The Long-Term Effects of Small Classes: A Five-Year Follow-Up of the Tennessee Class Size Experiment

Barbara Nye, Larry V. Hedges and Spyros Konstantopoulos

EDUCATIONAL EVALUATION AND POLICY ANALYSIS 1999; 21; 127
DOI: 10.3102/01623737021002127

Educational Evaluation and Policy Analysis
Summer 1999, Vol. 21, No. 2, pp. 127–142

The Long-Term Effects of Small Classes: A Five-Year Follow-Up of the Tennessee Class Size Experiment

Barbara Nye
Tennessee State University

Larry V. Hedges
Spyros Konstantopoulos
University of Chicago

Reduction of class size to increase academic achievement is a policy option that is currently of great interest. Although the results of small-scale randomized experiments and some interpretations of large-scale econometric studies point to positive effects of small classes, the evidence has been seen by some scholars as ambiguous. Project STAR in Tennessee, a 4-year, large-scale randomized experiment on the effects of class size, provided persuasive evidence that small classes had immediate effects on academic achievement. However, it was not clear whether these effects would persist over time as the children returned to classes of regular size or would fade, as have the effects of most other early education interventions. This article reports analyses of a 5-year follow-up of the students in that experiment. The analyses described here suggest that class size effects persist for at least 5 years and remain large enough to be important for educational policy. Thus, small classes in early grades appear to have lasting benefits.

Reduction of class size is a policy option that is gaining increasing attention throughout the nation. Some 18 states have recently adopted policies that reduce class sizes with the goal of improving achievement. Reduction of class size was also included in the president's recent education initiative.

Until recently, there was no consensus about whether class size reduction led to increases in academic achievement and other desired outcomes of schooling. Even today, there is not a perfect consensus among scholars about interpretation of the evidence on the effect of class size on academic achievement. However, we argue that important evidence was provided by Project STAR, the Tennessee class size experiment. This experiment provided rather strong evidence that class size reduction led to immediate increases in academic achievement in both reading and mathematics. However, it was not clear, on the basis of the STAR findings, whether these increases in academic achievement would disappear after the students were no longer in small classes or would persist for some or many years thereafter. The present study investigated whether assignment to small classes in the primary grades has lasting effects on academic achievement and, if so, how those effects are distributed across schools with different social compositions and teacher characteristics.

Research on the Effects of Class Size

Well over 100 experiments and quasi-experimental studies of the effects of class size have been conducted, each involving assignments of students to smaller or larger classes. This literature has been reviewed by Glass and Smith (1979); Glass, Cahen, Smith, and Filby (1982); Hedges and Stock (1983); and Mosteller, Light, and Sachs (1996). Although there is some disagreement about interpretation of the experimental research (e.g., Educational Research Service, 1980; Slavin, 1984), these syntheses of research suggest positive effects of class size

reduction, with the effects becoming larger as classes become smaller.

However, most of the class size experiments are small scale and short term. Consequently, it is difficult to determine whether the special circumstances surrounding the experiment caused the effects and whether they would have occurred in a more natural setting. That is, small-scale experiments may have high internal validity, but it is difficult to assess whether they will generalize to other settings.

Econometric work on education production functions (e.g., Hanushek, 1986) provides another means of studying the effects of class size. This research relies on the fact that naturally occurring class sizes vary across schools and models the relationship between class size and an outcome (usually academic achievement) while controlling for student characteristics such as social class or prior achievement. The considerable number of econometric studies of the effects of class size on achievement have been reviewed elsewhere (see, e.g., Hanushek, 1989; Hedges, Laine, & Greenwald, 1994). There is a certain degree of controversy over the interpretation of the econometric studies. Some researchers, such as Hanushek (1989), are persuaded that effects of class size must be small, since so few of the studies found statistically significant effects. Others, such as Greenwald, Hedges, and Laine (1996), argue that a better gauge of the size of the effects is the magnitude of the actual regression coefficients obtained in the studies regardless of their individual statistical significance.

Whatever the proper method of summary may be, the econometric studies are likely to be more generalizable than small-scale experiments, since they arise from data on operating schools. However, the internal validity of econometric studies is more questionable. That is, it is difficult to determine whether the relations between class size and achievement (controlling for student background) are causal. In many cases, the student background data are rather limited and may fail to fully account for individual differences between students assigned to classes of different sizes. For example, it may be that students are assigned to smaller classes because their achievement is low (e.g., in compensatory or remedial programs). Consequently, the relation between small classes and achievement may reflect the fact that low achievement causes certain students to be assigned to small classes in naturally operating schools.

An important addition to the research on class size was the Tennessee class size experiment, Project STAR. This study (labeled "one of the great experiments in education in U.S. history" by Mosteller et al., 1996, p. 814) substantially mitigated many of the problems of other class size research. It was a large-scale, randomized experiment that randomized both teachers and students into classrooms within each participating school. Consequently, it had high internal validity. The STAR project involved a broad range of schools from throughout a rather diverse state. It included both large urban school districts and small rural ones and both wealthy and poor districts. Roughly one quarter of the Project STAR students were from schools in urban areas, one quarter were from suburban schools, and one half were from schools in rural areas. Approximately one third of the sample was Black, and nearly half of the students were eligible for free or reduced-price lunches. Therefore, although Project STAR did not involve a national probability sample, it included essentially the entire range of educational conditions that occur in American education, and it is more likely to be generalizable than smaller, more circumscribed studies conducted in only one location. Moreover, it was conducted for 4 years as part of the everyday operation of the schools that participated and, thus, probably avoided the effects associated with new or experimental programs.

The analyses published during the original project (Word et al., 1990), some conducted since that time (e.g., Finn & Achilles, 1990), and reports based on these analyses (e.g., Mosteller, 1995) have important weaknesses. However, both these analyses and more recent (and satisfactory) examinations of data from Project STAR (e.g., Krueger, 1998) support the conclusion that small classes in kindergarten through Grade 3 lead to higher academic achievement. The effects of small classes were found to be remarkably consistent across schools, suggesting that small classes benefit students of all types in all kinds of schools (Nye, Hedges, & Konstantopoulos, 1999).

