American Educational Research Association

Some Findings from an Independent Investigation of the Tennessee STAR Experiment and from Other Investigations of Class Size Effects
Author(s): Eric A. Hanushek
Source: *Educational Evaluation and Policy Analysis*, Vol. 21, No. 2, Special Issue: Class Size: Issues and New Findings (Summer, 1999), pp. 143-163
Published by: American Educational Research Association
Stable URL: http://www.jstor.org/stable/1164297
Accessed: 19/10/2009 16:51

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at http://www.jstor.org/page/info/about/policies/terms.jsp. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at http://www.jstor.org/action/showPublisher?publisherCode=aera.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

American Educational Research Association is collaborating with JSTOR to digitize, preserve and extend access to *Educational Evaluation and Policy Analysis.*

http://www.jstor.org

Educational Evaluation and Policy Analysis
Summer1999, Vol. 21, No. 2, pp. 143–163

Some Findings From an Independent Investigation of the Tennessee STAR Experiment and From Other Investigations of Class Size Effects

Eric A. Hanushek
University of Rochester and National Bureau of Economic Research

While random-assignment experiments have considerable conceptual appeal, the validity and reliability of results depends crucially on a number of design and implementation issues. This paper reviews the major experiment in class size reduction—Tennessee's Project STAR—and puts the results in the context of existing nonexperimental evidence about class size. The nonexperimental evidence uniformly indicates no consistent improvement in achievement with class size reductions. This evidence comes from very different sources and methodologies, making the consistency of results quite striking. The experimental evidence from the STAR experiment is typically cited as providing strong support of current policy proposals to reduce class size. Detailed review of the evidence, however, uncovers a number of important design and implementation issues that suggest considerable uncertainty about the magnitude of any treatment effects. Moreover, there is reason to believe that the commonly cited results are biased upwards. Ignoring consideration of the uncertainties and possible biases in the experiment, the results show effects that are limited to very large (and expensive) reductions in kindergarten or possibly first grade class sizes. No support for smaller reductions in class size (i.e., reductions resulting in class sizes greater than 13–17 students) or for reductions in later grades is found in the STAR results.

The latest debate about the efficacy of class size reductions for improving student performance has renewed interest in the underlying evidence. The debate has been surprisingly lively, given the amount of research that has gone into understanding the impact of varying class sizes. From the vast magnitude of evidence on the subject, one specific piece—evidence from Project STAR, a random-assignment experiment in Tennessee conducted during the mid-1980s—has, with some justification, assumed disproportionate weight in the discussions.[1] Unfortunately, the precise nature of that evidence has been obscured in the public discussion of the policy proposals.

The policy debates surrounding class size were pushed to new heights with the aggressive incentives for reductions introduced in California in 1996. Propelled by the political popularity of this initiative, pressures to reduce class size have been felt across the states. And the negotiations over the federal budget for fiscal year 1999 hinged on inclusion of funds for broad class size reductions, even though educational policy is chiefly the province of the states. The policy debate, in turn, stimulated a reconsideration of the evidentiary base for class size policies.

Because the desirability of having smaller classes seems so obvious, the natural discussion would seem to hinge on whether or not class size reductions are worth their expense. Perhaps surprisingly, however, the current debate has seldom involved discussions of cost-effectiveness but instead has concentrated on the prior question of whether or not general class size reductions can be expected to yield significant performance gains, without regard to their costs. This article provides a brief review and assessment of the various pieces of relevant evidence. Because of the extent of available evidence and because of variation in the findings, the policy interpretations have largely come down

to how one should weight the different pieces of evidence—with proponents of class size reductions generally arguing implicitly or explicitly that the only relevant evidence is that from Project STAR. This analysis, which takes the view that all of the evidence should contribute to any conclusions, begins with an account of the findings of various components of the nonexperimental evidence. It then turns to a more detailed discussion of the STAR experiment and the evidence derived from it.

The nonexperimental evidence does not suggest that any substantial achievement gains would accrue to general class size reduction policies of the type recently discussed and implemented in various jurisdictions around the United States. Both pupil-teacher ratios and class sizes have fallen over some period of time, without any commensurate improvement in overall student performance. International test score evidence likewise suggests little if any aggregate relationship between the intensity of instruction and student achievement. And the largest body of evidence, that derived from detailed econometric investigations of student performance, provides little reason to support broad reductions in class size. Thus, the support for broad class size reduction policies that currently exists rests exclusively on the experimental evidence, particularly evidence from the Tennessee experiment.

The Tennessee experiment, while not the first experiment in reduced class sizes, is noteworthy for its scope and for its methodology. Because more than 6,000 students were assigned to small and large classes in kindergarten and this assignment was continued through the third grade, extensive data are available to shed light on the effects of small class settings on student performance. The now well-known results from STAR show students in small kindergarten classes on average outperforming those in larger kindergarten classes and show this aggregate performance gap persisting throughout the elementary grades.

Translation of these general research conclusions into policy statements must, nonetheless, be tempered by a set of less known uncertainties arising from the conduct of the Tennessee experiment and from disaggregation of the results. First, a number of design and implementation issues affect the inferences that can be drawn, making it clear that this is not the decisive evidence some have claimed it to be. Second, above the gains from an initial exposure to small classes, small classes do not lead to any further improvements in performance. Third, the apparent gains are only known to come from much larger reductions in class size than currently being discussed. Fourth, a substantial proportion of the schools in the experiment show student performance in small classes that is worse than performance in large classes, undoubtedly reflecting variations in teacher quality that are more important than any class size effects. Fifth, the costs of broad class size reductions are seldom if ever put into the context of other potential uses of funds. Simply put, the desirability of the research approach—random assignment experimentation—should not be confused with the reliability of this specific implementation or with the possible policy conclusions that can be supported by this specific experiment.

Nonexperimental Evidence

The bulk of evidence about class size reductions comes from analysis of nonexperimental data generated by the normal operations of schools.[2] Both cross-sectional and intertemporal variations in class size and teacher allocations offer potential insights into the impact of class size policies on student performance. While acknowledging the uncertainties inherent in the analysis of nonexperimental data, the striking aspect of the combined evidence on class size is the consistency with which it points to no systematic effects of class size reductions within the relevant policy range.

Aggregate Evidence

One undeniable feature of 20th-century U.S. schools has been the steady decline in pupil-teacher ratios. The increases in teacher intensity over the past three decades have been much larger than most current policy proposals to reduce class size. Given the magnitude of changes, one would expect a discernible effect on aggregate student performance, unless it is offset by opposing trends of similar magnitude.

Concentrating on the last half of the century, we see that average pupil-teacher ratios fell from about 27:1 to 17:1, or 35%, between 1950 and 1995 (U.S. Department of Education, 1997). These declines have clear and powerful implications for school costs, because salary expenditure is the largest budget item and increasing the intensity of teacher usage simply magnifies these costs. For example, Hanushek and Rivkin (1997) calculate

that 85% of the increase in instructional costs over the period 1970–1990 came from reductions in pupil-teacher ratios.

What has happened to student performance over this time? While we lack information about student achievement for the entire period, the information that we have from 1970 for the National Assessment of Educational Progress (NAEP) indicates that our 17-year-olds were performing roughly the same in 1996 as in 1970.[3] There are some differences by subject area. For science, the average scale score of 17-year-olds *fell* 9 points between 1969 and 1996. For math, 17-year-olds improved 3 points between 1973 and 1996. For reading, they improved 2 points between 1971 and 1996. Writing performance, available only since 1984, shows a fall of 7 points, by 1996. Only the fall in science (and in writing since 1984) represents a statistically significant difference. There have been improvements in NAEP scores for younger students, but they are not maintained and are not reflected in the skills that students take to college and to the job market. In summary, the overall picture is one of stagnant performance.

The aggregate data present a prima facie case that overall class size reductions are unlikely to lead to improvements in student performance, but there are several reasons why these data could be misleading. First, pupil-teacher ratios are not the same as class size, so class size per se may not have followed the same pattern as pupil-teacher ratios. Second, changes in the student population that affect the preparation or motivation of students could distort or mask any effects of class size changes. While these factors could each have some influence, the available evidence suggests that they are insufficient to reverse the aggregate picture.

The only data on teacher intensity that are available over long periods refer to pupil-teacher ratios. Pupil-teacher ratios are readily calculated from normal administrative data, and they have been available for the entire 20th century. Pupil-teacher ratios, however, differ from class size in a variety of ways. For example, if there are specialist teachers (e.g., music or bilingual teachers), if teachers typically meet a different number of classes than students generally take, or if teachers are assigned purely administrative duties, pupil-teacher ratios will differ from average class size. In fact, recent data for the country as a whole show that pupil-teacher ratios of approximately 17:1 have been less than the estimated average class size of approximately 24 (Lewit & Baker, 1997). For the aggregate picture, nonetheless, the issue is whether or not class sizes have tended to move with pupil-teacher ratios.[4] The best available evidence is that they do tend to move together in the aggregate (Lewit & Baker, 1997).[5]

The largest concern about the divergence of pupil-teacher ratios and class sizes that is typically raised concerns the expansion of special education instruction.[6] The growth in students with identified handicaps, coupled with legal requirements for providing educational services for them, has increased the size of the special education sector. Therefore, the expansion of the more highly staff-intensive special education sector could reduce the overall pupil-teacher ratio without commensurate decreases in mainstream class sizes. To the extent that mandated expenditure for disabled students is driving the fall in the pupil-teacher ratio, regular class sizes are not declining by much, and, by extension, one might not expect any improvement in measured student performance.[7]

The Education for All Handicapped Children Act of 1975[8] prescribed a series of diagnostics, counseling activities, and services to be provided for disabled students. Over time, there has been clear growth in the proportion of students classified as the special education population. The proportion of students in special education grew from about 8% in 1976 to more than 12% in 1995 (U.S. Department of Education, 1997). The number of special education teachers is rising even more rapidly than the student population. The growth in both the size of the special education population and the intensity of instruction could distort the picture of changes in average class sizes.

