the first (and only) large-scale experiment to address class size.[5] An experiment typically offers more compelling evidence than a nonexperimental study, because it allows researchers to isolate the impact of the policy they are testing. In the absence of an experiment, the effect of a policy may be confounded by other observable or unobservable factors that may be correlated with the policy. To solidify this idea, take the following model of student achievement as an example:

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

where $Y$ represents a measure of student achievement for student $i$ in school $j$, $S$ contains information on school-level inputs that affect achievement, $F$ contains family inputs, and $\varepsilon$ is an error term. Both $S$ and $F$ measure inputs over the child's entire lifetime and may contain inputs that are not observable to the econometrician. These omitted factors lead to biased coefficients if the omitted variables are correlated with included variables. For example, if students are assigned to small classes or better teachers in a compensatory manner—perhaps because of low baseline test scores or low levels of family inputs—but that information is not available to the researcher, the estimated impact of school resources will be biased. Similarly, bias will result if parents who are more involved in their child's education are more likely to push for a smaller class or better teachers and parental involvement is not measured in the data set.

The benefit of using a randomized experiment is that the treatment assignment is not related to any omitted characteristics. With a well-designed experimental assignment, a straightforward comparison of means by type of class will provide an unbiased estimate of the impact of class size on achievement. In the case of (an idealized version of) Project STAR, the equation to be estimated is as follows:

$$Y_{ics} = \beta_0 + \beta_1 SMALL_{cs} + \beta_2 AIDE_{cs} + \mathbf{X}_{ics}\gamma + \alpha_s + \upsilon_{ics}, \tag{2}$$

where *SMALL* and *AIDE* are binary variables indicating small-class and regular-aide-class assignments, respectively, and $c$ indexes class $c$ in school $s$. $\mathbf{X}$ is a vector of student-level characteristics.[6] When treatments are randomized, student-level covariates are not related to class assignment, and their inclusion should not change the estimated effect on class size, but their inclusion should contribute to the overall explanatory power of the model. A school-level fixed effect, $\alpha$, is included, so that class-type effects are identified from within-school comparisons. Finally, the error term $\upsilon$ contains class-level and individual-level components.

In practice, the nonrandom transitions and new entrants described above complicate the approach somewhat. Because of nonrandom transitions after initial assignment, it would be inappropriate to use current-year class type; instead, initial class-type assignment (the "intent-to-treat" measure) is used throughout all estimations in this paper.[7] That is, all impacts are measured with regard to the class that students were assigned to, not the class they actually attended. The intent-to-treat measure likely understates the impact of small classes by up to 15 percent.[8] As described above, new entrants into the program were randomly assigned to class types. So even though new entrants in first, second, and third grades are, on average, more disadvantaged than kindergarten entrants, randomization allows us to compare new entrants in each grade to other new entrants in the same school across class types. In practice, then, the school-level fixed effect in equation 2 is replaced with a fixed effect that combines school with a student's grade of entry (kindergarten, first grade, second grade, or third grade) to the experiment.

Impacts of reduced class size are straightforward to measure as the within-school (and entry wave) difference between class types, provided the randomization was done correctly. A compelling check of randomization is to examine a pretest to ensure that there are no measurable differences in the dependent variable between class types before the program began. Unfortunately, no baseline test was conducted. Another way to investigate whether randomization was done properly is to compare student characteristics that are related to student achievement but cannot be manipulated in response to treatment, such as student race, gender, and age. If there are no systematic differences in observable characteristics across class types, this provides support that the randomization was done properly. Table 2 presents student characteristics by entry wave and class type. The joint $p$ value for a test of equality across the columns is conditional on school fixed effects. The first four rows show that there are no systematic differences in background characteristics between class type along race, gender, free-lunch status, and age. It is also apparent that later entrants to Project STAR are more disadvantaged, with a substantially higher fraction of later entrants on free lunch and likely to be older (which may signal that they were retained in grade).[9]

Another drawback is that initial random assignment was not recorded; rather initial *enrollment* was measured. If parents successfully lobbied for a class change in the days between class assignments and the beginning of school, this would be masked in the data. To test whether this is a serious limitation, Krueger collected data on initial assignment from eighteen participating schools for 1,581 students.[10] He finds that only 0.3 percent of stu-

**Table 2. Mean Characteristics, by Entry Wave**

| *Entry wave and characteristic* | *Small (1)* | *Regular (2)* | *Regular-aide (3)* | *P value (4)* |
|---|---|---|---|---|
| *Students who entered STAR in kindergarten* | | | | |
| Free lunch | 0.47 | 0.48 | 0.50 | 0.46 |
| White or Asian | 0.68 | 0.67 | 0.66 | 0.66 |
| Age in 1985 | 5.44 | 5.43 | 5.42 | 0.38 |
| Female | 0.49 | 0.49 | 0.48 | 0.87 |
| Attrition rate | 0.49 | 0.52 | 0.53 | 0.01 |
| Days absent | 10.00 | 10.50 | 10.90 | 0.01 |
| Class size in kindergarten | 15.10 | 22.40 | 22.80 | 0.00 |
| Standardized test score in kindergarten | 0.17 | 0.00 | 0.00 | 0.00 |
| *Students who entered STAR in first grade* | | | | |
| Free lunch | 0.59 | 0.62 | 0.61 | 0.29 |
| White or Asian | 0.62 | 0.56 | 0.64 | 0.28 |
| Age in 1985 | 5.78 | 5.86 | 5.88 | 0.12 |
| Female | 0.49 | 0.44 | 0.46 | 0.33 |
| Attrition rate | 0.53 | 0.51 | 0.47 | 0.07 |
| Days absent | 8.20 | 7.70 | 7.70 | 0.95 |
| Class size in first grade | 15.90 | 22.70 | 23.50 | 0.00 |
| Standardized test score in first grade | -0.04 | -0.24 | -0.09 | 0.00 |
| *Students who entered STAR in second grade* | | | | |
| Free lunch | 0.66 | 0.63 | 0.66 | 0.58 |
| White or Asian | 0.53 | 0.54 | 0.44 | 0.15 |
| Age in 1985 | 5.94 | 6.00 | 6.03 | 0.48 |
| Female | 0.43 | 0.45 | 0.46 | 0.56 |
| Attrition rate | 0.37 | 0.34 | 0.35 | 0.58 |
| Days absent | n.a. | n.a. | n.a. | n.a. |
| Class size in second grade | 15.50 | 23.70 | 23.60 | 0.00 |
| Standardized test score in second grade | -0.11 | -0.16 | -0.27 | 0.40 |
| *Students who entered STAR in third grade* | | | | |
| Free lunch | 0.60 | 0.64 | 0.69 | 0.18 |
| White or Asian | 0.66 | 0.57 | 0.55 | 0.21 |
| Age in 1985 | 5.95 | 5.92 | 5.99 | 0.40 |
| Female | 0.43 | 0.47 | 0.46 | 0.62 |
| Attrition rate | n.a. | n.a. | n.a. | n.a. |
| Days absent | 6.00 | 7.60 | 7.60 | 0.00 |
| Class size in third grade | 16.00 | 24.10 | 24.40 | 0.00 |
| Standardized test score in third grade | -0.10 | -0.20 | -0.30 | 0.01 |