Lasting Benefits of Early Education Interventions

Although a number of early education interventions are able to demonstrate immediate benefits, the effects typically disappear over time so that the achievement of students who receive the intervention is no higher than that of students who do

not. For example, McKey et al. (1985) found that initial positive effects of Head Start had completely faded 3 years after the intervention. Haskins (1989) found similar results in a comprehensive review of model preschool and Head Start programs. An analysis of 300 early intervention programs by White (1985) revealed a pattern of initial effects that faded over time, as did an investigation by Barnett (1992). The Perry Preschool Project appears to be a notable exception to these findings, producing lasting gains in achievement (see, e.g., Barnett, 1985). However, research on many other early interventions suggests that achievement effects disappear by Grade 8 (Barnett, 1992), although a somewhat more optimistic view of possible lasting effects of early interventions was offered by Barnett (1995).

The Present Study

This article addresses the question of the long-term effects of small classes by examining the achievement of students who were involved in Project STAR for the 5 years after the experiment ended, when these students were in Grades 4 to 8. Examination of long-term effects of small classes using the Project STAR data is complicated by at least three factors that have not been adequately addressed in earlier analyses of these data.

The first—and undoubtedly most important—factor is attrition. Even though the students and teachers were initially assigned to classes at random, differences in attrition between small and regular-sized classes could bias the results of long-term follow-up comparisons of students initially in small and regular-sized classes. Attrition, of course, is a problem that vexes not only Project STAR but nearly every longitudinal field study.[1]

The second factor is that although there was random assignment of students to classes, there was a small amount of switching among treatment groups, which could bias results. Although analyses of this switching suggest that it did not substantially bias the initial estimates of treatment effects (see Nye et al., 1999), switching could conceivably have a larger effect among the (somewhat smaller) sample available for long-term follow-up.

The third factor is that, because new students were enrolled (and randomized) each year, all students did not spend the same amount of time in the experiment. That is, some of the students were part of the experiment for only 1 year (e.g., Grade 3), while others were part of the experiment for 4 years (kindergarten to Grade 3). If the effects are at all cumulative, the effects on students with 4 years of small classes would be expected to be larger than those on students with only 1 year of small classes.

To address these complicating factors, we carried out several analyses to help understand the effects of each of them. We first conducted an analysis in which the treatment was defined as class size type (small or regular) at Grade 3. This analysis included students with at least 1 year of the small (or regular) class size and might be considered a typical type of analysis for a longitudinal experiment.

To examine the effects of switching among class types, we conducted an analysis in which the treatment was defined as the class size type to which students were initially assigned, even though they may have received the other class size type (e.g., students who were assigned to a regular-sized class but actually enrolled in a small class). This analysis, labeled in medicine an "intention to treat" analysis, should underestimate the effects of small classes unless they actually produce lower achievement than regular classes (which seems implausible).

To examine the effects of a greater or lesser number of years of treatment, we conducted analyses in which the treatment was defined as being in small classes for at least 1, 2, 3, or all 4 years (kindergarten through Grade 3) of the experiment. By comparing the results of these analyses with those of the first analysis, it is possible to obtain some insight about whether treatment effects are larger for students who receive more years of small classes.

Because the pattern of attrition was slightly different in each of the analyses just outlined, we examined the pattern of differential attrition between treatment groups for each analysis.

The Intervention: Project STAR

Project STAR (Student/Teacher Achievement Ratio) was a randomized experiment commissioned in 1985 by the Tennessee state legislature and implemented by a consortium of Tennessee universities and the Tennessee State Department of Education. The total cost of the experiment, including the cost of hiring new teachers and classroom aides, was approximately $12 million.

Initially, all Tennessee school districts were asked to participate in Project STAR, and about 180 schools in about 50 of the 141 school systems in

the state expressed interest in participating. Only about 100 schools had enough students in each grade to meet the size criterion (at least 57 students per grade necessary to form one small and two regular-sized classes) for participation. This size criterion, which was necessary to permit assignment to class types within schools, excluded very small schools from the study. Ultimately, 79 elementary schools in 42 school districts became sites in the STAR experiment. Districts had to agree to participate for 4 years and to allow site visitations for verification of class sizes, interviewing, and data collection, including extra student testing. They also had to allow random assignment of pupils and teachers to class types from kindergarten through Grade 3.

The state paid for the additional teachers and classroom aides, and only class size conditions changed within schools. School districts and buildings followed their own policies, curricula, and so forth. It is important that the study design provided that no student would receive any less service than would normally be provided by the state as a consequence of being in the STAR project. Thus, there was no incentive for any student not to participate, and participating schools, as a whole, had an incentive (in the form of overall greater resources) to take part in the STAR project.

The experiment randomly assigned kindergarten students to small classes (13–17 students), larger classes (22–26 students), or larger classes with a full-time classroom aide. Teachers were also randomly assigned to classes of different types. These class-type assignments of students and teachers were maintained through the third grade. Some students entered the study in the first grade and subsequent grades and were randomly assigned to classes at that time.

The Project STAR database and a long-term follow-up of the students who were involved in the project (the Lasting Benefits Study, Nye et al., 1994) are parts of a larger program of research on class size conducted by the Center of Excellence for Research and Policy on Basic Skills at Tennessee State University. This article is based on the data collected as part of the Lasting Benefits Study.

Statistical Methods

The design of the STAR experiment involved randomly assigning students and teachers to treatments within schools. The study was conceptually a series of within-school experiments conducted with the same procedures and outcome variables. Since the variance in student achievement within schools is typically different from the variance between schools, the sampling design involved clustering, which should be taken into account in determining the type of analysis conducted. One such analysis involves the use of hierarchical linear models (Bryk & Raudenbush, 1992). These models permit analysis and pooling of school-specific regressions (including, for example, treatment effects) in a manner that takes into account the clustering of the sample by school.