From the aggregate trends, it is difficult to discern any significant or distinct effect of special education legislation on the general pattern of pupil-teacher ratios, in part simply reflecting its limited overall scope (cf. Lewit & Baker, 1997). Hanushek and Rivkin (1997) provide an upper bound on how much the changes in special education could have affected the observed pupil-teacher ratios during the 1980s and conclude that no more than one third of the change could be due to changes in special education. Thus, the aggregate picture of mismatch between teacher intensity and student performance cannot be attributed simply to overall changes in special education.

The second concern with the aggregate evidence is that the student population might have gotten

worse over time in terms of motivation or preparation, so more intensive instruction would be needed just to hold even in performance. For example, between 1970 and 1990, children living in poverty families rose from 14.9% to 19.9%, while children living with both parents declined from 85% to 73%. Over the same period, however, there were offsetting trends. Adults 25–29 years old with a high school or greater level of schooling went from 74% to 86% (up from 61% in 1960). Moreover, among all families with children, the percentage with three or more children fell from 36% to 20%.

It is difficult to know how to net out these opposing trends with any accuracy. While differences in families are very important for student achievement, most studies of student performance have not focused their primary attention on families and have not explicitly dealt with measuring and testing the importance of specific aspects of family inputs. Mayer (1997) suggests that the direct causal impact of family income might be fairly small and that the past works have more identified associations than true causal impacts. This analysis, nonetheless, cannot conclusively indicate whether or not there have been trends in the underlying causal factors (that are correlated in cross sections with income). Grissmer, Kirby, Berends, and Williamson (1994) do attempt to sort out the various factors. Using econometric techniques to estimate how various family factors influence children's achievement, they apply cross-sectionally estimated regression coefficients as weights to the trended family background factors identified earlier. Their overall findings are that Black students performed better over time than would be expected from the trends in Black family factors, but White students performed worse over time than would be expected. In other words, for the nation as a whole, student backgrounds appear to have improved, not gotten worse.[9]

Thus, while changes in family inputs make it possible that a portion of the increased school resources has gone to offset adverse factors, the evidence is quite inconclusive about even the direction of any trend effects, let alone the magnitude. The only available quantitative estimates indicate that changing family effects are unable to offset the large observed changes in pupil-teacher ratios and school resources. Indeed, for the nation as a whole, these trends are estimated to have worked in the opposite direction, making the performance of schools appear better than it was.

Grissmer, Flanagan, and Williamson (1998) extend the prior trend analysis in Black-White achievement to argue that reduced pupil-teacher ratios might explain a portion of the relative gain by Black students during the 1980s. Since the detailed econometric analysis of Cook and Evans (1996) demonstrates that variations in school level resources do not explain the changes in relative NAEP performance, the simple trend analysis of Grissmer et al. (1998) is plausible only if Blacks are much more sensitive to variations in pupil-teacher ratios than Whites. (Even accepting this, the trend analysis has difficulty explaining why relative convergence of scores stopped during the 1990s). In any event, since Whites comprise 80% of the student population and receive a roughly proportionate share of any past (or proposed) reductions in class size, the overall aggregate findings provide the appropriate evidence about potential effects of general policy proposals.

In summary, a very large natural experiment in class size reduction has been ongoing for a long period of time, and, although producing larger aggregate policy changes than typically advocated today, overall achievement data do not suggest that it has been a productive policy to pursue.

International Evidence

Similar kinds of results are found if one looks across countries at the relationship between pupil-teacher ratios and student performance. While it is clearly difficult to develop standardized data across countries, to control for the many differences in populations and schools, and to describe actual classroom organization, international variations in class sizes and pupil-teacher ratios are larger than those found within the United States and thus offer some promise for detecting effects.

The Third International Mathematics and Science Study, conducted in 1995, provides mathematics and science tests for a group of voluntarily participating nations. To highlight the role of pupil-teacher ratios, the eighth-grade math and science scores can be correlated with the primary school pupil-teacher ratio in each country.[10] For the 17 nations with consistent test and pupil-teacher ratio data, there is a *positive* relationship between pupil-teacher ratio and test scores, and it is statistically significant at the 10% level for both tests (although the statistical significance disappears when Korea, the sampled country with the largest pupil-teacher ratio, is dropped).

A more systematic attempt to investigate the relationship between student performance and pupil-teacher ratios uses the six prior international tests in math or science given between 1960 and 1990 (Hanushek & Kim, 1996). When the 70 country-test-specific observations of test performance that are available are used, there is a positive but statistically insignificant effect of pupil-teacher ratios on performance after allowing for differences in parental schooling. Again, while there are very large differences in pupil-teacher ratios, they do not show up as significantly influencing student performance.

Uniform data are not available on international differences in class sizes, but some intensive investigations have shown that class size differences vary more internationally than pupil-teacher ratios. Specifically, Japan and the United States have quite similar pupil-teacher ratios, but, because of choices in how to organize schools and to use teachers, Japanese class sizes are much larger than U.S. class sizes (Stevenson & Stigler, 1992). Japanese student performance is, on average, much better than U.S. student performance.

In summary, the very large international differences in teacher intensity provide no evidence of systematic influence on student performance, as measured by common math and science tests.

Econometric Evidence

Beginning with the "Coleman Report" (Coleman et al., 1966), there has been an intensive effort to identify the effects of school resources on student performance. The large body of literature that has accumulated over the past three decades provides a number of insights into the relationship between resources and achievement. The overall picture of school resources has been developed elsewhere (Hanushek, 1997), and this discussion refers only to the relevant findings for class size.

The econometric estimates relate class size or teacher intensity to measures of student performance while also separating out the influence of family and other inputs into education. The precise sampling, specification of the relationships, measurement of student performance, and estimation techniques differ across studies, but here I concentrate on the summary of any relationship across studies.[11]

The econometric studies of the determinants of student performance available through 1994 provide 277 separate estimates of the effect of class size or teacher-pupil ratios on student outcomes.[12] Studies are aggregated according to the estimated sign and statistical significance of the relationship.[13] The analysis begins with all of the combined evidence but subsequently focuses on just the best of the studies that consider variations in class size across individual classrooms.

Table 1 summarizes the available results for estimates of the effects of teacher-pupil ratios on student outcomes. The top row of the table shows that only 15% of all studies find a positive and statistically significant relationship between teacher intensity and student performance—the expected result if class size systematically matters. Even though conventional wisdom suggests that increasing the teacher-student ratio should have a positive effect on student performance, 13% of all studies show negative and statistically significant relationships with student performance. Ignoring the statistical significance, or the confidence that we have that there is any true relationship, we find that the estimates are almost equally divided between those suggesting that small classes are better and those suggesting that they are worse.[14] This distribution of results, symmetrically distributed about a zero effect, is what one would expect if there were no systematic relationship between class size and student performance. Fully 85% of the

TABLE 1
Percentage Distribution of Estimated Influence of Teacher-Pupil Ratio on Student Performance, by Level of Schooling

	Statistically significant			Statistically insignificant		
School level	No. of estimates	Positive (%)	Negative (%)	Positive (%)	Negative (%)	Unknown sign (%)
All schools	277	15	13	27	25	20
Elementary schools	136	13	20	25	20	23
Secondary schools	141	17	7	28	31	17

Note. A positive sign implies that smaller classes enhance student performance.

studies suggest either that fewer teachers per student are better (i.e., yield negative estimates) or that there is less confidence than usually required that there is any relationship at all (i.e., the effects are statistically insignificant).

Some people have suggested that the effect of class size may differ by point in the schooling process (including the interpretation of the STAR study discussed subsequently). To consider this possibility, the overall estimates of the effects of teacher-pupil ratios are divided into elementary and secondary schools. As Table 1 shows, there is little difference between the estimated effects in elementary and in secondary schools, but, if anything, there is less support for increasing teacher-pupil ratios at the elementary level. For elementary schools, more estimated effects (both for all studies and for ones with statistically significant estimates) are negative as opposed to positive (i.e., indicating that smaller classes are worse). There are, nonetheless, too few studies to permit looking at individual grades as opposed to all elementary grades combined.