Source: Author's calculations. See table 1 for details.

n.a. Not available.

a. The $p$ value is for $F$ test of equality of all three groups, conditional on school fixed effects. Sample size ranges as follows: for students who entered in kindergarten, 6,299–6,324; for students who entered in first grade, 2,240–2,314; for students who entered in second grade, 1,585–1,679; for students who entered in third grade, 1,202–1,283. Free lunch pertains to the fraction receiving free lunch when they were randomly assigned. Test scores are scaled in units of the standard deviation of non-small classes, with mean non-small class score = 0.

dents failed to attend their initially assigned class type in kindergarten, and only one of those was moved into a small class from a regular-size one. If rates were similar at the other schools, then this would not appear to be a serious limitation.

If families feel that their child is well served by attending smaller classes (or are upset that their child has been randomly assigned to a regular class), this might yield a differential attrition rate or better attendance rates by class type.[11] Table 2 provides some evidence that this is true for the cohort entering in kindergarten. Small-class students are 3 to 4 percentile points more likely to stay at a Project STAR school through third grade and, on average, miss a fraction of a day less during their kindergarten year.[12] If students who gain the most from small classes are the ones induced to stay, then the impact of small classes may be overstated during the experimental period. Long-term follow-up data, which add back in the early exits from Project STAR, alleviate this problem. The differential attrition rate subsides in the entering waves after kindergarten. Table 2 confirms that, indeed, there was a "program"—students assigned to small classes have about seven fewer students in their class than those assigned to non-small classes. There is no difference in class size between regular and regular-aide classrooms. Finally, the table previews the results described in the next section. With the exception of the second-grade entry wave, by the end of the first year in Project STAR, the students in small classes statistically significantly outperform those in non-small classes.

Finally, it is crucial that teachers were randomly assigned. If the most effective teachers were disproportionately placed with small (or regular) classes, then the class-size effect would pick up this effect as well. Only limited data are available to confirm random assignment of teachers, which are displayed in table 3. Overall, almost all of the teachers are female, and about 80 percent are white. Average years of experience range between nine and fifteen, depending on the class type and year. In most cases, there is no within-school difference across the teachers' race, gender, experience level, or highest level of education.[13] Where there is a significant difference in teachers' characteristics across class type (for total experience in first grade and master's degree in third grade), the "best" attributes (more experience and higher degrees) are not more likely to be found in small classes, so there is no evidence that assignment was done based on seniority, measured qualifications, or anything else that would violate random assignment.

**Table 3. Teacher Characteristics, by Grade**

| *Characteristic* | *Small* *(1)* | *Regular* *(2)* | *Regular-aide* *(3)* | *P value* *(4)* |
|---|---|---|---|---|
| *Kindergarten teachers* | | | | |
| White | 0.866 | 0.780 | 0.838 | 0.327 |
| Female | 1.000 | 1.000 | 1.000 | n.a. |
| Master's degree | 0.323 | 0.364 | 0.374 | 0.653 |
| Total experience | 9.0 | 9.1 | 9.7 | 0.404 |
| Sample size | 127 | 99 | 99 | n.a. |
| *First-grade teachers* | | | | |
| White | 0.813 | 0.835 | 0.820 | 0.462 |
| Female | 0.976 | 1.000 | 1.000 | 0.139 |
| Master's degree | 0.358 | 0.313 | 0.390 | 0.674 |
| Total experience | 12.2 | 10.7 | 12.7 | 0.069 |
| Sample size | 123 | 115 | 100 | n.a. |
| *Second-grade teachers* | | | | |
| White | 0.794 | 0.770 | 0.804 | 0.583 |
| Female | 0.992 | 0.980 | 0.972 | 0.291 |
| Master's degree | 0.344 | 0.340 | 0.439 | 0.305 |
| Total experience | 13.0 | 12.7 | 13.9 | 0.542 |
| Sample size | 131 | 100 | 107 | n.a. |
| *Third-grade teachers* | | | | |
| White | 0.775 | 0.798 | 0.776 | 0.650 |
| Female | 0.964 | 0.944 | 0.972 | 0.465 |
| Master's degree | 0.377 | 0.427 | 0.523 | 0.077 |
| Total experience | 13.3 | 13.6 | 14.8 | 0.368 |
| Sample size | 138 | 89 | 107 | n.a. |

Source: Author's calculations. See table 1 for details.
n.a. = Not applicable.
a. *P* values for *F* test of equality across the three class types, conditional on school fixed effects.

## K–3 Test Score Results

Because of randomization, the impact of being assigned to a small class can be measured by comparing average test scores across class types. As described above, students were randomly assigned to small and regular classes within schools by entry wave, so all analysis controls for separate school effects for each entering cohort while estimating the treatment effect of being assigned to a small class. Results in this section report the estimated coefficient on small-class treatment from a regression estimated using the following equation:

$$Y_{igs} = \beta_{0g} + \beta_{1g} SMALL_{is} + \beta_{2g} AIDE_{is} + \beta_{3g} \mathbf{X}_{is} + \alpha_{sw} + \varepsilon_{igs}, \tag{3}$$

where $g$ indexes the grade (K–8) of the test score. Both the *SMALL* and *AIDE* variables are measured as initial assignment, not actual class attendance. The fixed effect varies by school and entry wave $w$. The coefficient on the control for classes with a teacher aide is not reported here. Because there is no impact of teacher aides on student performance, the small-class effect is similar whether or not aide classes are controlled.

The dependent variable is the mean math and reading score on the Stanford Achievement Test (SAT-9) for each grade. In cases that have missing test scores for one test but not both, the score for the nonmissing test is used. Test scores are presented as $z$ scores—that is, the scores are standardized by subtracting the mean score for non-small classes and dividing by the standard deviation. The coefficient on the indicator variable for small class can be interpreted as the standard deviation impact of the treatment.