The within-school model used in our primary analysis treated student achievement as a function of student characteristics (gender and socioeconomic status [SES; operationalized by free-lunch eligibility]), treatment group assignment, and the interaction of assignment and gender. The specific model for the achievement test score Y_{ij} of the ith student in the jth school was

$$Y_{ij} = \beta_{0j} + \beta_{1j}\text{FEMALE}_{ij} + \beta_{2j}\text{SES}_{ij} + \beta_{3j}\text{SMALL}_{ij} + \beta_{4j}\text{FS}_{ij} + \varepsilon_{ij},$$

where FEMALE_{ij} is a dummy variable for gender, SES_{ij} is a dummy variable for free-lunch eligibility, SMALL_{ij} is an indicator variable for small class size, FS_{ij} is the interaction of FEMALE and SMALL, and ε_{ij} is a student-specific sampling error.

We modeled variation across schools in each of the school-specific regression coefficients as depending on the geographic location of the school and the percentage of Black students in the school. The specific two-level model for the mth coefficient in the jth school β_{mj} was therefore

$$\beta_{mj} = \gamma_{0m} + \gamma_{1m}\text{INNER}_j + \gamma_{2j}\text{RURAL}_j + \gamma_{3j}\text{URBAN}_j + \gamma_{4j}\text{MINORITY}_j + \eta_{mj},$$

where INNER_j, RURAL_j, and URBAN_j are indicators of the geographic location of the school; MINORITY_j is the proportion of students in the school who are Black; and η_{mj} is a Level 2 residual (random effect). Therefore, the objective of the statistical analysis was to estimate the five fixed effects (the γ_{mj}) determining the β_{ij}s and the corresponding between-school variance components (variances of the η_{mj}s). In the analysis of eighth-grade data reported here, the geographic location of the school was not available in the same format used in Grades 4 and 6, so we did not include geographic location of school as a school-level predictor.

We conducted separate analyses for each of the three dependent variables (California Test of Basic Skills [CTBS] mathematics, reading, and science test scores) at each of the three grade levels (Grades 4, 6, and 8). The statistical analysis was repeated nine times in each of the major analyses. In each case, the outcome variable was standardized so that the overall variance in the entire sample was one. Thus, the small class effects estimated are interpretable as the effects of being in a small class, expressed in standard deviation units.[2] We conducted similar analyses for follow-up data collected in Grades 5 and 7; since the results were quite similar, however, we do not report them here.

Results and Discussion

The results of each of the different types of analyses are reported in the subsections to follow, along with interpretive comments.

Analyses of the Effects of Small Classes in Grade 3

The most conventional analysis was an examination of the treatments as received by the students. The treatment group included students who attended 1 or more years of small classes. In this analysis, the small number of students who may have been initially assigned to receive a different treatment than the one they actually received were treated as if they were assigned to the treatment they received. We discuss later the issue of students who (contrary to the design of the experiment) switched from one treatment to another and therefore received treatments other than those to which they were initially assigned.

Attrition from the study. In this study, as in other longitudinal studies, there was attrition in that some of the students who began the study did not remain in the schools studied for one reason or another or were absent when the achievement tests were given. Attrition could be a source of bias in estimating treatment effects if the students who drop out of one treatment group are systematically different from those in other treatment groups. In particular, if the dropouts from the regular-sized classes evidenced higher average achievement than the dropouts from the small classes, then the attrition could lead to a positive bias in the estimated effect of small classes; that is, small classes could seem more favorable than they actually were. However, if the dropouts from the smaller classes evidenced higher achievement than those in the regular classes, then differential attrition could not make the small classes appear to have higher achievement.

There was some attrition from each year of the follow-up period; to simplify the presentation, however, we describe only the analysis for attrition from the (final) eighth-grade data set.[3] By the eighth grade, follow-up data could not be obtained on approximately one third of the students who were in the STAR study in Grade 3. Table 1 shows the average Stanford Achievement Test (SAT) mathematics, reading, and science achievement test scores for students who participated in Grade 3 of the STAR experiment, broken down by whether the eighth-grade test data were available. Those whose test scores were available provided the data analyzed in this study. The students whose eighth-grade data were not available were the "dropouts" whose characteristics could bias the results.[4]

Table 1 shows, for example, that 568 (29.3%) of the 1,938 third graders assigned to small classes (who had third-grade math test scores) dropped out of the study before they were tested in the eighth grade and that the average mathematics test score of the dropouts from small classes was 610.7, while the average mathematics test score of the dropouts from the regular-sized classes was 604.0.

TABLE 1

Mean Third-Grade Mathematics, Reading, and Science Achievement Scores of Students Who Were and Were Not Present in the Grade 8 Follow-Up Sample: Treatment Defined by Actual Assignment

	Present at follow-up				Not present at follow-up			
	Small class at Grade 3		Regular-sized class at Grade 3		Small class at Grade 3		Regular-sized class at Grade 3	
Grade 3 measure	*N*	*M*	*N*	*M*	*N*	*M*	*N*	*M*
SAT math	1,370	628.0	2,792	621.2	568	610.7	1,347	604.0
SAT reading	1,350	625.7	2,746	617.5	565	610.0	1,339	603.0
SAT science	1,435	630.7	2,844	623.0	621	618.1	1,421	608.6

The percentage of dropouts was slightly smaller in the small classes than in the regular-sized classes. Moreover, the average achievement of the students who dropped out of small classes was higher than the achievement of the students who dropped out of the larger classes.

The dropouts from both the small classes and the regular-sized classes appeared to have slightly lower average achievement than the students who were present in their respective class types at follow-up. However, the difference in achievement between the dropouts and those present for the follow-up from small classes was about the same as the difference between the mean achievement of the dropouts and those present at follow-up in the regular-sized classes.

The achievement gap between small and regular-sized classes, the crude treatment effect, was almost identical among dropouts and those present at follow-up. The differences between these gaps were not statistically significant for any of the three achievement areas. Thus, there was no difference in treatment effect at third grade between the dropouts and those present at follow-up. This suggests that differential attrition cannot explain any positive effect of small classes on achievement. That is, it is implausible that attrition made small classes appear more favorable than if there were no attrition.