With these data, it is also possible to address explicitly the distinction between pupil-teacher ratios and class size. As discussed earlier, while these two concepts differ, they are highly related. The overall estimates previously summarized contain a mixture of studies that explicitly measure class size and those that contain aggregate measures of teacher-pupil ratios for a school, district, or state. In fact, studies that investigate performance within individual classrooms invariably measure class size, while those at higher levels of aggregation most often measure average teacher-pupil ratios. In particular, studies that are highly aggregated, such as those investigating performance across entire districts or entire states, are almost always forced to consider just the overall teacher-pupil ratio.[15]

The issue of measuring actual class size, as opposed to only the teacher-pupil ratio, can be explicitly considered by focusing on just studies that analyze actual class size. Furthermore, by restricting attention to the best of the studies—those estimating value-added models for individual students—the effects of other potential problems with the estimation can be minimized.[16] Table 2 provides a summary of value-added results, both for all 78 separate estimates of class size effects and for the 23 estimates that come from samples in a single state. Clearly, the number of these estimates is very much reduced from the overall set that is available, and thus any conclusions are subject to more uncertainty simply as a result of the limited number of underlying investigations. The restriction to samples within single states corrects for differences in state school policies to avoid the biases previously discussed (cf. Hanushek, Rivkin, & Taylor, 1996). Because of the superiority of these analyses, each study deserves more weight than one of the general studies reviewed previously.

The more refined results in Table 2 provide little reason to believe that smaller classes systematically lead to improvements in student achievement. Of the best available studies (single-state, value-added studies of individual classroom achievement), only 1 of 23 (4%) shows smaller classes to have a statistically significant positive effect on student performance. More studies actually suggest that small classes are harmful.

The econometric evidence as a whole gives little support to the idea that smaller classes will lead to general improvements in performance. The available studies observe the effects of class size over a broad range (roughly 15 to 40 students per class) and, within that range, show little consistency of effects. There are of course a number of individual studies that suggest small classes are better (see Table 1), but there is no reason to put more weight

TABLE 2
Percentage Distribution of Effect of Class Size on Student Performance, Based on Value-Added Models of Individual Student Performance

	Statistically significant			Statistically insignificant		
Universe of studies	No. of estimates	Positive (%)	Negative (%)	Positive (%)	Negative (%)	Unknown sign (%)
All value-added studies	78	12	8	21	26	35
Value-added studies within a single state	23	4	13	30	39	13

Note. A positive sign implies that smaller classes enhance student performance.

on these than on the almost similar number of studies finding the opposite. In fact, when Table 2 restricts the results to those with the best analytical design, the support for gains from small classes actually falls.

Most of the econometric studies do not directly address the underlying mechanism for establishing small and large classes. If, for example, a school district used a subjective method of assigning "weaker" students to small classes and "stronger" students to large classes, the econometric methods might not provide an accurate assessment of the direct, causal influence of class size. This problem arises only when decisions are made on the basis of unmeasured student characteristics. If, for example, students are assigned to specific classes on the basis of their early test scores and if these test scores are controlled for in the econometric analysis as in the value-added estimation, such problems do not arise. The statistical analysis of Texas schools by Rivkin, Hanushek, and Kain (1998) also goes further by incorporating an econometric approach to deal with any remaining selection concerns. That analysis allows for individual student fixed effects in terms of achievement growth, virtually ruling out the hypothesized selection effects. Other studies have explicitly considered exogenous factors affecting class size within the context of instrumental variable estimators for the effects of class size: Akerhielm (1995) and Boozer and Rouse (1995) for a national sample of schools (National Education Longitudinal Study of 1988 [NEL88]), Hoxby (1998) for schooling in Connecticut, and Angrist and Lavy (1999) for schooling in Israel. These studies provide no clear conclusions about the impact of class size, and their generalizability to all U.S. schools is unclear.

It is possible that detection of small effects of class size is difficult in a number of these studies. With small sample sizes or correlations of small classes with a variety of other teacher, school, and family influences on learning, the statistical methods may not be powerful enough to identify reliably an effect, even when the effect exists (cf. Krueger, 1997). Some support for this hypothesis is found in Rivkin et al. (1998). When very large samples of Texas students (more than 500,000 repeated observations across 3,000 schools) are used, small positive effects of reduced classes can be detected for low-income students in earlier grades. But, as discussed later, small effects are small. While the focus of attention has turned to an issue of whether there are *any* effects, the policy discussion must also consider whether or not these effects are sufficiently large to justify the large expenditure needed to reduce overall class sizes.

In summary, the extensive statistical investigations of differences in teacher intensity and class size provide no consistent or clear indication that overall class size reductions will lead to improved student performance. The best studies that concentrate on differences in performance across individual classrooms with varying numbers of students and that separate out other possible influences on student performance offer no support whatsoever for general gains in achievement through class size policies.

Project STAR

The prior evidence comes from analyses of data generated naturally by the operations of schools. Inferences about the effect of altered class size rely on statistical adjustments for factors other than class size that might affect student performance. The primary alternative to these approaches is the use of random-assignment experimentation. Random-assignment experiments in principle have considerable appeal, and there is a powerful case for more extensive use of this approach. The underlying idea is that we can obtain valid evidence about the impact of a given well-defined treatment by randomly assigning subjects to treatment and control groups, eliminating the possible contaminating effects of other factors and permitting conceptually cleaner analysis of the outcomes of interest across these groups. Randomized trials have been employed extensively and productively in medical research. With observations derived from natural variations in individual selection, one must be able to distinguish between the treatment and other differences in patients and doctors that might directly affect the outcomes and that might be related to whether or not the treatment is received. The random assignment is employed to circumvent problems of otherwise having to measure and model all of the various factors that might affect outcomes in addition to the treatment. Randomization seeks to eliminate any relationship between selection into a treatment program and other factors that might affect outcomes.

The ultimate appeal (and validity) of any experimental results nonetheless depends crucially on design and implementation. There is no dispute that high-quality random-assignment experi-

ments offer the potential for dramatically expanding our knowledge of effective policies in schools, and an argument can be made that we continue to invest too little in such experiments (cf. Hanushek & Associates, 1994). But even medical experiments with well-designed protocols and well-defined treatment programs frequently require more than one set of clinical trials to ensure valid and reliable results. Social experiments, which tend to be much more complex, are very difficult to design and implement, making it even less likely that a single trial will provide definitive answers.

Much of the recent debate on class size policy has focused on the results from the Tennessee class size experiment of the mid-1980s. This experiment, mandated by the Tennessee legislature, has been used to justify the class size reduction programs in California and in a variety of states emulating California since 1996, as well as to motivate the federal debates. Some have suggested explicitly or implicitly that this one experiment, with its superior analytical design, supersedes all of the results from nonexperimental analyses reported earlier. Unfortunately, because the underlying data from the experiment have not been widely available, little of the discussion considers the details of the design and implementation of Project STAR.

Two central themes run through the discussion here. First, closer attention to the details of the STAR experiment points to considerable uncertainty about the results. Second, even ignoring that uncertainty, the evidence does not provide strong support for the policy proposals currently being discussed.

Issues of Design and Implementation

This discussion concentrates on a few key issues in the design and implementation of the STAR program. The design of the experiment and its history are described elsewhere (see, for example, Mosteller, 1995; Word et al., 1990) and are not discussed here except as relevant to the interpretation of the results.

The ideal experiment would randomly assign a large group of students and a large group of teachers to different class size treatments. These students would be followed over time, and their achievement would be recorded. Variations in the grade level of treatment, in the number of grades in which students were assigned to different class sizes, and in the amount of teacher training would provide additional information about key aspects of potential class size reduction programs. Project STAR had some but not all of these design features and included a series of implementation problems that introduce uncertainty about the interpretation of any of its results. The experiment was designed to begin with kindergarten students and to follow them for 4 years. Three treatments were initially included: small classes (13–17 students), regular classes (22–25 students), and regular classes (22–25 students) with a teacher's aide. Schools were solicited for participation, with the stipulation that any school participating must be large enough to have at least one class in each treatment group.

Table 3 displays some key elements of the samples that were constructed for the experiment. The initial sample included 6,324 kindergarten

TABLE 3
Project STAR Sample Sizes, by Treatment Group and Grade

	Grade level			
	K	1	2	3
Total students	6,324	6,829	6,840	6,802
In experiment in prior years	0	4,515	5,049	5,413
New to experiment	6,324	2,314	1,791	1,389
Students in small class treatment	1,900	1,925	2,016	2,174
In experiment in prior years	0	1,540	1,627	1,771
In regular class treatment previous year	0	248	192	207
New to experiment	1,900	385	389	403
Students in regular class treatment	4,424	4,904	4,824	4,628
In experiment in prior years	0	2,975	3,422	3,642
In small class treatment previous year	0	108	47	72
New to experiment	4,424	1,929	1,402	986
Number of schools	79	76	75	75
Number of teachers	326	339	339	335

students. These students were split between 1,900 in small classes and 4,424 in regular classes. (For most of this discussion, the two separate regular class treatment groups are aggregated together. After the first year, these treatments were effectively combined.) For subsequent grades of the experiment, slightly larger sample sizes—around 6,800—were maintained. The published sample sizes reported for the analysis are frequently smaller, reflecting the fact that 3%–12% of students in any year did not have valid test scores. The initial sample included 79 schools, although this subsequently fell to 75. The initial group of 326 teachers grew slightly to reflect the increased sample size in subsequent grades, although of course most teachers were new to the experiment at each new grade.