Table 4 reports the impact of initial assignment to a small class on student test scores, and figure 1 represents the small-class impact graphically. As many researchers have found, overall students benefit about 0.15 standard deviation from assignment to a small class.[14] When the results are disaggregated by race, it appears that black students benefit about twice as much as white students (0.24 versus 0.12 standard deviation) from being assigned to a small class. Krueger and Whitmore find that this result is driven primarily by a larger treatment effect for all students in predominantly black schools, regardless of race, suggesting that benefits from additional resources are higher in such schools.[15] There is also a small, positive within-school interaction between small class and an indicator variable for black students, which means that black students gain a little more from small classes than their white classmates. Both of these findings suggest that smaller classes might be an effective strategy to narrow the black-white achievement gap.

Similar—but less stark—differences appear between free-lunch-eligible students (who must have a family income less than 185 percent of the poverty line to qualify) and non-free-lunch-eligible students. In third grade, free-lunch-eligible students gain about 0.055 standard deviation more than non-free-lunch-eligible students. Boys appear to have slightly larger small-class gains than girls, but the difference is not statistically significant.[16] There is also considerable heterogeneity in size of impact based on teacher experience. Whether or not the results are conditioned on school fixed effects, students with more experienced teachers show large, statistically significant gains. In contrast, students who have a teacher with fewer than five years of experience show smaller and often not statistically significant gains from small classes. This could help to explain the difference between the large impacts

**Table 4. Small-Class Effects on Test Scores[a]**

| *Subgroup* | *Kindergarten (1)* | *First grade (2)* | *Second grade (3)* | *Third grade (4)* |
|---|---|---|---|---|
| *Overall* | 0.187 | 0.189 | 0.141 | 0.152 |
| | (0.039) | (0.035) | (0.034) | (0.030) |
| *Race* | | | | |
| Black | 0.214 | 0.249 | 0.207 | 0.242 |
| | (0.074) | (0.063) | (0.054) | (0.060) |
| White | 0.172 | 0.161 | 0.105 | 0.115 |
| | (0.042) | (0.040) | (0.042) | (0.034) |
| *Free-lunch eligibility* | | | | |
| Free lunch | 0.188 | 0.195 | 0.174 | 0.174 |
| | (0.046) | (0.042) | (0.041) | (0.039) |
| No free lunch | 0.177 | 0.194 | 0.126 | 0.118 |
| | (0.051) | (0.047) | (0.047) | (0.041) |
| *Gender* | | | | |
| Male | 0.209 | 0.192 | 0.144 | 0.172 |
| | (0.041) | (0.040) | (0.039) | (0.400) |
| Female | 0.157 | 0.180 | 0.132 | 0.122 |
| | (0.049) | (0.040) | (0.042) | (0.040) |
| *Teacher experience* | | | | |
| Less experience | 0.310 | 0.057 | 0.073 | 0.171 |
| (less than or equal to four years) | (0.121) | (0.081) | (0.064) | (0.100) |
| More experience (more than five years) | 0.181 | 0.269 | 0.179 | 0.154 |
| | (0.044) | (0.056) | (0.037) | (0.034) |

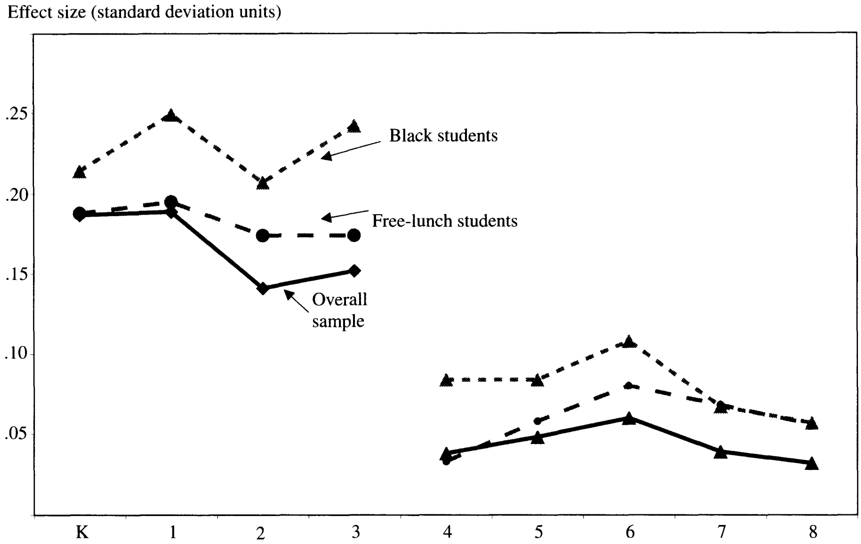
Source: Author's calculations. See table 1 for details.

a. Standard errors clustered on classroom are in parentheses. Coefficients represent the impact of being randomly assigned to a small class on test scores, in terms of standard deviation units. Each cell represents a separate regression. Other covariates included in the model (where appropriate) are indicators for white, female, and free-lunch eligibility and school-by-entry-wave fixed effects.

found in Project STAR and the disappointing results from California's statewide class-size-reduction effort.[17] In California, many districts were forced to hire new, inexperienced teachers in order to reduce class size, and there is evidence in Project STAR that these inexperienced teachers are not particularly more effective when given small classes.

One concern about the results is that they may be driven by Hawthorne effects, meaning that students and teachers behave differently simply because they are given special treatment. If this were the case, then benefits from small classes would dissipate if the class-size reduction were made permanent and the students and teachers were no longer being studied and made to feel special. Krueger addresses this by investigating differences in achievement using the variation in regular-size classes.[18] There is little reason to think that

**Figure 1. Small-Class Impacts on Test Scores**




Source: Author's calculations.

Hawthorne effects would cause some classes to behave differently than other classes in the treatment group. Regular-size classes range from sixteen to thirty students, but the bulk of the distribution is between twenty and twenty-six students. Whether or not school effects are controlled, students in a regular class with slightly fewer members outscore students in larger regular classes. The estimated magnitude of a one-student reduction in class size is consistent with the magnitude of the experimental results (which estimate the impact of a seven-student reduction).