Treatment effects. The results of the two-level analyses (in which treatment assignment was regarded as the treatment actually received) are summarized in Table 2. These results demonstrate that, across all schools, female students had significantly higher achievement in reading and mathematics in all grades and significantly lower achievement in science in Grades 4 and 8 (but not in Grade 6). Low-SES students evidenced significantly lower achievement in reading, mathematics, and science at all grade levels.

Perhaps more important for this experiment is that the average effect of small classes was statistically significant and positive for both mathematics and reading achievement at every grade level, ranging from 0.11 to 0.20 standard deviation units. The small class effect was positive for science achievement at all grades (ranging from 0.10 to 0.17 standard deviation units) and was statistically significant for both Grades 6 and 8. There was little evidence of interaction between gender and class size, meaning that small classes did not differentially favor one gender or the other. Moreover, except for science achievement in Grade 4, there was no evidence that small class effects varied across schools, since the between-school variance component was not statistically significant.

There is evidence that the mean level of achievement (adjusted for SES) varied across schools, as indicated by the statistically significant variance components for these effects. There seemed to be relatively little variation across schools in the effects of gender and the interaction of gender and small class size, as evidenced by the fact that this variance component was not statistically significant in any of eight analyses.

Few of the predictors in the between-school model other than the intercepts explained much of the variance in model coefficients. For example, none of the school-level predictors explained as much as 10% of the between-school variation in small class effects in any grade or subject matter. However, the percentage of minority students in a school was a significant predictor of SES-adjusted achievement at each grade level.

We carried out a series of other hierarchical linear model analyses using related analytic models; for example, we eliminated all of the predictors that were not statistically significant and examined slightly different codings of the variables. None of these analyses suggested results that were qualitatively different from those reported here. In each case, the small class effect had about the same positive magnitude and was statistically significant. There did not appear to be any consistent relation between the small class effects and any of the school-level predictors.

Analysis of the Effects of Treatments as Initially Assigned

In order to investigate the potential effects of students who switched from the treatment in which they were originally assigned to a different treatment, we conducted an analysis of the effects of treatments as they were initially assigned (regardless of the treatment received). If the treatment is not actually harmful, the actual effect should be underestimated. Here, as in the previous analysis, the treatment group included students with different numbers of years (and, in some cases, no years or only 1 year) of small classes.

Attrition from the study. Although the intention to treat analysis attempted to control for switching among treatment groups, it did not control for the effects of differential attrition. Consequently, we

had to examine the pattern of attrition as in the previous analysis. Since the definition of treatment groups in this analysis was not identical to that of the previous analysis, it was not necessary that the pattern of attrition be the same for these treatment groups as in the previous analysis. To determine whether differential attrition might be responsible for the positive estimates of treatment effects in this analysis, we computed mean third-grade achievement test scores for the students who dropped out and did not drop out of each treatment group before the follow-up.

To simplify the presentation, we describe only the analysis of attrition from the (final) eighth-grade data set.[5] By the eighth grade, follow-up data could not be obtained on approximately one third of the students who were in the STAR study in Grade 3. Table 3 shows the average third-grade SAT mathematics, reading, and science achievement test scores for students who participated in Project STAR, broken down by whether eighth-grade test data were available. Those whose test scores were available provided the data analyzed in this study. The students whose eighth-grade data were not available were the "dropouts" whose characteristics could bias the results. Table 3 shows, for example, that 490 (30.7%) of the 1,594 third graders initially assigned to small classes (who had third-grade math test scores) dropped out before they were tested in the eighth grade and that the average mathematics test score of the dropouts from small classes was 611.3, while the average mathematics test score of the dropouts from the regular-sized classes was 604.2.

It can be seen in Table 3 that the percentage of dropouts was slightly smaller among those initially assigned to small classes than among those initially assigned to regular-sized classes. Once again, the dropouts evidenced lower average achievement than those who were present at Grade 8, but the difference between those who dropped out and those who did not was about the same for students in the two assignment groups. Moreover, the average achievement of the students who dropped out of small classes was higher than the achievement of the students who dropped out of the regular-sized classes.

The achievement gap between small and regular-sized classes, the crude treatment effect, was almost identical among dropouts and those present at follow-up. The differences between these gaps were not statistically significant for any of the three achievement areas. Thus, there was no difference in treatment effect at third grade between the dropouts and those present at follow-up. As a result, it is implausible that attrition made small classes appear more favorable than if there were no attrition.

Treatment effects. The results of the two-level analyses (in which treatment assignment was regarded as the treatment initially assigned, regardless of the treatment actually received) are summarized in Table 4. The pattern of these results was quite similar to that obtained in the previous analysis (which defined treatment as the actual treatment received). The results demonstrate that, across all schools, female students had significantly higher achievement (adjusted for SES) in reading and mathematics at all grades and significantly lower achievement in science only in Grade 8. Low-SES students evidenced significantly lower achievement in reading, mathematics, and science at all grade levels. There was evidence that the mean level of achievement (adjusted for SES) varied across schools, as indicated by the statistically significant variance components for the intercept in models for achievement at all grade levels.

Perhaps more important for this experiment is that the average effect of small classes was statistically significant and positive for both mathematics and science achievement at every grade level, ranging from 0.13 to 0.22 standard deviation units. The small class effect was positive for reading achievement at all grades (ranging from 0.11 to 0.17 standard deviation units), and it was statistically significant at both Grades 4 and 8 and just missed statistical significance at Grade 6. Again, there was little evidence of interaction between gender and class size, meaning that small classes do not differentially favor one gender or the other. Moreover, there was no evidence that small class effects in mathematics and reading vary across schools, since the between-school variance component was not statistically significant. There was, however, statistically significant variation in small class effects on science achievement at Grades 4 and 6.

As in the previous analyses, few of the predictors in the between-school model other than the intercepts explained much of the variance in model coefficients. However, the percentage of minority students in a school was a significant predictor of SES-adjusted achievement in Grade 8 but not in Grades 4 or 6.