In each year of the experiment, there was sizable attrition from the prior year's treatment groups, and these students were replaced with new students. In first grade, 2,314 new students were added; in second grade, 1,791 new students were added; and, in third grade, 1,389 new students were added. Of the initial experimental group starting in kindergarten, 48% remained in the experiment for the entire four years.

Randomization. A key element for the entire research design is that students in experimental schools are randomly assigned across treatment groups. With large samples and random assignment, the difference in performance between students in two treatment groups is frequently presumed to reflect the causal impact of those treatments. It is, however, very difficult to verify the randomization in the STAR experiment. Even with a random-assignment design, it is useful to assess the assignment outcomes in terms of differences across treatment groups in measurable aspects. For this experiment, the largest concerns would arise from differences in entering achievement levels, but such verification is not possible in the Tennessee experiment because no pretest of achievement was given to the students. While this is explicable in the kindergarten sample, given the difficulty of testing at very young ages, it is less easy to understand in the subsequent years of the study. During the course of the study, 5,276 new students were added in the later grades, but none were given pretests at their enrollment in the experiment even though appropriate tests were available through the experiment itself.

Students entering the experiment in the first grade as opposed to kindergarten have noticeably lower performance at the end of the first grade. This finding is, however, complicated by the fact that kindergarten attendance was not compulsory in Tennessee at the time of the experiment, so many entering in the first grade probably did not have kindergarten.[17] Presuming that new samples are appropriately randomized across treatment groups, the addition of new experimental subjects still complicates the interpretation, largely because of unknown treatments in prior years. If, for example, class size reductions were to have a cumulative effect on student performance, it would be difficult to ascertain the full effect of reduced K–3 classes without knowledge of the new students' prior schooling experiences (unless prior achievement measures were also available so that value-added estimates could provide information about the specific experimental small classes).

Krueger (1997) considers whether the new experimental subjects in each grade were significantly different across treatment groups for measurable attributes (race, free-lunch status, age, and attrition rates). There are significant overall differences by race, but the statistical significance disappears if school effects are first removed. Significant differences in attrition rates exist in the earlier grades, even within schools (as described later).

The issue of randomness also has two other potentially important dimensions. First, the schools in the experiment are not random. As noted, they had to volunteer to participate, and they had to be large enough to accommodate at least three classes in each grade.[18] This sample selection has implications for the population to which any results can be generalized. Given the description of the schools that is available, it is not possible to provide any detailed analysis of the experimental schools. On simple grounds, however, the sample does differ from the student population in the state: 33% of the experimental students were Black, as compared with 23% for public school students in Tennessee in fall 1986.[19] Note also that the schools in the sample do change over the experiment, with four schools not remaining in the experiment for the full 4 years. The reasons for their withdrawal are not reported.

The sampling of schools is especially important if, as has been suggested, class size has a differential effect on students. Specifically, if low income or minority students are more sensitive to variations in class size, any overall estimates of the average effects of class size will depend upon the

sample weighting of the relevant subpopulations. This problem exists even when (appropriately) the analysis of treatment effects is conducted within individual schools (e.g., as in Krueger, 1997).

Second, and most important, the results depend fundamentally on the choice of teachers. While the teachers were to be randomly assigned to treatment groups, there is little description of how this was done. Nor is it easy to provide any reliable analysis of the teacher assignment, because only a few descriptors of teachers are found in the data and because there is little reason to believe that they adequately measure differences in teacher quality.[20] Yet, the huge differences generally found among teachers could dramatically influence the results, implying that the reality of teacher assignment is crucial. (A reasonable application of reduced class size policies would take into account differences among teachers in making assignments, but the policy application differs from the correct design of an experiment that is attempting to uncover the average effects of an across-the-board class size reduction.) Because of the ongoing need to assign new teachers to the various treatment groups across the 4 years, it seems entirely plausible that elements of teacher and principal preferences for different classes entered.[21]

Krueger (1997) checks the randomness by assessing whether the race, average experience, or degree level of teachers differs by treatment group, and he cannot reject similarity at the .05 level. He interprets this as implying that teachers were randomly assigned. These characteristics are, however, not very correlated with teacher quality, if at all (Hanushek, 1997), and variations in overall teacher quality have much larger effects than any measurable teacher characteristics (Rivkin, Hanushek, & Kain, 1998). If these measured characteristics are orthogonal to teacher quality, directly assigning teachers to treatment groups on the basis of quality would yield no correlation between treatment group and measured characteristics. Therefore, the finding that treatment group and these measured teacher characteristics are uncorrelated is compatible with both random teacher assignment and systematic assignment based on (unmeasured) teacher quality, thus yielding little information.

It is difficult to learn much about the distribution of teachers from just the data within this experiment. At the same time, it would be valuable to compare the teachers in this experiment with the value-added estimates for individual teachers that have been independently constructed for Tennessee (Sanders & Horn, 1995).[22]

Issues of random assignment introduce uncertainties into the results that cannot be sufficiently resolved with the available data. The direction of any bias that might result, however, cannot be readily ascertained without more detailed data.

Possible sources of nonrandomness in implementation. The STAR experiment, as pointed out by Mosteller (1995), is an extraordinarily important event in educational research history. Without doubt, more experimentation would also be valuable (cf. Hanushek & Associates, 1994). None of the discussions should minimize the innovative nature of the STAR experiment. At the same time, it must be recognized that conducting such experiments is very difficult. The end result often differs significantly from the ideal, and this can clearly affect the reliability and interpretation of findings from an experiment. Such is the case with the STAR experiment.

The clearest story of the sampling is the large attrition of students at each grade in the experiment. As noted, slightly less than half of the original students in the experiment in kindergarten remained in the experiment until the end of the third grade. As Table 3 indicates, the 1-year attrition rates are between 20% and 30% of the prior grade samples. Some attrition is clearly to be expected in any social experiment, but these rates of attrition appear very high. Moreover, as found by Goldstein and Blatchford (1998) and Krueger (1997), the attrition is not random. For example, Goldstein and Blatchford show that those dropping out of the experiment in the first grade had kindergarten achievement noticeably below average, and the differential below average was larger for those who started in regular classroom treatment groups in kindergarten (-0.35 standard deviations in math and -0.33 standard deviations in reading) than for those in the small kindergarten classroom treatment group (-0.25 standard deviations in math and -0.17 standard deviations in reading).[23] My calculations of performance gaps indicate that the difference between those leaving and those staying is even larger for the last year of the experiment. The calculation of simple treatment effects throughout the experiment can be adjusted for some of the observed attrition, but such adjustments necessarily rest on a number of strong but untestable assumptions.

Another potentially serious introduction of

TABLE 4
Test Taking in Project STAR, by Treatment Group and Grade

	Grade level							
	K		1		2		3	
	n	%	*n*	%	*n*	%	*n*	%
Total students	6,324		6,829		6,840		6,802	
No score on reading test	536	8.5	434	6.4	763	11.2	802	11.8
No score on math test	454	7.2	231	3.4	775	11.3	725	10.7
Students in small class treatment	1,900		1,925		2,016		2,174	
No score on reading test	161	8.5	102	5.3	222	11.0	259	11.9
No score on math test	138	7.3	58	3.0	227	11.2	236	10.9
Students in regular class treatment	4,424		4,904		4,824		4,628	
No score on reading test	375	8.5	332	6.8	541	11.2	543	11.7
No score on math test	316	7.1	173	3.5	548	11.4	489	10.6

nonrandomness arises from significant movement between treatment groups through the course of the experiment: unplanned, nonrandom "treatment crossover." As seen in Table 3, the gross flows were substantially larger for movement from regular to small classes than for the reverse—suggesting that principals responded to pressure from parents to get their children into the small classes. For Grades 1–3, between 9% and 12% of the students in the small classes had been in regular classes during the prior year; only 1%–2% of the students in regular classes had been in small classes during the prior year.[24] If the treatment crossover were random, any estimates of the cumulative effects of class size would tend to be lessened, because the treatment and control groups are not receiving completely distinct schooling programs over time. The experimental analysis (Word et al., 1990) indicates nonetheless that the crossover decisions were not random. For analytical purposes, these students who switch treatment groups in the middle of the experiment can be retained in the group to which they were initially assigned (cf. Krueger, 1997), but again the experiment as implemented is getting farther from the ideal.

Table 4 displays information about rates of test taking within the experiment. Large numbers of students in the experiment did not take the tests in each year (3%–12% across the test years). While these losses of test information might appear reasonable with absences, moves, and the like, the available information suggests some treatment bias in the lack of test taking. The rates of test taking are very similar across treatment groups except for Grade 1, where test taking in regular classes is noticeably below that in small classes. For students who were in the experiment for both kindergarten and first grade, it is possible to compare kindergarten scores for those who did and did not take the first-grade test. For students in both small and regular classes, kindergarten achievement of those taking the first-grade test exceeds that of those not taking it. Importantly, in math this differential score by test-taking status is noticeably larger for students in small kindergartens than for those in regular kindergartens, biasing the estimated treatment effects upward.