Another concern might arise regarding generalizability of the findings to a larger population, because of the size restrictions in place for participation in the experiment. Schools were required to be large enough to support three classrooms per grade, and on average Project STAR schools were about 30 percent larger than schools across Tennessee or the United States. If larger schools are somehow more or less effective if given additional resources, then the findings in Project STAR may not be generalizable to smaller school settings.[19] One way to address this using the available data is to compare the small-class advantage across the wide range of school sizes in Project STAR. In third grade, the largest school had more than 200 students, while the smallest participating school had only fifty-six students. To test whether school size is related to the magnitude of the class-size effect, I separate schools into enrollment quartiles. The lowest quartile had an average of about sixty students, while the highest had more than twice that number, at 130 students. There is no statistically significant relationship between school size and test performance or between school size and small-class impact. This suggests that limiting Project STAR to relatively large school sizes is not a major drawback in terms of generalizability to other samples.

## Follow-up Studies on Test Performance

In fourth grade, the class-size-reduction experiment concluded, and all students returned to regular-size classes. At the same time, the assessment test was changed from the SAT-9 to the Comprehensive Test of Basic Skills (CTBS). Both tests are multiple-choice standardized tests that measure reading and math achievement and are taken by students at the end of the school year. The CTBS results are scaled in the same manner as the SAT-9, in terms of standard deviation units. One important difference in the data is that all students in public schools statewide who participated in Project STAR were included in the follow-up study, even if they were retained a grade.[20] As a result, some students took the fourth-grade test in 1990, while others took it in later years or even took it more than once. In the analysis reported here, all scores for a given grade—no matter what year a student was in that grade—are compared. In the event that a student was retained in a particular grade and took that grade's test more than once, the first available score is used.[21] As before, all estimates are conditional on school-by-entry-wave fixed effects.

Results for grades four to eight are reported in table 5 and illustrated in figure 1. Overall, there is a persistent positive impact of small-class assignment that is statistically significant (or borderline significant) through eighth grade, as found in previous studies.[22] The magnitude of the gain is one-third to half the size observed while the students were in the experimental classes. When the results are disaggregated, though, the impact appears to remain stronger with black and free-lunch-eligible students than with more advantaged students.

Another potential measure of achievement is whether students take either the SAT or the ACT college entrance exam, which can be used as an early proxy for college attendance.[23] In order to measure this, the Educational Testing Service and ACT matched Project STAR student data to their national databases of test records, as described by Krueger and Whitmore.[24] To examine whether assignment to a small class influences the rate of taking a college

**Table 5. Small-Class Effects on Long-Term Test Scores[a]**

| *Subgroup* | *Fourth grade[b] (1)* | *Fifth grade[b] (2)* | *Sixth grade[b] (3)* | *Seventh grade[b] (4)* | *Eighth grade[b] (5)* | *Took college entrance test[c] (6)* |
|---|---|---|---|---|---|---|
| *Overall* | 0.035 | 0.048 | 0.060 | 0.040 | 0.036 | 0.024 |
| | (0.025) | (0.024) | (0.025) | (0.025) | (0.025) | (0.010) |
| *Race* | | | | | | |
| Black | 0.078 | 0.080 | 0.105 | 0.066 | 0.063 | 0.050 |
| | (0.048) | (0.043) | (0.045) | (0.042) | (0.046) | (0.018) |
| White | 0.026 | 0.028 | 0.043 | 0.031 | 0.027 | 0.011 |
| | (0.027) | (0.029) | (0.029) | (0.031) | (0.030) | (0.013) |
| *Free-lunch eligibility* | | | | | | |
| Free lunch | 0.029 | 0.058 | 0.080 | 0.067 | 0.064 | 0.031 |
| | (0.036) | (0.031) | (0.034) | (0.031) | (0.034) | (0.014) |
| No free lunch | 0.048 | 0.036 | 0.027 | 0.003 | -0.012 | 0.018 |
| | (0.035) | (0.034) | (0.036) | (0.038) | (0.038) | (0.017) |
| *Gender* | | | | | | |
| Male | 0.046 | 0.063 | 0.065 | 0.086 | 0.061 | 0.029 |
| | (0.032) | (0.034) | (0.036) | (0.032) | (0.035) | (0.014) |
| Female | 0.011 | 0.026 | 0.039 | -0.017 | 0.002 | 0.015 |
| | (0.037) | (0.033) | (0.031) | (0.035) | (0.034) | (0.016) |

Source: Author's calculations. See table 1 for details.

a. Standard errors clustered on initial school are in parentheses. Coefficients represent the impact on test scores of being randomly assigned to a small class, in terms of standard deviation units. Each cell represents a separate regression. Other covariates included in the model (where appropriate) are indicators for white, female, and free-lunch eligibility and school-by-entry-wave fixed effects.

b. *z* score.

c. 1 = yes.

entrance exam, a binary variable indicating that a college entrance exam is taken is the dependent variable in equation 3. The impact of small-class assignment on college test taking is included as the final column in table 5. Overall, test-taking rates are about 2 percentage points higher for small-class students. For white students, the impact is small and not statistically significant. However, black students are 5 points more likely to take the SAT or ACT if they were assigned to a small rather than regular-size class.[25] This corresponds to a rate of 38 percent of black students assigned to small classes who take at least one of the college entrance exams, compared with 33 percent in regular classes. The chance of such a large difference in test-taking rates between the small- and regular-class students occurring by chance is less than 1 in 10,000. Krueger and Whitmore interpret the magnitude of these effects as a reduction in the black-white test-taking gap.[26] For students in regular classes, the black-white gap in taking a college entrance exam is 12.9 percentage points, compared to 5.1 percentage points for students in small classes. Thus assigning all students to a small class is estimated to reduce the black-white gap in the test-taking rate by an impressive 60 percent. After controlling for increased selection into the test among small-class students, the impact on test scores for blacks is 0.15 standard deviation, about the same as the test-score impact in third grade.

## Follow-up Studies on Non-Test Outcomes

Increased investments in school quality may also affect the frequency of negative social outcomes such as crime, welfare receipt, and teen pregnancy. For example, the Perry Preschool Project was an intervention that increased preschool quality and yielded large, persistent effects on outcomes of participants through age forty, despite disappointing impacts on standardized test scores when participants were younger.[27] To date, only limited outcomes have been studied in the Project STAR data, but as the sample ages, more research should be done on outcomes such as earnings, criminal behavior, and welfare utilization. To the extent that reductions are measured along these lines, the social cost-benefit analysis of reducing class size in the early grades would increase.