TABLE 2

Coefficients and Variance Components From Hierarchical Linear Modeling Analyses: Treatment Defined by Actual Assignment at Grade 3

	Grade 4			Grade 6			Grade 8		
	Math	Reading	Science	Math	Reading	Science	Math	Reading	Science
Intercept									
Intercept	–0.008	–0.009	–0.020	–0.023	–0.030	–0.020	–0.001	0.039	0.039
Inner-city school	0.056	–0.009	–0.156	0.027	–0.000	–0.146			
Rural school	0.016	–0.175*	–0.062	–0.118	–0.120	–0.168*			
Urban school	–0.051	–0.194	–0.069	–0.231	–0.174	–0.105			
Percentage of Black students	–0.452*	–0.723*	–0.658*	–0.328	–0.554*	–0.680*	–0.537*	–0.627*	–0.638*
Residual variance component	0.085*	0.036*	0.056*	0.095*	0.045*	0.030*	0.092*	0.054*	0.072*
Female									
Intercept	0.191*	0.135*	–0.083*	0.271*	0.192*	0.030	0.153*	0.117*	–0.157*
Inner-city school	–0.339	0.109	–0.149	–0.028	–0.195	–0.311*			
Rural school	–0.078	–0.123	–0.141	–0.024	–0.174	–0.220			
Urban school	–0.020	0.079	0.076	–0.052	0.047	0.043			
Percentage of Black students	0.020	–0.155	–0.030	0.069	–0.020	0.079	–0.030	–0.074	–0.018
Residual variance component	0.000	0.013	0.009	0.015	0.006	0.008	0.009	0.005	0.002
Low SES at Grade 3									
Intercept	–0.442*	–0.522*	–0.434*	–0.405*	–0.488*	–0.396*	–0.417*	–0.461*	–0.400*
Inner-city school	0.043	–0.108	0.078	0.054	0.013	0.051			
Rural school	0.093	–0.114	0.084	–0.041	0.047	0.062			
Urban school	0.151	–0.296	–0.101	0.012	0.031	0.054			
Percentage of Black students	0.237	0.041	0.181	0.135	0.302	0.211	0.284*	0.236*	0.164*
Residual variance component	0.003	0.013	0.011	0.007	0.015	0.027	0.001*	0.027*	0.001
Small class at Grade 3									
Intercept	0.126*	0.112*	0.098	0.203*	0.126*	0.167*	0.158*	0.133*	0.140*
Inner-city school	–0.505*	–0.243	–0.263	0.216	–0.038	0.160			
Rural school	–0.163	–0.161	–0.163	–0.068	–0.123	–0.106			
Urban school	–0.140	–0.169	–0.053	–0.041	–0.032	–0.016			
Percentage of Black students	0.330	0.189	0.054	–0.109	–0.088	–0.127	–0.103	–0.066	–0.088
Residual variance component	0.020	0.029	0.053*	0.015	0.034	0.028	0.004	0.023	0.018

TABLE 2
(continued)

	Grade 4			Grade 6			Grade 8		
	Math	Reading	Science	Math	Reading	Science	Math	Reading	Science
Female–small class interaction									
Intercept	–0.024	0.013	0.008	–0.185*	–0.050	–0.128	–0.084	–0.008	–0.027
Inner-city school	0.778*	0.037	0.187	–0.253	–0.040	–0.146			
Rural school	0.120	0.148	0.194	–0.073	0.187	0.164			
Urban school	0.063	0.075	0.066	0.031	0.023	0.047			
Percentage of Black students	–0.317	0.093	0.261	0.182	0.203	0.233	0.194	0.097	0.200
Residual variance component	0.013	0.024	0.042	0.019	0.009	0.015	0.007	0.005	0.030

* $p < .05$.

TABLE 3
Mean Third-Grade Mathematics, Reading, and Science Achievement Scores of Students Who Were and Were Not Present in the Grade 8 Follow-Up Sample: Treatment Defined by Initial Assignment

	Present at follow-up				Not present at follow-up			
	Initially assigned to small class		Initially assigned to regular class		Initially assigned to small class		Initially assigned to regular class	
Grade 3 measure	*N*	*M*	*N*	*M*	*N*	*M*	*N*	*M*
SAT math	1,104	629.1	3,058	621.4	490	611.3	1,425	604.2
SAT reading	1,089	627.1	3,007	617.7	486	610.4	1,418	603.3
SAT science	1,160	630.9	3,119	623.6	537	618.5	1,505	609.0

Again, we carried out a series of other hierarchical linear model analyses using related analytic models; for example, we eliminated all of the predictors that were not statistically significant and examined slightly different codings of the variables. None of these analyses suggested results that were qualitatively different from those reported here. In each case, the small class effect had about the same positive magnitude and was statistically significant. There did not appear to be any consistent relation between small class effects and any of the school-level predictors.

Analyses of the Effects of Different Amounts of Exposure to Small Classes

In order to get some idea of whether longer exposure to small classes produced greater effects than fewer years of small classes, we conducted four analyses in which the treatment group was defined as those who received 1 or more, 2 or more, 3 or more, or 4 years of small classes. In each case, the comparison group consisted of all other children in the experiment, regardless of whether they were assigned to the regular-sized class group or whether they had entered the experiment later than kindergarten, been assigned to the small class group at some point, and received up to 3 years of small classes.

Attrition from the study. Since the definition of treatment groups in these analyses was quite different from that of the previous analyses, it was not necessary that the pattern of attrition be the same for these treatment groups as in the previous analyses. To determine whether differential attrition might have biased the results of estimates of treatment effects in this analysis, we computed the mean third-grade achievement test scores for the students who dropped out and did not drop out of each treatment group before the follow-up (Grade 4, 6, or 8).