Teacher expectations and reactions to the experiment itself could also enter. Unlike medical experiments, the assignment to treatment groups is not a blind process. Instead, everybody in the school (and probably in the homes) knows that the experiment is happening, and many are likely to have prior views about the efficacy of smaller classes. The results of the experiment could also have been reasonably expected to have serious resource implications, given that it was mandated by the state legislature to provide evidence for proposed policy initiatives. Teachers and principals could react to this, in part by simply playing out their expectations that students in the small classes should perform better. This concern has an element of Hawthorne effect in it, but it also includes more direct motivation and incentives of teachers and principals that could bias the results of the different treatment groups. The significant reassignment of students across treatment groups with the predominant flow from regular to small classes clearly indicates that school personnel reacted to participant desires in this nonblind experiment.

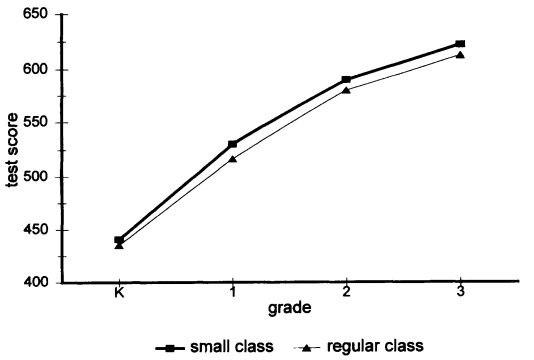

FIGURE 1. *Stanford Achievement Test—Reading.*

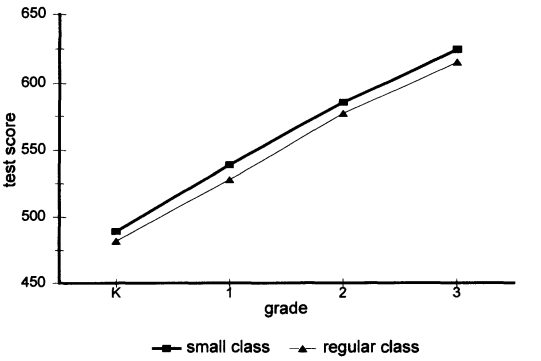

FIGURE 2. *Stanford Achievement Test—Math.*

In reviewing the implementation, it is clear that there are serious compromises to the desired protocols for the experiment. Yet, it is difficult to obtain any precise estimates of the direction of many of the biases, let alone of the magnitude of the various problems. The important issue is whether any of the nonrandomness is differentially associated with treatment groups, so that simple estimates of treatment effects would be biased. One clear effect of these various factors is simply to elevate the uncertainty surrounding any estimated effects, although subsequent analysis also suggests that there is an overall bias toward finding larger effects of small classes.

Summary of STAR Results

The results of the STAR experiment have been widely publicized. The simplest summary is found in Figures 1 and 2. These plots provide the average published performance in reading and math of students in the small class and regular class treatment groups for the four grade levels of the experiment. The regular classes with and without an aide are combined, because these two groups were virtually indistinguishable at the end of kindergarten and students were quite freely transferred across these two treatments in later grades (Word et al., 1990). Both figures yield similar conclusions.

1. Students in small classes perform better than those in regular classes or regular classes with aides starting in kindergarten.

2. The kindergarten performance advantage of small classes widens some in first grade but then either remains quantitatively the same (reading) or narrows (math) by third grade.

3. Taking each grade separately, the difference in performance between small and regular classes is statistically significant.

These figures reflect the typical reporting of the results, which focuses on the differences in performance at each grade and concludes that small classes are better than large (e.g., Finn & Achilles, 1990; Mosteller, 1995). These pictures also form the key evidence that has been employed to justify state and federal incentive programs to reduce class size.

But what is wrong with these pictures? First, they ignore the questions of nonrandomness raised earlier, since each grade comparison relies on the tested students enrolled in the different treatment groups each year. Second, the common interpretation of these ignores the fact that one would expect the differences in performance to become wider through the grades because they continue to get more resources (smaller classes), and that should keep adding an advantage.[25] While there are different perspectives that can reconcile this finding, the implication remains that a commensurate policy would reduce just class size in kindergarten and, possibly, first grade.[26]

Before discussing the implications of this analysis, however, it is useful to recast the analysis in terms of the nonrandomness issues brought out earlier. Each of the issues of sample selection—attrition, treatment reassignment, test taking, and the like—is difficult to analyze completely. Nonetheless, one simple and powerful comparison is available. The overall year-by-year results can be compared with the results that come from restricting the sample to those students who remained in the sample for the entire 4 years. (As discussed, this sample is roughly half of the total kindergarten sample and somewhat more than one quarter of all students observed during the full experiment.)

Consider what the comparison of estimated treatment effects should look like between the annual samples employed earlier and the 4-year sample *if* any nonrandomness from attrition, retention in grade, and sample additions is unrelated to treatment group. In such a case where the implementation problems are innocuous vis-à-vis the analysis of treatment effects, the average differences across grades should be identical if class size reduction has a one-time effect. If there is a cumulative effect, the differences for the 4-year sample should be increasingly larger than for the year-to-year sample, because students in the 4-year sample have larger average treatments.

Table 5 provides comparisons for the annual samples and the 4-year sample of students. For these comparisons, the reading and math scores have been converted into z scores, so that the differences will indicate relative position in the distribution of overall performance in each grade. The annual samples, which correspond to the data in Figures 1 and 2, show that the difference between performance in class sizes of 22–25 and 13–17 is 0.17 standard deviations in both math and reading. Students in small classes perform 0.12 standard deviations above the overall kindergarten mean, and those in regular classes perform 0.05 standard deviations below the mean. Attrition from the experiment is, however, concentrated in low-performing students; ignoring treatment group, the average student who remains in the experiment for the full 4 years is 0.24 (0.26) standard deviations above the kindergarten mean in reading (math). The interesting issue, however, is the differential impact on treatment groups: The kindergarten differential for the 4-year sample is slightly larger than that for the annual sample in both reading (0.18) and math (0.19).

Figures 3 and 4 plot the estimated advantage of small classes for the annual samples and for the 4-year sample. The estimates show distinctly contrasting patterns of achievement differences. The annual samples show both larger differences (presumably reflecting sample selection, differential attrition, or differential test taking) and different patterns over time. Over time, the differential effect of small versus regular classes appears to be significantly less, especially at the end of third grade, for those in the experiment all 4 years relative to the annual samples. Ceteris paribus, the differential effect should be larger if there is a cumulative effect of reduced class size and no treatment bias from the factors mentioned earlier.

Ignoring any possible initial biases in sampling for the experiment, the results are consistent with a one-time effect of smaller classes that either erodes or can be made up for over time in regular classes. One ambiguity exists, nonetheless. It could be that the gains in performance obtained in kindergarten would be expected to erode over time if further small-class treatment is not maintained.[27] The experimental design did not call for investigation of this possibility, which would be covered by randomly placing some of the students in small kindergartens back into regular classrooms at varying later grades. (Note that there was some movement across treatment cells, but this was not random, and achievement differences generated by this movement could well reflect characteristics other than class size.) One insight is available from the follow-on study to the Tennessee experiment. Students in the STAR experiment during the third grade were tracked in later grades—when the experiment ended and only regular educational settings were available. According to the sixth-grade report from the Lasting Benefits Study (Nye, Zaharias, Fulton, & Achilles, 1993), students previously in small classrooms in STAR outperformed the students previously in regular classrooms by 0.21 standard deviations in reading and 0.16 standard deviations in math. These differences are very close to those for the third-grade annual sample reported in Table 5 (0.22 standard deviations and 0.18 standard deviations, respectively). Nonetheless, any results from the reports of the Lasting Benefits Study should be taken as highly tenuous, because the investigators will not release their data and, as seen with STAR, the sample definitions and analytical decisions have large impacts.[28]

The important point is that the differential performance across treatment groups is unaffected by whether students are assigned to small classes (Grades 1–3) or not (Grade 4 and later). Finn and Achilles (this issue) transform test scores into estimated grade equivalents, which they identify as widening in later grades (i.e., the spread of grade equivalents for any number of standard deviations of test performance is larger as the grade level increases). While this metric would make the lines in Figures 1 and 2 fan out as grade level increases, it does not remove the reality that the differences in performance between small and large classes appear unaffected by whether or not added resources are applied.

TABLE 5

Test Performance (z-Scores) by Treatment Group, Grade, and Time in Experiment

	Grade level							
	K		1		2		3	
	Annual sample[a]	Four-year sample[b]	Annual sample[a]	Four-year sample[b]	Annual sample[a]	Four-year sample[b]	Annual sample[a]	Four-year sample[b]
Reading								
Small classes	0.12	0.38	0.17	0.41	0.14	0.34	0.15	0.32
Regular classes	–0.05	0.19	–0.07	0.26	–0.06	0.23	–0.07	0.18
Difference[c]	0.17	0.18	0.23	0.15	0.20	0.11	0.22	0.14
Math								
Small classes	0.12	0.40	0.19	0.44	0.13	0.32	0.12	0.29
Regular classes	–0.05	0.20	–0.08	0.22	–0.06	0.20	–0.06	0.19
Difference[c]	0.17	0.19	0.26	0.22	0.19	0.13	0.18	0.10

[a] Includes all students tested in given grade.