One measure of outcome is criminal arrest data, which come from Tennessee Department of Corrections records.[28] The data match was only available for Shelby County, Tennessee, which contains the city of Memphis. All criminal arrests through April 2001 were included. Data on arrests for criminal offenses were collected and analyzed for 3,300 students who originally attended one of the 20 Memphis elementary schools included in Project STAR. Criminal arrests in this sample are rare (but much more likely for males than for females): only 6.3 percent of male Project STAR students are reported as having been arrested compared with 4.3 percent for females. Because the outcome is so rare, results are estimated using a probit model, where the dependent variable equals 1 if a student has been arrested or, in some cases, arrested for a particular type of crime. Using this approach, it is found that males assigned to small classes are 2.2 percentage points less likely to be arrested for any crime than those in regular-size classes ($p$ value = 0.12). Reductions were particularly strong for violent and property crimes, in which small-class attendance decreased criminal arrest rates by 55 ($p$ value = 0.09) and 57 percent ($p$ value = 0.02), respectively. No statistically meaningful decreases were apparent for drug or traffic crime arrests. Results are similar when both males and females are included. Another outcome that has been measured is the teen birth rate. Birth records, like crime records, are matched in Tennessee only. Birth records are only available by calendar year. The analysis is restricted to births during 1997 and 1998 because most students graduated from high school in 1998. Unlike all other outcomes discussed so far, birth records are not available for individuals and are aggregated up to a school-by-class-type level. Small-class assignment is associated with a statistically significant 1.6 percentage point (or 33 percent impact; $t$ ratio = 2.29) lower teen birth rate for white females but has no measured impact on black females.

In addition to investigating straightforward impacts of the program, several researchers have exploited aspects of the randomization to answer other important questions about the education production function that are not directly related to class size. Dee uses the random assignment of teachers and students to investigate the impact on achievement of having a teacher who is the same race as the student.[29] He finds that having an own-race teacher increases students' performance by a statistically significant amount (3–5 percentile rank points) for both black and white students. Dee and Keys find that students perform better in math when they are randomly assigned to a teacher who is receiving merit pay.[30] Several other recent papers use randomization into class types to identify the impact of peer composition on student achievement. Graham finds that being randomly assigned to a classroom with average peer test scores in the seventy-fifth percentile leads to a 1.1 standard deviation increase in student performance, while peers with average scores in the twenty-fifth percentile decrease their own performance by 0.9 standard deviation.[31] Schanzenbach finds that if a child's peers are made up of a higher-than-average fraction of girls, there is about a 2 percentile point positive impact on the student's own test score.[32] The effect seems to work through two channels: girls, on average, have higher test scores, but even after factoring that part out, there is a positive effect of having more girls in the classroom. This suggests that girls may change the culture of a classroom—at least in the early grades—to facilitate more learning for both boys and girls. Cascio and Schanzenbach find that conditional on a student's own age, being older than average relative to one's classmates is associated with higher test scores.[33]

## Why Might Small Classes Matter?

As described previously, the Project STAR experiment only manipulated class size. There were no changes in curriculum, there was no additional teacher training at most schools, and few, if any, new teachers needed to be hired given the limited scope of the project. What, then, caused the impact on student performance?

Lazear puts forth a useful theory of educational production.[34] In it, gains from class-size reduction are driven by a decrease in the amount of time that the classroom is being disrupted. A simple summary of the model is as follows: a child is behaving in class at a given moment with probability $p$ and misbehaving with probability $(1 - p)$. In the model, misbehaving might mean disrupting class by talking or fighting or be as benign as asking questions that slow down the class or monopolizing the teacher's time. When there are $n$ children in the classroom, $p^n$ is the probability that the entire class is behaving and learning is taking place (assuming that $p$ is independent across children). Assuming a constant disruption rate, having fewer students in the class means that learning is taking place a larger fraction of the time.

In the model, the impact of reducing class size depends not only on the size of the class but also on the behavior of the students in it. As a result, Lazear's theory predicts that class-size effects should be larger for classes with more poorly behaved students. There is some evidence that teachers report higher rates of misbehavior in predominantly black schools compared to predominantly white schools, so the theory fits well with the observation that class size has a larger impact in black schools.

Another potential mechanism is that early interventions improve noncognitive skills in addition to the cognitive skills measured by standardized test scores.[35] During kindergarten (at least during the period studied here), a primary focus in the classroom is to build noncognitive skills such as listening, sitting still, and cooperating. To the extent that kindergarten teachers could be more effective in teaching these skills in small classes, an improvement in noncognitive skills might spill over to later outcomes even if cognitive gains are modest. I test two approaches to isolating the impact of small classes on noncognitive skills. An indirect test is to compare the observed increase in a non-test outcome to the increase that would be predicted by the improvement in test scores. For example, there is a large increase in college test taking for black students assigned to a small class, which is a signal that more small-class students are going on to college. Is that increase larger than what would be predicted by the test-score increase observed in third grade (or kindergarten)? To test this, the observed relationship between test-taking rates is compared with standardized test scores in the early grades. Within regular-size classes only, the probability that a black student takes a college entrance exam increases by 0.6 percentage point for each additional percentile rank attained on the third-grade test ($R^2 = 0.17$). As reported in Krueger and Whitmore, the percentile rank increase for black students in third grade is 7.6 points for small classes.[36] The point estimate implies that 4.5 points of the observed 8-point increase in college test taking for small-class students can be attributed to improved test scores in third grade alone.[37] Taking account of sampling variability, it is possible that most of the increase in college test taking is explained by higher cognitive test scores in third grade. On the other hand, part of the increase in college test taking could be explained by improvements in (unmeasured) noncognitive skills that might have been associated with small-class attendance in the early grades.

A more direct way to isolate the impact of small-class size is to investigate a few measures of noncognitive skills collected in the data (such as listening, self-concept, and motivation). Measures of self-concept and motivation are from the Self-Concept and Motivational Inventory (SCAMIN), which was given at the end of the school year in all grades. The measure has been found to be only moderately reliable and is thought to be most reliable for middle-income, suburban students.[38]

Results are presented in table 6. Columns 1–4 present self-concept and motivation scores in standard deviation units. As with other results, this table shows the coefficient on an indicator for initial small-class assignment in a regression framework that controls for school-by-entry-wave fixed effects. By the end of kindergarten, there are some apparent increases in self-concept overall and for females, blacks, and students with free-lunch status and an increase in motivation for females. This impact dissipates entirely by the end of third grade, which could be due to lack of validity in the test for older students or could reflect a catching up of students in regular classes.