The results of these analyses were substantially the same as those of the previous analyses of attrition. The dropout rate was smaller among the small classes, and dropouts from small classes evidenced higher achievement in third grade than did dropouts from regular-sized classes, particularly when the treatment group was defined as having several years of small classes. This was not surprising since treatment group students had to remain in the same school during the several-year period in question. It seems highly plausible that students who had low mobility during this period would also have low mobility from Grades 4 to 8 and therefore be more likely to be captured in the follow-up sample. The achievement gap between small and regular-sized classes was almost identical among dropouts and those present at follow-up in each analysis. None of the differences between these gaps were statistically significant, and thus it is implausible that attrition made small classes look more favorable than if there were no attrition in any of the analyses.

Treatment effects. The results of the two-level analyses (in which the small class group was regarded as those who received 1 or more, 2 or more, 3 or more, or 4 years of small classes) are summarized in Table 5. Since the results of many of these analyses were similar to those in other analyses, we present only the estimates of the small class effects.

The average effect of small classes was statistically significant and positive for mathematics and reading achievement at every grade level. The effect of small classes was statistically significant for science achievement only among those who had at least 2 years of small classes. Moreover, the effects become larger the longer students are in small classes; the effects were roughly twice as large for the students who had 4 years of small classes as for

students who had as few as 1 year of small classes. As in the previous analyses, few of the predictors in the between-school model other than the intercepts explained much of the variance in model coefficients.

It is important to recognize that, in this analysis, students were not randomly assigned to receive 4 (as opposed to 1, 2, or 3) years of small classes. Therefore, the students who did receive 4 years of small classes, for example, were likely to be different (in ways not entirely accounted for by race, gender, and social class) from those who did not. For instance, they could have come from families less likely to move. Assuming that students who come from more stable homes may have higher levels of achievement, these students might be better prepared to benefit from small classes, which could exaggerate effects. On the other hand, the group against which the students who received 4 years of small classes were compared included substantial numbers of students who received at least some small classes and who received as many as 3 years of small classes, which would tend to reduce the apparent effects of small classes.[6]

Comparisons With Effects at Grade 3

One of the purposes of this study was to determine whether the effects of small classes fade after 5 years. While the analyses reported here provide evidence that the effects were not null in Grades 4, 6, and 8, they did not directly address the question of whether effects have decreased substantially from those observed at the end of the STAR experiment when the students were in Grade 3. Table 6 gives the small class effects estimated at the end of Grades 4, 6, and 8, along with the effects estimated from similar analyses at the end of Grade 3 (from Nye et al., 1999). Note, however, that the analysis at Grade 3 was based on a slightly larger group of students than those in the analyses in Grades 4, 6, and 8, since test scores for some students that were available at Grade 3 were not available at one or more of the follow-ups.

First consider the results when the class size used in the analysis was defined by actual class size at Grade 3 (regardless of initial assignment) or by initial assignment (regardless of the actual class size), shown in the first two rows of each panel of Table 6. The effects in mathematics were almost identical at the posttest in Grade 3 and at the eighth-grade follow-up 5 years later. The effects in science were smaller in eighth grade than in third grade. In reading, the effect defining class size by actual class size was smaller, but the effect defined by initial assignment was actually larger at Grade 8 than at Grade 3. Note that the largest change over the 5-year period was about 1.5 standard errors of estimate.

The changes in the small class effects among the group that received 4 years of small classes appeared to be somewhat different. There seemed to be little change in the effect for mathematics achievement, a slight decrease (about 16%) for reading achievement, and an actual increase for science achievement (25%). This may indicate that the effects of small classes experienced over 4 years were larger and more lasting. However, interpretation of this finding is complicated because (as we mentioned earlier) not all of the factors determining which students received 4 (as opposed to fewer) years of small classes were controlled as part of the experiment.

Taken together, these findings suggest that, while the small class effect may have been somewhat smaller at Grade 8 follow-up than at Grade 3 posttest, the effect remained substantial in comparison with the initial effect (no less than 70% of the initial effect). That is, the small class effect on achievement may diminish somewhat, but it definitely does not fade to statistical or practical insignificance after 5 years.

Conclusions

The STAR experiment demonstrated that small classes lead to significantly higher achievement for students in reading and mathematics. Analyses of the data collected by the Lasting Benefits Study demonstrate that the positive effects of small classes in early grades result in mathematics, reading, and science achievement gains that persist at least through the eighth grade. As in the case of the initial effects, these effects were remarkably consistent across schools, so lasting benefits were found for all kinds of students in all kinds of schools. Moreover, the achievement effects were similar in magnitude to those observed in Grade 3. That is, the effects of small classes in kindergarten through Grade 3 on achievement did not appear to be "fading out" by Grade 8.

This 5-year follow-up study was also subject to substantial attrition, but the students who dropped out of the small classes actually evidenced higher achievement than those who dropped out of the larger classes, suggesting that the observed differ-

TABLE 4

Coefficients and Variance Components From Hierarchical Linear Modeling Analyses: Treatment Defined by Initial Assignment