[b] Includes only students who remained in the experiment through all 4 years.

[c] Adjusted for rounding of treatment means.

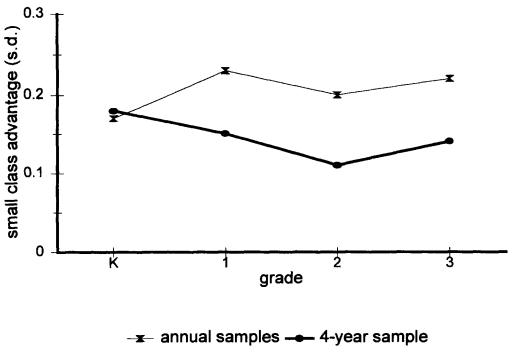

FIGURE 3. *Reading scores by sample.*

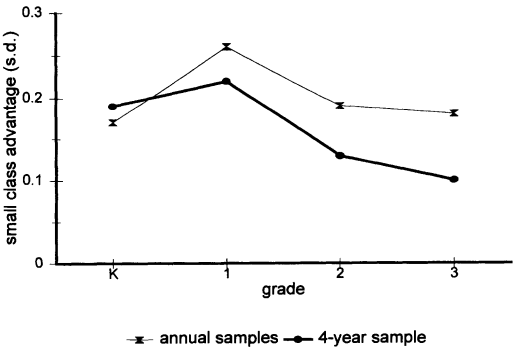

FIGURE 4. *Math scores by sample.*

Variation Across Schools

The STAR experiment was not a random selection of schools. The only randomization was across students within selected schools. Therefore, it is useful to ask, "How frequently is the small class achievement above that in the other treatments?"[29] If we look at kindergarten performance, where the composition of students in the treatment groups was clearest, we can compare performance directly. There were 79 schools with kindergarten experiments, and each had at least one small classroom, one regular classroom, and one regular classroom with a teacher's aide. Comparing reading scores, we see that the small classrooms outperformed both the regular and regular with aide classrooms in *40* schools.[30] While this is above the number that would be expected if performance were completely independent of class size, it also demonstrates that other things are very important in determining achievement.

The variations in performance within schools highlight one of the most important issues. A variety of previous studies have demonstrated the large performance differences across teachers. For example, Rivkin et al. (1998) find that, even in instances in which small effects can be obtained by reducing class size, these typical effects will be completely overshadowed by differences in teacher quality. At this point, differential costs have not been taken into account, but only if one presumes that nothing can be done about teacher quality would it make sense to ignore this differential impact.

Conclusions About Research Design

The STAR experiment has received justified recognition for its importance in educational research. The use of random-assignment experimentation deserves more attention in educational research. It is an especially appropriate technique for the analysis of well-defined treatments such as reduced class size.

While random-assignment experiments such as STAR appear expensive relative to other research approaches, such comparisons are frequently incorrectly made. First, because of the potential power of random-assignment approaches, they offer the possibility of much more reliable information than other research designs about the effects of alternative policy. Thus, the correct standard for judging different research approaches should be one that standardizes for the quality of research outcomes. Second, the appropriate comparison is often not alternative research approaches but a full-blown program. Here is where the comparison is easiest. The STAR experiment cost $12 million, which amounts to $16 million in 1996 prices. Compare this with running a statewide class size reduction program in California, which is currently costing in excess of $1 billion per year for K–3 reductions and, as the evidence tends to suggest, likely to be generally ineffective.

The advantages of random-assignment methodology do not, however, imply that other evidence has no value, particularly when there are significant uncertainties in the experimental evidence. Indeed, as discussed earlier, extensive investigations of the impact of reduced class sizes conducted over a long period of time have yielded strong and consistent evidence. The STAR experiment, while generally not contradicting the previous evidence, had a series of implementation problems that introduce ambiguity and uncertainty. It is difficult to assess with any precision the impact of the various sampling and selection issues that arise in STAR. The evidence indicates that a number but

not necessarily all of the resulting biases go in the direction of inflating the calculated impact of class size reductions. Thus, the evidence from the one currently available experiment is hardly sufficient to overturn that from different, nonexperimental approaches.

Several issues arise that would be important in any future experimentation with class size policies. First, simply replaying the problems with the original Tennessee design, it would be valuable to consider different magnitudes of class size reductions, to develop alternative treatments covering the grade level of introduction of small classes and the pattern by which individual students are in small and regular classes, and to provide further information about differences among the treatment groups of students and teachers. Second, from the implementation phase, it is clear that complicated experiments such as STAR require more attention to sampling, selection, and attrition issues. An important element of the work is maintaining the experimental design throughout the study and not permitting high levels of student or school interventions to counteract the randomization.

Without doubt, the decision to employ a random-assignment experimental design to study class size reductions was extremely innovative and important. The fact that it is hard to do and hard to derive definitive results, particularly in the initial large-scale attempt, should not detract from its importance as an underused research technique. Conversely, its importance as a research methodology should not lead to the conclusion that any results from a study are definitive.

Conclusions About Policy

This article has concentrated on the limited task of reviewing the evidence on whether or not there is any systematic impact of reducing class sizes. The surprising finding is that the evidence does not offer much reason to expect a systematic effect from overall class size reduction policies. This conclusion comes from both nonexperimental and experimental evidence.

The nonexperimental evidence has been generally understood as not supporting overall policies of class size reduction. The broad array of approaches, with different methodologies and sources of evidence, has provided a quite consistent message that broad reductions in class size are unlikely to produce significant improvements in student achievement. The aggregate evidence sketched here indicates that beneficial effects cannot be seen from the large increases in teacher intensity that have occurred over the past three decades. While some of the changes in overall pupil-teacher ratios undoubtedly went into programmatic additions that did not reduce the number of children in the typical regular classroom, there is little doubt that there were also broad reductions in class sizes. From a different vantage point, the enormous differences in teacher intensity observed internationally appear to have nothing to do with international differences in math and science performance. Finally, from yet another analytical perspective, almost 300 econometric investigations of the determinants of student achievement have failed to provide any consistent evidence that higher teacher-pupil ratios have a positive effect. When disaggregated to the smaller set of high-quality studies with detailed measures of class size within individual classrooms, there is even less support for general class size reduction policies.

In contrast, the experimental evidence from Tennessee has been generally understood to strongly support existing and proposed class size reduction policies. Therefore, it is useful to review exactly what can be inferred from the STAR experiment. First, the evidence applies to a specific set of larger elementary schools, and it is not known whether there are larger populations to which it can be generalized.

Second, the evidence refers just to a very large class size reduction that moves classes down considerably below the levels mentioned in California or in recent policy proposals. While it might appear reasonable simply to interpolate the results for less aggressive reduction programs, the early motivation for the STAR experiment was the conclusion of Glass and Smith (1979) that there was little achievement effect until class sizes got down to around 15 students. That analysis explicitly identified a nonlinear relationship—one that would suggest that policies of reducing class sizes to around 20, as has been done in California, would be expected to have little or no impact.

Third, the positive impacts of class size reduction appear limited to kindergarten and, possibly, first grade. Specifically, the annual samples of STAR data indicate no further impact on achievement of class size reductions after the first grade. The 4-year sample, which gets around a number (but not all) of the selection problems, isolates just a kindergarten effect.[31]

Fourth, the evidence casts considerable doubt on the efficacy of teacher aides for permitting the classroom teacher to provide more individualized attention and thus for raising achievement. (This evidence may also add to questions about the efficacy of general class size reduction policies, because the addition of an adult aide would seem to represent one kind of class size reduction.)

Fifth, the huge variation in effectiveness of small class instruction, even in kindergarten, appears to reflect underlying variation in teacher quality that far exceeds any average effects of reduced class size. It is only slightly better than an even bet from the STAR data that the small class achievement will exceed that of the regular and the regular with aide classes in any of the sampled schools.

The evidence does not say that small classes never matter. Nor does it say that small classes can never be used to elevate achievement. To the contrary, one way of reading the econometric evidence is that sometimes small classes are useful and other times they are not. This also is a potential finding of STAR, although it is difficult to identify any specific circumstances in which small classes are particularly effective. If there truly is a range of effects, one of the real challenges of school management is figuring which students, teachers, or subject matters may be most affected by reduced class sizes and which would not be affected by increased class sizes. One important example is that disadvantaged students may be more sensitive to variations in class size than are advantaged students. This result is clear in Rivkin, Hanushek, and Kain (1998) where class size variation has no significant impact on students ineligible for free or reduced-price lunch but has some impact (although again small) on eligible students. Grissmer et al. (1998) similarly suggest that Black students are much more sensitive to reduced pupil-teacher ratios than are White students. These examples indicate that, if implemented, policy applications must focus on strategic use of reduced class sizes.

Currently, however, the policy debate and the prevailing management tendency are geared to reducing classes across the board—typically on a "fairness" argument as pertains to either students or teachers. Redirecting attention to performance seems to be an issue of getting the incentives in schools correct so that teachers and school personnel are rewarded for improved student achievement (Hanushek & Associates, 1994).