Columns 5 and 6 present scores for the listening subsection of the Basic Skills First test, a curriculum-based test given to all children in Project STAR. Like the self-concept and motivation measures, listening is measured imperfectly and may reflect a combination of cognitive and noncognitive skills. Here there are strong positive impacts of small classes on listening skills for most subgroups of students in kindergarten. By third grade, the impacts are smaller and only marginally significant (in contrast to the robust findings over time for the cognitive tests of math and reading), but this may suggest that one important mechanism for improved outcomes later on is that small classes increase a student's noncognitive skills. Overall, these approaches seem to indicate that the improvement in noncognitive skills that is due to small classes may play a role in the positive outcomes, but that the results are also consistent with the outcomes' being driven by cognitive skills alone.

**Table 6. Small-Class Impacts on Potential Measures of Noncognitive Skills[a]**

| | *Self-concept* | | *Motivation* | | *Listening* | |
|---|---|---|---|---|---|---|
| *Subgroup* | *Kindergarten (1)* | *Third grade (2)* | *Kindergarten (3)* | *Third grade (4)* | *Kindergarten (5)* | *Third grade (6)* |
| *Overall* | 0.135 | 0.024 | 0.031 | -0.084 | 0.172 | 0.104 |
| | (0.055) | (0.035) | (0.030) | (0.046) | (0.075) | (0.055) |
| *Race* | | | | | | |
| Black | 0.251 | -0.032 | 0.080 | -0.141 | 0.274 | 0.141 |
| | (0.093) | (0.071) | (0.063) | (0.085) | (0.132) | (0.106) |
| White | 0.083 | 0.052 | 0.008 | -0.058 | 0.112 | 0.089 |
| | (0.061) | (0.036) | (0.031) | (0.055) | (0.086) | (0.063) |
| *Free-lunch eligibility* | | | | | | |
| Free lunch | 0.218 | 0.003 | 0.051 | -0.111 | 0.253 | 0.125 |
| | (0.076) | (0.053) | (0.048) | (0.073) | (0.098) | (0.078) |
| No free lunch | 0.063 | 0.034 | 0.006 | -0.002 | 0.090 | 0.064 |
| | (0.062) | (0.045) | (0.034) | (0.068) | (0.087) | (0.062) |
| *Gender* | | | | | | |
| Male | 0.088 | 0.061 | -0.024 | 0.005 | 0.146 | 0.145 |
| | (0.063) | (0.042) | (0.035) | (0.055) | (0.085) | (0.072) |
| Female | 0.178 | 0.000 | 0.086 | -0.153 | 0.204 | 0.067 |
| | (0.064) | (0.050) | (0.038) | (0.066) | (0.083) | (0.068) |

Source: Author's calculations. See table 1 for details.

a. Standard errors clustered on classroom are in parentheses. All tests are scaled in standard deviation units. Self-concept and motivation scores are from the SCAMIN test, and listening scores are from the Basic Skills First assessment test.

## How Do Small-Class Impacts Compare to Other Interventions?

As a policy intervention, is reducing class size an economically worthwhile investment? One way to measure this is to compare the long-term benefits of higher test scores to the costs of reducing class size. To do this, I update Krueger and Whitmore's cost-benefit analysis and solve for $r$, the internal rate of return, in the following equation:

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

where $C_t$ is the cost of reducing class size in year $t$ and $E_t$ is annual earnings in year $t$.[39] Additionally, I calculate net present value for various assumptions about the rate of return. For this calculation, it is assumed that students begin kindergarten at age five, begin working at age eighteen, and retire at age sixty-five. The $\beta$ term relates a 1 standard deviation gain in test scores during high school to an increase in future earnings. It would be preferable to have a parameter that relates increases in elementary school test scores to later earnings, but I have found no such estimate in the literature. The $\delta$ term is the increase in test scores for students assigned to small classes.

The left-hand side of the equation represents the present discounted value of the costs of reduced class size. $C_t$ is the additional cost of reducing class size by seven students, which requires a 47 percent increase in the number of classes. Assuming that the additional cost is proportional to current average spending per pupil, the additional cost each year is 47 percent of $10,551, which is the average national spending based on 2001–02 figures, inflated by the consumer price index to 2005 prices. Since students, on average, receive 2.3 years of class-size reduction, those costs are reflected as full years of class-size reduction in kindergarten and first grade and 0.3 times the additional cost in second grade. In this calculation, no costs are borne in third grade.

The right-hand side of the equation measures benefits of reduced class size over a student's working life. The test score increase ($\delta$) is 0.152 standard deviation, from column 4 of table 4, and the estimate of $\beta$ is a 0.20 increase in wages from a 1 standard deviation increase in test scores.[40] To forecast future earnings, average earnings for each age between eighteen and sixty-five are calculated for 2005 from the 2006 March *Current Population Survey*, and future real earnings growth is assumed to be 0, 1, or 2 percent a year.[41] Results are presented in table 7.

Using these assumptions, the estimated internal rate of return from the class-size effect in Project STAR ranges from 5 to 10 percent. The net present value of the investment ranges from $3,000 to $50,000.[42] Of course, exact numbers should be viewed with caution, as the calculation is based on many assumptions that may or may not prove to be reasonable. The benefits only include future increased earnings and ignore potential impacts on crime and other behavior since those results are on the margin of statistical significance. If future follow-ups of Project STAR find lasting impacts on other outcomes such as crime or welfare use, then those benefits should be added to the equation and the internal rate of return would increase.

Another way to think about whether the investment in smaller classes is worthwhile is to compare it to other proposed interventions. Some reforms such as improved teacher training have been shown to be ineffective and therefore do not make an appropriate comparison.[43] The most promising interventions aside from smaller class size come from the new literature on school choice through either vouchers or charter schools. When the Project STAR results are compared to the largest positive effects found in voucher

**Table 7. Cost-Benefit Analysis of Small Classes[a]**

| | Net present value, assuming annual real wage growth of | | |
|---|---|---|---|
| Discount rate assumption | 0 percent | 1 percent | 2 percent |
| Discount rate (r) | | | |
| 0.02 | 16,617 | 29,267 | 48,335 |
| 0.03 | 8,421 | 16,892 | 29,492 |
| 0.04 | 2,924 | 13,654 | 17,162 |
| 0.05 | -821 | 3,178 | 8,965 |
| Internal rate of return (IRR), percent | 4.75 | 5.82 | 9.95 |

Source: Author's calculations based on data in Census Bureau and Bureau of Labor Statistics, Current Population Survey.
a. All figures in 2005 dollars. Assumes 2.3 years of exposure to small classes and a test score increase of 0.152 standard deviation. A 1 standard deviation increase in average test scores is associated with a 0.2 standard deviation increase in earnings.

experiments—the New York City results found by Howell and others—small classes yield 30 percent higher test scores but also cost 50 percent more.[44] Compared to recent work by Hoxby and Rockoff on three charter schools in Chicago that employed random assignment, small classes yield about a 35 percent improvement in test scores.[45] It is difficult to compare the costs associated with charter schools, which are allotted approximately the same in per-pupil revenues from the school district but generally raise substantial external funding. In order to determine which would be a better investment—school choice or reduced class size—we need to have some measure of willingness to pay for improved test scores and compare the smaller improvements associated (so far) with choice to the larger cost associated with class-size reduction.