	Grade 4			Grade 6			Grade 8		
	Math	Reading	Science	Math	Reading	Science	Math	Reading	Science
Intercept									
Intercept	–0.006	–0.010	–0.018	–0.025	–0.029	–0.020	0.000	0.039	0.039
Inner-city school	0.089	0.010	–0.154	0.022	–0.001	–0.166			
Rural school	0.012	–0.171*	–0.065	–0.117	–0.117	–0.163			
Urban school	–0.042	–0.184	–0.061	–0.217	–0.164	–0.090			
Percentage of Black students	–0.470*	–0.733*	–0.657*	–0.330	–0.545*	–0.661*	–0.536*	–0.622*	–0.635*
Residual variance component	0.085*	0.036*	0.056*	0.098*	0.045*	0.031*	0.093*	0.052*	0.067*
Female									
Intercept	0.204*	0.146*	–0.068	0.262*	0.188*	0.013	0.152*	0.124*	–0.143*
Inner-city school	–0.258	0.115	–0.138	–0.089	–0.233	–0.336*			
Rural school	–0.102	–0.139	–0.154	–0.061	–0.154	–0.233*			
Urban school	–0.035	0.013	0.036	–0.112	–0.005	0.017			
Percentage of Black students	0.010	–0.123	0.000	0.045	0.049	0.067	–0.013	–0.042	–0.008
Residual variance component	0.019	0.010	0.013	0.017	0.003	0.019	0.010	0.012	0.004
Low SES at Grade 3									
Intercept	–0.443*	–0.528*	–0.435*	–0.401*	–0.494*	–0.400*	–0.423*	–0.463*	–0.405*
Inner-city school	0.015	–0.113	0.073	0.074	0.019	0.075			
Rural school	0.080	–0.094	0.076	–0.034	0.058	0.070			
Urban school	0.137	–0.262	–0.104	0.027	0.040	0.081			
Percentage of Black students	0.216	0.061	0.171	0.143	0.304	0.209	0.293*	0.222*	0.152*
Residual variance component	0.009	0.007	0.019*	0.005	0.013	0.022	0.006	0.028*	0.013
Small class at Grade 3									
Intercept	0.137*	0.128*	0.128*	0.218*	0.110	0.134*	0.146*	0.173*	0.178*
Inner-city school	–0.379	–0.191	–0.324	0.021	–0.133	0.008			
Rural school	–0.191	–0.131	–0.183	–0.157	–0.116	–0.152			
Urban school	–0.135	–0.183	–0.085	–0.164	–0.112	–0.147			
Percentage of Black students	–0.250	0.284	0.147	–0.147	0.065	–0.171	–0.033	0.103	–0.015
Residual variance component	0.020	0.008	0.081*	0.010	0.050*	0.072*	0.021	0.019	0.018

TABLE 4
(continued)

	Grade 4			Grade 6			Grade 8		
	Math	Reading	Science	Math	Reading	Science	Math	Reading	Science
Female–small class interaction									
Intercept	–0.081	–0.025	–0.058	–0.203*	–0.058	–0.118	–0.105	–0.039	–0.094
Inner-city school	0.676*	0.031	0.364	–0.019	0.128	–0.037			
Rural school	0.230	0.227	0.273	0.072	0.151	0.259			
Urban school	0.082	0.263	0.161	0.282	0.174	0.134			
Percentage of Black students	–0.306	–0.026	0.047	0.284	–0.043	0.325	0.142	–0.083	0.130
Residual variance component	0.023	0.006	0.086	0.010	0.024	0.080	0.032	0.012	0.028

* $p < .05$.

TABLE 5
Average Cumulative Effects of Small Class Assignment Across Grades

	Grade 4			Grade 6			Grade 8		
	Math	Reading	Science	Math	Reading	Science	Math	Reading	Science
Small class in any grade: intercept	0.134*	0.116*	0.114*	0.202*	0.140*	0.189*	0.059	0.071	0.078
Small class in two or more grades: intercept	0.242*	0.229*	0.203*	0.237*	0.222*	0.237*	0.204*	0.185*	0.160*
Small class in three or more grades: intercept	0.292*	0.268*	0.227*	0.290*	0.249*	0.239*	0.283*	0.261*	0.249*
Small class in all grades: intercept	0.344*	0.366*	0.329*	0.412*	0.373*	0.363*	0.368*	0.324*	0.300*

* $p < .05$.

TABLE 6
Summary of Small Class Effects at the End of the STAR Experiment (Grade 3) and for 5 Years Thereafter

Definition of small classes	Grade 3	(*SE*)	4	(*SE*)	6	(*SE*)	8	(*SE*)
Mathematics								
Actual assignment	0.150	(0.064)	0.126	(0.046)	0.203	(0.051)	0.158	(0.042)
Initial assignment	0.141	(0.046)	0.137	(0.048)	0.218	(0.053)	0.146	(0.046)
All grades K–3	0.352	(0.065)	0.344	(0.057)	0.412	(0.069)	0.368	(0.058)
Reading								
Actual assignment	0.175	(0.057)	0.112	(0.048)	0.126	(0.053)	0.133	(0.044)
Initial assignment	0.154	(0.044)	0.128	(0.046)	0.110	(0.058)	0.173	(0.046)
All grades K–3	0.386	(0.064)	0.366	(0.062)	0.373	(0.070)	0.324	(0.057)
Science								
Actual assignment	0.176	(0.044)	0.098	(0.052)	0.167	(0.052)	0.140	(0.043)
Initial assignment	0.184	(0.043)	0.128	(0.058)	0.134	(0.060)	0.178	(0.045)
All grades K–3	0.400	(0.059)	0.329	(0.067)	0.363	(0.066)	0.300	(0.057)

Note. Grade 3 effects are from Nye, Hedges, and Konstantopoulos (1999).

ences in achievement between students who had been in small and larger classes were not due to attrition. The effect of switching between classes is more difficult to determine qualitatively, but an analysis estimating the small class effect using the initial assignment of students resulted in estimates of small class effects that were almost identical to those obtained when actual class type was used.

The STAR and Lasting Benefits studies provide important (and perhaps the strongest) pieces of the converging evidence about the effectiveness of small classes in promoting achievement. The magnitudes of effects in the STAR study are quite consistent with those obtained in small-scale randomized experiments (whose generalizability might be questioned) and with the results of econometric studies (whose internal validity might be questioned). The present study provides evidence that these effects do not disappear over time but, rather, provide students from small classes in early grades with achievement benefits that last at least until high school. Together, all of this evidence points to positive effects of small classes on achievement that are large enough and of sufficient duration to support policies of reduction of class size to result in small-sized (15–17 pupils) classes in the primary grades.

This study also demonstrates that students who experience more years of small classes in kindergarten through Grade 3 have higher levels of achievement (adjusted for social class) 5 years later than students who have fewer years of small classes. While this does not represent definitive evidence that the effects of small classes compound, it is strongly suggestive. Therefore, it is likely that the lasting benefits increase with the number of years in small classes.