The emphasis on whether or not there are any significant positive effects from class size reductions is also misleading from a policy viewpoint. Class size reduction represents one of the most costly reform policies actively discussed. Even if there are positive effects, they must be sufficiently large to justify the expenditure. Because of the very small (if any) effects of general class size reduction policies that have been found, a thorough analysis of costs and effectiveness relative to other policies does not seem to be required here. It is important, however, to remember costs when, as the popular discussion sometimes introduces, the argument is made that "even if we are unsure about the size of any effects, we should proceed, because surely reductions in class size could not hurt."

It also appears that the ultimate effect of any large-scale program to reduce class size will depend much more importantly on the quality of new teachers hired than on the effects of class size reductions per se. Variations in teacher quality have been shown to be extraordinarily important for student achievement, and the econometric studies providing such results indicate that these variations completely dominate any effects of altered class size. Rivkin, Hanushek, and Kain (1998) demonstrate that class size variation can explain just a very small portion of the variation in student achievement and that variations in teacher quality are much more significant. Hanushek (1992), for example, estimates variations in total teacher differences (measured and unmeasured) and shows that the differences in student achievement with a good versus a bad teacher can be more than a whole grade level of achievement within a single school year. Thus, if new hires resulting from a class size reduction policy are above the average quality of existing teachers, average student performance is likely to increase. If below, average student performance is likely to fall with class size reductions. From past experience, there is little reason to believe that the quality of new teachers will be significantly different from that of existing teachers unless incentives facing schools also change.[32] But consideration of possible hiring outcomes does speak to the assertion that "surely reductions in class size could not hurt."

Notes

Charles Achilles, Alan Krueger, and two referees provided helpful comments on an earlier version. The data for the analysis were kindly provided by Helen Pate Bain,

chair of HEROS, Inc. She of course bears no responsibility for the analysis here.

[1]STAR is an acronym for Student/Teacher Achievement Ratio. The experiment, as described in Word et al. (1990), covered kindergarten through third-grade classes and was conducted during 1985–1989.

[2]Detailed discussion of the evidence, along with a more complete bibliography of studies, can be found in Hanushek (in press).

[3]A longer time series can be constructed from the Scholastic Aptitude Test (SAT), although using those data introduces added interpretive issues. Average SAT scores fell dramatically from the mid-1960s until the end of the 1970s, suggesting that the achievement picture in the NAEP data neglects an earlier period of achievement falloff and thus starts at a lower level than historically achieved. The voluntary nature of the SAT and the increase in the proportion of high school seniors taking the test do introduce uncertainties about the precise magnitude of any change. The SAT is taken by a selective group of students who wish to enter competitive colleges and universities. As the proportion taking the test rises, so the hypothesis goes, an increasingly lower achieving group will be drawn into the test, leading to lower scores purely because of changes in test taking. While the exact magnitude of any such effects is uncertain, it seems clear that this change in selectivity has caused some of the SAT decline but not all of it (e.g., see Congressional Budget Office, 1986; Wirtz, 1977).

[4]Class sizes also differ dramatically within districts and across states and districts at any point in time. In the subsequent consideration of statistical analyses of student performance, attention is given to these measurement issues.

[5]The class size trends come from a National Education Association survey of teachers that has been conducted every 5 years since 1956. There is no information about the sampling design, validity of the responses, or the range of classroom situations included in this survey.

[6]Title I spending for compensatory education, which began in the 1960s, would also affect class sizes and teacher utilization for both compensatory and regular education. Changes in class size from this source, however, do not cause the same potential problems as any attributable to special education. Title I students are regularly tested, while a number of special education students are not (see Note 7). Therefore, if smaller classes aid disadvantaged students (indeed, some suggest an even greater impact for disadvantaged than for nondisadvantaged students), any reductions in class size should show up directly in average student performance.

[7]While little evidence is available, it is frequently asserted that special education students do not get included in tests and other measures of performance. Therefore, in assessing performance, it would be appropriate to link expenditure on regular-instruction students with their test performance. On the performance side, however, if a larger proportion of students are identified as special education students and if these are generally students who would perform poorly on tests, the shift to increased special education over time should lead to general increases in test scores ceteris paribus. (States also vary in their inclusion of special education students in state testing programs; see Hanushek, Kain, and Rivkin [1998].)

[8]This act, PL 94-142, is commonly identified as having direct and significant effects on the cost and methods of delivery of local education. See discussion and evaluation in Singer and Butler (1987) and Monk (1990).

[9]These estimates are themselves subject to criticism. They do not observe or measure differences in schools but instead simply attribute unexplained residual differences in the predicted and observed trends to school factors. The statistical complications of this estimation are likely to yield biased regression estimates, which in turn would provide incorrect weights for the trends in family backgrounds. Also, one must believe either that the factors identified are the true causal influences (cf. Mayer, 1997) or that they maintain a constant relationship with the true causal influences.

[10]Test scores are reported in Beaton et al. (1996a, 1996b). Primary pupil-teacher ratios for public and private schools are found in Organization for Economic Co-operation and Development (1996).

[11]The summary presented here describes all of the separate estimates of the effects of resources on student performance that could be found. For tabulation purposes, a "study" is a separate estimate of the class size effect in a published analysis of an educational production function. The overall sample of studies and description of criteria for inclusion can be found in Hanushek (1997). The entire collection of production function estimates includes 90 individual publications with 377 separate estimates of some resource parameter, from which the studies of teacher-pupil ratios are extracted. While a large number of studies were produced as a more or less immediate reaction to the "Coleman Report," half of the available studies have been published since 1985.

[12]Estimates of the effect of class size or pupil-teacher ratios are reversed in sign to yield the effects of teacher-pupil ratios, so that conventional wisdom would call for a positive effect in the reported estimates. The distinction between teacher-pupil ratios, pupil-teacher ratios, and class size is taken up later.

[13]More details about the methodology and the available studies can be found in Hanushek (1979, 1997). Some controversy also exists about the best way to summarize the results of different studies, but these issues have little bearing on the discussions here; see Greenwald, Hedges, and Laine (1996) and Hanushek (1996a, 1997). Other discussions and controversies about the estimation strategies can be found in Card and Krueger (1996); Heckman, Layne-Farrar, and Todd (1996); and Hanushek (1996b). The issues raised in those latter discussions, while relevant to some of the considerations in this article, are very

technical and, in my opinion, do not affect the policy conclusions here.

[14]Twenty percent of the studies do not report the sign of any estimated relationship. Instead, they simply note that the estimates were statistically insignificant.

[15]As described in Hanushek, Rivkin, and Taylor (1996), the more aggregated analyses are subject to a series of specification problems (independent of the measurement issue considered here) that are exacerbated by the aggregation of the analysis. In particular, the more aggregated analyses leave out consideration of state-by-state differences in school policies, and this omission appears to bias the results toward finding stronger effects of teacher-pupil ratios and school resources in general.

[16]One type of statistical investigation—that employing a value-added specification—is generally regarded as being conceptually superior and likely to provide the most reliable estimates of education production functions. These studies relate a student's current performance to the student's performance at some prior time and to the school and family inputs during this intervening time. The superiority of this approach comes from the use of prior achievement to ameliorate any problems arising from missing data about past school and family factors and from differences in innate abilities of students (Hanushek, 1979).

[17]Note that for students within the regular class treatment group, the difference in average first-grade performance by grade of entry into the experiment exceeds the difference in average first-grade performance between small and regular classes, perhaps indicating the importance of kindergarten.

[18]The initial applicant pool included 180 schools in 50 separate districts (out of 141). The final pool of schools, spread by geographic type (inner city, urban, suburban, and rural), was drawn from 42 districts (Word et al., 1990). Schools were compensated for any extra teachers or aides that needed to be hired but were responsible for all remaining costs.

[19]From 1980 through 1995, Black elementary students remained less than 24%, so the disparity cannot arise from changing demographic patterns in Tennessee.

[20]The teacher data include race, gender, teaching experience, highest degree, and position on the Tennessee career ladder. While there is no information about the effect of career ladder position on student performance, none of the other measures have been found to be reliable indicators of quality (Hanushek, 1997).

[21]See later discussion of student reassignment to small classes, which provides some prima facie evidence of actions that override the experimental design.

[22]The development of the full Tennessee Value-Added Assessment System (TVAAS) came after the end of the STAR experiment. Nonetheless, data on teachers remaining in Tennessee schools after STAR could be analyzed within TVAAS. In addition, some of the early model development and early analysis of test scores (Sanders & Horn, 1995) may also overlap with the experiment itself.

[23]Some of the attrition may be due to retention in grade. The impact on the experimental results of such explicit selection out of the experiment would depend on whether it was applied differentially across treatment groups. Large retention in grade for regular class size students, for example, would tend to bias the treatment results against small class sizes.

[24]Note that, since there are more students in regular classrooms, the probabilities of a random student leaving a treatment group are more similar than the proportions in the treatment crossover groups.

[25]A similar conclusion is reached by Prais (1996), who frames the discussion in terms of the value added in each grade but relies on just the published aggregate data.

[26]A discussion of alternative underlying learning models is found in Hanushek (in press). The most consistent model suggests that there is a one-time gain from "learning how to do the business of school." This motivates, for example, the Krueger (1997) parameterization that estimates a first-year effect and subsequent gains and is consistent with simple value-added calculations of Prais (1996).