## Conclusion

Mosteller describes Project STAR as "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."[46] Given the scarcity of large-scale educational experiments like Project STAR, it is important to learn as much as possible from the experiment. Researchers have combed through the experiment to learn not only about the effects of reduced class size on test scores but also to gain insight into classroom dynamics. Overall, Project STAR indicates that reducing class size is a reasonable economic investment: the benefits are sizable and long lasting, especially for black students, and the overall benefits outweigh the costs.

## Notes

1. Elizabeth J. Word and others, *Student/Teacher Achievement Ratio (STAR): Tennessee's K–3 Class Size Study* (Nashville: Tennessee State Department of Education, 1990); Frederick Mosteller, "The Tennessee Study of Class Size in the Early School Grades," *Future of Children* 5, no. 2 (Summer-Fall 1995): 113–27.

2. Eric A. Hanushek, "Some Findings from an Independent Investigation of the Tennessee STAR Experiment and from Other Investigations of Class Size Effects," *Educational Evaluation and Policy Analysis* 21, no. 2 (1999): 154–64.

3. Of all transitions, 25 percent were into small (more desirable) classes. See Alan B. Krueger, "Experimental Estimates of Education Production Functions," *Quarterly Journal of Economics* 114, no. 2 (1999): 497–532.

4. Elizabeth Word and others, *The State of Tennessee's Student/Teacher Achievement Ratio (STAR) Project: Final Summary Report 1985–1990* (Nashville: Tennessee State Department of Education, 1990).

5. Eric A. Hanushek, "The Economics of Schooling: Production and Efficiency in Public Schools," *Journal of Economic Literature* 24, no. 3 (1986): 1141–77; Eric A. Hanushek, "Assessing the Effects of School Resources on Student Performance: An Update," *Educational Evaluation and Policy Analysis* 19, no. 2 (Summer 1997): 141–64.

6. Krueger, "Experimental Estimates of Education Production Functions."

7. Some early work uses current class size instead of initial assignment. See Jeremy D. Finn and Charles M. Achilles, "Answers and Questions about Class Size: A Statewide Experiment," *American Educational Research Journal* 27 (Fall 1990): 557–77; Word and others, *Student/Teacher Achievement Ratio*.

8. This conservative "intent-to-treat" measure based on random assignment is typically considered preferable to models that measure the impact of actual class type attended in cases in which there is nonrandom movement between classes. A simple example may help to illustrate this: if a child were moved from a regular class to a small class because his parents insisted on the move, it is reasonable to assume that the parents are especially active in other aspects of the student's education, say, by monitoring homework especially closely or providing other education-enhancing opportunities. The problem arises because we do not have perfect measures of the home environment. In the ideal case in which class type is randomly assigned, these home environment measures are not correlated with class type, and their impacts are absorbed in the error term in equation 2. When the effect of actual (nonrandom) class attended is measured instead, some of the impacts of the active home environment also may be picked up because attendance may be correlated with this "home environment" component of the error term. Using the experimentally induced variation—even though not all students attended their assigned class type, and some students' test scores "count" toward the regular class they were assigned to, even though they actually attended small classes—circumvents this problem but understates the true impact. Krueger provides a more detailed discussion of this matter; see Krueger, "Experimental Estimates of Education Production Functions."

9. In addition, tests of the interaction between race, gender, and free-lunch status and small-class assignment and a model saturated with all the interaction terms between race, gender, and free-lunch status show no statistically significant relationship between baseline characteristics and small-class assignment ($p$ values of 0.94 and 0.15, respectively.)

10. Krueger, "Experimental Estimates of Education Production Functions."

11. Better attendance rates in small classes might also be caused by fewer classmates from whom to pick up germs and illnesses.

12. Another potential source of bias is selective withdrawal prior to entering kindergarten. Krueger reports that, of initially assigned students, 10.4 percent assigned to small kindergarten

classes failed to enroll in kindergarten in the fall. Comparable figures for regular classes are 14.3 percent (difference with regard to small classes has $t = 1.86$), and for regular-with-aide classes are 12.2 percent (difference between small classes $t = 0.86$). See Krueger, "Experimental Estimates of Education Production Functions."

13. As Hanushek points out, though, these characteristics—and most other observable characteristics—are poor predictors of teacher effectiveness. Recent work by Jacob and Lefgren indicates that principals generally can identify which teachers are at the extremes of effectiveness across the entire school but are not able to distinguish between the middle 60–80 percent of teachers in terms of effectiveness. It may be reasonable to conclude that teachers are not allocated to classes based on their potential impact if principals are not able to distinguish easily between most teachers. See Hanushek, "Some Findings from an Independent Investigation of the Tennessee STAR Experiment and from Other Investigations of Class Size Effects"; Brian A. Jacob and Lars Lefgren, "Principals as Agents: Subjective Performance Measures in Education," Working Paper 11463 (Cambridge, Mass.: National Bureau of Economic Research, 2005).

14. Finn and Achilles, "Answers and Questions about Class Size"; Word and others, *Student/Teacher Achievement Ratio*; Elizabeth R. Word and others, *The State of Tennessee's Student/Teacher Achievement Ratio (STAR) Project: Technical Report 1985–1990* (Nashville: Tennessee State Department of Education, 1994); Krueger, "Experimental Estimates of Education Production Functions"; Alan B. Krueger and Diane M. Whitmore, "The Effect of Attending a Small Class in the Early Grades on College-Test Taking and Middle School Test Results: Evidence from Project STAR," *Economic Journal* 111, no. 468 (2001): 1–28; Barbara Nye, Larry V. Hedges, and Spyros Konstantopoulos, "Do Low-Achieving Students Benefit More from Small Classes? Evidence from the Tennessee Class Size Experiment," *Educational Evaluation and Policy Analysis* 24, no. 3 (2002): 201–17.

15. Alan B. Krueger and Diane M. Whitmore, "Would Smaller Classes Help Close the Black-White Achievement Gap?" in *Bridging the Achievement Gap*, edited by John E. Chubb and Tom Loveless (Brookings, 2002).