This research has not answered all of the important questions about the effects of small classes on achievement and other desirable outcomes of schooling. For example, the mechanisms by which small classes lead to higher achievement are not clear (although there are several obvious hypotheses). Understanding of the mechanism could lead to more effective ways to implement class size reductions and to improvements in their effectiveness. Such understanding is obviously desirable. Similarly, the effects of small changes in class size or of altering teacher-pupil ratios (without necessarily changing class sizes) are not immediately obvious from this experiment.

The exact effects of small classes on achievement growth are not yet clear. While it appears that students in small classes experience an increase in achievement that persists over time, growth trajectories of students from small and regular classes have not been studied in detail. This, too, would be desirable.

Other questions concern the effects of small classes on the other outcomes of schooling, such as student engagement, motivation, and persistence. Of particular concern is whether students who were in small classes in early grades will be more likely to graduate from high school, more likely to go to college, and more likely to partici-

pate successfully in the labor force. Some of these questions may ultimately be answered through the use of data that is continuing to be collected by the Lasting Benefits Study.

Notes

[1]The possible effects of attrition within Project STAR before Grade 3 were considered in a separate paper and found to be negligible (Nye et al., 1999).

[2]The choice of the overall standard deviation for standardization, rather than the within-class standard deviation (which would be smaller), results in smaller standardized effects than would have been obtained with the (pooled) within-class standard deviation.

[3]The pattern of attrition to Grades 4 and 6 was similar.

[4]These dropouts did not leave school but either transferred to a school not involved in the STAR study or were absent on the days that testing occurred, and thus their eighth-grade achievement test scores were not available.

[5]The pattern of attrition to Grades 4 and 6 was similar.

[6]The exact numbers varied by year of follow-up and outcome variable.

References

Barnett, W. S. (1985). Benefit-cost analysis of the Perry Preschool Program and its policy implications. *Educational Evaluation and Policy Analysis, 7,* 333–342.

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

Barnett, W. S. (1995). The long term effects of early childhood programs on cognitive and school outcomes. *The Future of Children: Long-Term Outcomes of Early Childhood Programs, 5,* 25–250.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models.* Newbury Park, CA: Sage.

Educational Research Service. (1980). *Class size research: A critique of recent meta-analysis.* Arlington, VA: Author.

Finn, J. D. & Achilles, C. M. (1990). Answers and questions about class size: A statewide experiment. *American Educational Research Journal, 27,* 557–577.

Glass, G. V., Cahen, L. S., Smith, M. L., & Filby, N. N. (1982). *School class size: Research and policy.* Beverly Hills, CA: Sage.

Glass, G. V., & Smith, M. E. (1979). Meta-analysis of research on class size and achievement. *Educational Evaluation and Policy Analysis, 1,* 2–16.

Greenwald, R., Hedges, L. V., & Laine, R. (1996). The effect of school resources on student achievement. *Review of Educational Research, 66,* 361–396.

Hanushek, E. A. (1986). The economics of schooling: Production and efficiency in public schools. *Journal of Economic Literature, 24,* 1141–1177.

Hanushek, E. A. (1989). The impact of differential school expenditures on school performance. *Educational Researcher, 18*(4), 45–65.

Haskins, R. (1989). Beyond metaphor: The efficacy of early childhood education. *American Psychologist, 44,* 274–282.

Hedges, L. V., Laine, R. D., & Greenwald, R. (1994). Does money matter? A meta-analysis of the effects of differential school inputs on student outcomes. *Educational Researcher, 23*(3), 5–14.

Hedges, L. V. & Stock, W. (1983). The effects of class size: An examination of rival hypotheses. *American Educational Research Journal, 20,* 63–85.

Krueger, A. (1998). *Experimental estimates of education production functions* (NBER Working Paper No. 379). Princeton, NJ: Princeton University, Industrial Relations Section.

McKey, R. H., Condelli, L., Ganson, H., Barret, B. J., McConkey, C., & Plantz, M. C. (1985). *The impact of Head Start on children, families, and communities* (DHHS Publication No. OHDS 85-31193). Washington, DC: U.S. Government Printing Office.

Mosteller, F. (1995). The Tennessee study of class size in the early school grades. *The Future of Children, 5,* 113–127.

Mosteller, F., Light, R. J., & Sachs, J. A. (1996). Sustained inquiry in education: Lessons learned from skill grouping and class size. *Harvard Educational Review, 66,* 797–842.

Nye, B. A., Hedges, L. V., & Konstantopoulos, S. (1999). *The effects of small classes on academic achievement: The results of the Tennessee class size experiment.* Manuscript under review.

Nye, B. A., Zaharias, J. B., Fulton, B. D., Achilles, C. M., & Cain, V. A. (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.* Nashville: Tennessee State University, Center of Excellence for Research in Basic Skills.

Slavin, R. (1984). Meta-analysis in education: How has it been used? *Educational Researcher, 13*(8), 6–15, 24–25.

White, K. R. (1985). The efficacy of early intervention. *Journal of Special Education, 19,* 401–416.

Word, E., Johnston, J., Bain, H., Fulton, D. B., Boyd-Zaharias, J., Lintz, M. N., Achilles, C. M., Folger, J., & Breda, C. (1990). *Student/Teacher Achievement Ratio (STAR): Tennessee's K–3 class-size study.* Nashville: Tennessee Department of Education.

Authors

BARBARA NYE is a the principal investigator for longitudinal research on class size and Executive Director of the Center of Excellence for Research and Policy on Basic Skills, Tennessee State University, 330 10th Avenue, North Box 141, Nashville, TN 37203-3401. She

specializes in research on academic achievement, early intervention, and evaluation and implementation of systemic reform of K–12 and higher education curriculum/programs.

LARRY V. HEDGES is the Stella M. Rowley Professor of Education, Psychology, and Sociology at the University of Chicago, 5835 S. Kimbark Avenue, Chicago, IL 60637. His specialties are statistical methods for social science research and policy analysis.

SPYROS KONSTANTOPOULOS is a graduate student in the Department of Education, University of Chicago, 5835 S. Kimbark Avenue, Chicago, IL 60637. He specializes in statistical methods for social science research.

Manuscript received June 30, 1998
Revision received March 1, 1999
Accepted March 2, 1999