[27]Such an expectation might come from common interpretations of the falloff in gains observed in early evaluations of the Head Start preschool program. The evidence from preschool programs actually is quite varied in terms of estimated effects, duration of programs, and research methodology. See Barnett (1992) for a review and critique.

[28]The Lasting Benefits Study data, along with the Project STAR data, are controlled by Barbara Nye, director of the Center of Excellence for Research in Basic Skills at Tennessee State University. Despite repeated requests, the center has refused to release any of the data about the Tennessee class size experiment even though more than a decade has passed. The data for Project STAR used here were kindly provided by Helen Pate Bain, an original investigator no longer associated with the center.

[29]I thank Charles Achilles for suggesting this analysis to me.

[30]In 19 cases, the advantage of the small class average over the combined regular class average is statistically significant. In 14 cases, the small class advantage over both the regular and regular with aide classes is statistically significant.

[31]These two conclusions may not be incompatible. The annual samples introduced large numbers of new students in the first grade who probably had not had any kindergarten. If the effects are "first-year" effects, this group would have a clear impact on the annual sample data but not the 4-year sample. See also Krueger (1997).

[32]Under some circumstances, such as the large unexpected hiring from the California class size reductions in 1996, one might expect the average quality to fall. In general, however, there is no shortage of trained teachers, and the real issue is simply the selection from the substantial pool of trained teachers not currently employed in the

schools. See Ballou and Podgursky (1997) and Murnane, Singer, Willett, Kemple, and Olsen (1991).

References

Akerhielm, K. (1995). Does class size matter? *Economics of Education Review, 14,* 229–241.

Angrist, J. D., & Lavy, V. (1999). Using Maimondides' rule to estimate the effect of class size on scholastic achievement. *Quarterly Journal of Economics, 114*(2), 533–575.

Ballou, D., & Podgursky, M. (1997). *Teacher pay and teacher quality.* Kalamazoo, MI: W. E. Upjohn Institute for Employment Research.

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

Beaton, A. E., Martin, M. O., Mullis, I. V. S., Gonzalez, E. J., Smith, T. A., & Kelly, D. L. (1996a). *Science achievement in the middle school years: IEA's Third International Mathematics and Science Study (TIMSS).* Boston: Center for the Study of Testing, Evaluation, and Educational Policy, Boston College.

Beaton, A. E., Mullis, I. V. S., Martin, M. O., Gonzalez, E. J., Kelly, D. L., & Smith, T. A. (1996b). *Mathematics achievement in the middle school years: IEA's Third International Mathematics and Science Study (TIMSS).* Boston: Center for the Study of Testing, Evaluation, and Educational Policy, Boston College.

Boozer, M., & Rouse, C. (1995). *Intraschool variation in class size: Patterns and implications.* (Working Paper No. 5144). Cambridge, MA: National Bureau of Economic Research.

Card, D., & Krueger, A. B. (1996). School resources and student outcomes: An overview of the literature and new evidence from North and South Carolina. *Journal of Economic Perspectives, 10,* 31–50.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., & York, R. L. (1966). *Equality of educational opportunity.* Washington, DC: U.S. Government Printing Office.

Congressional Budget Office. (1986). *Trends in educational achievement.* Washington, DC: Author.

Cook, M. D., & Evans, W. N. (1996). *Families or schools? Explaining the convergence in White and Black academic performance.* Unpublished manuscript.

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

Goldstein, H. & Blatchford, P. (1998). Class size and educational achievement: A review of methodology with particular reference to study design. *British Educational Research Journal, 24,* 255–268.

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

Grissmer, D., Flanagan, A., & Williamson, S. (1998). Why did the Black-White score gap narrow in the 1970s and 1980s? In C. Jencks & M. Phillips (Eds.), *The Black-White test score gap* (pp. 181–226). Washington, DC: Brookings Institution Press.

Grissmer, D. W., Kirby, S. N., Berends, M., & Williamson, S. (1994). *Student achievement and the changing American family.* Santa Monica, CA: Rand Corporation.

Hanushek, E. A. (1979). Conceptual and empirical issues in the estimation of educational production functions. *Journal of Human Resources, 14,* 351–388.

Hanushek, E. A. (1992). The trade-off between child quantity and quality. *Journal of Political Economy, 100,* 84–117.

Hanushek, E. A. (1996a). Measuring investment in education. *Journal of Economic Perspectives, 10*(4), 9–30.

Hanushek, E. A. (1996b). School resources and student performance. In G. Burtless (Ed.), *Does money matter? The effect of school resources on student achievement and adult success* (pp. 43–73). Washington, DC: Brookings Institution.

Hanushek, E. A. (1997). Assessing the effects of school resources on student performance: An update. *Educational Evaluation and Policy Analysis, 19,* 141–164.

Hanushek, E. A. (in press). The evidence on class size. In S. E. Mayer & P. Peterson (Eds.), *Earning and learning: How schools matter.* Washington, DC: Brookings Institution.

Hanushek, E. A., & Associates. (1994). *Making schools work: Improving performance and controlling costs.* Washington, DC: Brookings Institution.

Hanushek, E. A., Kain, J. F., & Rivkin, S. G. (1998). *Does special education raise academic achievement for students with disabilities?* (Working Paper No. 6690). Cambridge, MA: National Bureau of Economic Research.

Hanushek, E. A., & Kim, D. (1996). *Schooling, labor force quality, and economic growth* (Working Paper No. 5399). Cambridge, MA: National Bureau of Economic Research.

Hanushek, E. A., & Rivkin, S. G. (1997). Understanding the twentieth-century growth in U.S. school spending. *Journal of Human Resources, 32,* 35–68.

Hanushek, E. A., Rivkin, S. G., & Taylor, L. L. (1996). Aggregation and the estimated effects of school resources. *Review of Economics and Statistics, 78,* 611–627.

Heckman, J. S., Layne-Farrar, A., Todd, P. (1996). Does measured school quality really matter? An examination of the earnings-quality relationship. In G. Burtless (Ed.), *Does money matter? The effect of school resources on student achievement and adult success* (pp. 192–289).

Washington, DC: Brookings Institution.

Hoxby, C. M. (1998). *The effects of class size and composition on student achievement: New evidence from natural population variation* (Working Paper No. 6869). Cambridge, MA: National Bureau of Economic Research.

Krueger, A. B. (1997). *Experimental estimates of education production functions* (Working Paper No. 6051). Cambridge, MA: National Bureau of Economic Research.

Lewit, E. M., & Baker, L. S. (1997). Class size. *The Future of Children, 7,* 112–121.

Mayer, S. E. (1997). *What money can't buy: Family income and children's life chances.* Cambridge, MA: Harvard University Press.

Monk, D. H. (1990). *Educational finance: An economic approach.* New York: McGraw-Hill.

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

Murnane, R. J., Singer, J. D., Willett, J. B., Kemple, J. J., & Olsen, R. J. (1991). *Who will teach?* Cambridge, MA: Harvard University Press.

Nye, B. A., Zaharias, J. B., Fulton, B. D., & Achilles, C. M. (1993). *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: Sixth grade technical report.* Nashville: Center of Excellence for Research in Basic Skills, Tennessee State University.

Organization for Economic Co-operation and Development. (1996). *Education at a glance: OECD indicators.* Paris: Author.

Prais, S. J. (1996). Class-size and learning: The Tennessee experiment—What follows? *Oxford Review of Education, 22,* 399–414.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (1998). Teachers, schools, and academic achievement. (Working Paper No. 6691). Cambridge, MA: National Bureau of Economic Research.

Sanders, W. L., & Horn, S. P. (1995). The Tennessee Value-Added Assessment System (TVAAS): Mixed model methodology in educational assessment. In A. J. Shinkfield & D. L. Stufflebeam (Eds.), *Teacher evaluation: Guide to effective practice* (pp. 337–376). Boston: Kluwer Academic.

Singer, J. D., & Butler, J. A. (1987). The Education for All Handicapped Children Act: Schools as agents of social reform. *Harvard Educational Review, 57,* 125–152.

Stevenson, H. W., & Stigler, J. W. (1992). *The learning gap: Why our schools are failing and what we can learn from Japanese and Chinese education.* New York: Summit Books.

U.S. Department of Education. (1997). *Digest of education statistics.* Washington, DC: National Center for Education Statistics.

Wirtz, W. (1977). *On further examination: Report of the advisory panel and the Scholastic Aptitude Test score decline.* New York: College Entrance Examination Board.

Word, E., Johnston, J., Bain, H. P., Fulton, B. D., Zaharias, J. B., Lintz, M. N., Achilles, C. M., Folger, J., & Breda, C. (1990). *Student/Teacher Achievement Ratio (STAR), Tennessee's K–3 class size study: Final summary report, 1985–1990.* Nashville: Tennessee State Department of Education.

Author

ERIC A. HANUSHEK is a professor of economics and public policy at Wallis Institute of Political Economy and Department of Economics, University of Rochester, Harkness Hall 108, Rochester, NY 14627-0158. He specializes in public finance and education policy.

Manuscript received January 15, 1999
Revisions received February 16, 1999
Accepted February 17, 1999