16. Diane Whitmore, "Resource and Peer Impacts on Girls' Academic Achievement: Evidence from a Randomized Experiment," *American Economic Review* 95, no. 22 (2005): 199–203.

17. G. W. Bohrnstedt and B. M. Stecher, *What We Have Learned about Class Size Reduction in California* (Sacramento: California Department of Education, 2002).

18. Krueger, "Experimental Estimates of Education Production Functions."

19. Thanks to Dan Goldhaber for pointing this out.

20. An estimated 20 percent of students had been retained a grade by eighth grade, but this probability did not vary with initial class assignment.

21. Krueger and Whitmore use a different approach and deduct a small number of points from the test scores of students who were retained. Since class type is not correlated with the probability of grade retention, the results are robust to either approach. See Krueger and Whitmore, "The Effect of Attending a Small Class in the Early Grades on College-Test Taking and Middle School Test Results."

22. Charles M. Achilles and others, *The Lasting Benefits Study (LBS) in Grades 4 and 5 (1990–1991): A Legacy from Tennessee's Four-Year (K–3) Class-Size Study (1985–1989), Project STAR* (Nashville: HEROS, 1993); Barbara Nye and others, *The Lasting Benefits Study: Eighth Grade Technical Report* (Tennessee State University, Center of Excellence for Research in Basic Skills, 1995); Krueger and Whitmore, "The Effect of Attending a Small Class in the Early Grades on College-Test Taking and Middle School Test Results."

23. Susan Dynarski and I have been collecting college performance data, so a more direct measure of college behavior will be available soon.

24. Krueger and Whitmore, "The Effect of Attending a Small Class in the Early Grades on College-Test Taking and Middle School Test Results"; Krueger and Whitmore, "Would Smaller Classes Help Close the Black-White Achievement Gap?"

25. The numbers reported here are slightly different from those reported in Krueger and Whitmore, "The Effect of Attending a Small Class in the Early Grades on College-Test Taking and Middle School Test Results," because at that time data were only available for students who graduated on track with their kindergarten class. Updated data allow a study of the impact of class size on students who were retained.

26. Krueger and Whitmore, "Would Smaller Classes Help Close the Black-White Achievement Gap?"

27. Lawrence J. Schweinhart and others, *Lifetime Effects: The High/Scope Perry Preschool Study through Age 40* (Ypsilanti, Mich.: High/Scope Press, 2005).

28. Krueger and Whitmore, "Would Smaller Classes Help Close the Black-White Achievement Gap?" Diane W. Schanzenbach, "The Impact of Early School Intervention on Crime" (University of Chicago, 2007).

29. Thomas S. Dee, "Teachers, Race, and Student Achievement in a Randomized Experiment," *Review of Economics and Statistics* 86, no. 1 (2004): 195–210.

30. Thomas S. Dee and Benjamin J. Keys, "Does Merit Pay Reward Good Teachers? Evidence from a Randomized Experiment," *Journal of Policy Analysis and Management* 23, no. 3 (2005): 471–88.

31. Bryan S. Graham, "Identifying Social Interactions through Excess Variance Contrasts," University of California, Berkeley, 2005.

32. Diane Whitmore Schanzenbach, "Classroom Gender Composition and Student Achievement: Evidence from a Randomized Experiment," Working Paper (Harris School at the University of Chicago, 2006).

33. Elizabeth U. Cascio and Diane Whitmore Schanzenbach, "First in the Class? Academic Redshirting and Education Production," Dartmouth College, 2006.

34. Edward P. Lazear, "Educational Production," *Quarterly Journal of Economics* 116, no. 3 (2001): 777–803.

35. James J. Heckman, "Skill Formation and the Economics of Investing in Disadvantaged Children," *Science* 312, no. 5782 (2006): 1900–02; James J. Heckman and Alan B. Krueger, *Inequality in America: What Role for Human Capital Policies?* (MIT Press, 2005).

36. Krueger and Whitmore, "The Effect of Attending a Small Class in the Early Grades on College-Test Taking and Middle School Test Results."

37. Another approach is to predict college test taking based on third-grade test scores and a small-class indicator. Using this approach, the small-class indicator does not predict college test taking in a statistically significant manner (coefficient = 0.028, standard error = 0.027), which is consistent with the entire test-taking effect's being driven by higher test scores. The approach similarly yields not significant findings predicting test-taking behavior on eighth-grade scores and a small-class indicator (coefficient = 0.030, standard error 0.020).

38. Finn and Achilles, "Answers and Questions about Class Size"; Kimberly A. Gordon Rouse and Susan E. Cashin, "Children's Self-Concept and Motivation Assessment: Initial Reliability and Validity," paper presented at the annual conference of the American Psychological Association, New Orleans, 2001.

39. Krueger and Whitmore, "The Effect of Attending a Small Class in the Early Grades on College-Test Taking and Middle School Test Results."

40. From Derek Neal and William R. Johnson, "The Role of Pre-Market Factors in Black-White Wage Differences," *Journal of Political Economy* 104, no. 5 (October 1996): 869–95.

41. Census Bureau and Bureau of Labor Statistics, *Current Population Survey,* March 2006 (www.census.gov/cps/ [October 2006]).

42. Another approach to estimating a cost-benefit analysis would be to multiply the increased earnings associated with postsecondary education by the increased likelihood that a student takes the SAT or ACT. This assumes that the 0.024 increased test-taking rate translates to the same size increase in postsecondary attendance. Using a similar approach to that described above and measuring wages for high school only versus at least some postsecondary schooling, the internal rate of return ranges from 4.5 percent (in the no-growth scenario) to 6.7 percent (if real growth is assumed to be 2 percent). Further information is available from the author upon request.

43. Brian A. Jacob and Lars Lefgren, "The Impact of Teacher Training on Student Achievement: Quasi-Experimental Evidence from School Reform Efforts in Chicago," *Journal of Human Resources* 39, no. 1 (2004): 50–79.

44. William Howell and others, "School Vouchers and Academic Performance: Results from Three Randomized Field Trials," *Journal of Policy Analysis and Management* 21, no. 2 (2002): 191–218; Krueger and Whitmore, "Would Smaller Classes Help Close the Black-White Achievement Gap?"

45. Caroline M. Hoxby and Jonah E. Rockoff, "The Impact of Charter Schools on Student Achievement," Columbia Business School, 2004.

46. Mosteller, "The Tennessee Study of Class Size in the Early School Grades."