Institute for Research on Poverty
Discussion Paper no. 1242-02

How Close Is Close Enough?
Testing Nonexperimental Estimates of Impact against Experimental Estimates of Impact with Education Test Scores as Outcomes

Elizabeth Ty Wilde
Monitor Group
E-mail: Ty-Wilde@Monitor.com

Robinson Hollister
Department of Economics
Swarthmore College
E-mail: rhollis1@swarthmore.edu

January 2002

This work benefitted greatly from assistance from Diane Whitmore, Alan Krueger, and Tom Dee in dealing with the Project Star data. Roberto Agodini and Mark Dynarski provided help in developing the propensity-score-matching procedures and useful comments on earlier drafts.

Abstract

In this study we test the performance of some nonexperimental estimators of impacts applied to an educational intervention—reduction in class size—where achievement test scores were the outcome. We compare the nonexperimental estimates of the impacts to "true impact" estimates provided by a random-assignment design used to assess the effects of that intervention. Our primary focus in this study is on a nonexperimental estimator based on a complex procedure called propensity score matching.

Previous studies which tested nonexperimental estimators against experimental ones all had employment or welfare use as the outcome variable. We tried to determine whether the conclusions from those studies about the performance of nonexperimental estimators carried over into the education domain.

Project Star is the source of data for the experimental estimates and the source for drawing nonexperimental comparison groups used to make nonexperimental estimates. Project Star was an experiment in Tennessee involving 79 schools in which students in kindergarten through third grade were randomly assigned to small classes (the treatment group) or to regular-size classes (the control group). The outcome variables from the data set were the math and reading achievement test scores.

We carried out the propensity-score-matching estimating procedure separately for each of 11 schools' kindergartens and used it to derive nonexperimental estimates of the impact of smaller class size. We also developed proper standard errors for the propensity-score-matched estimators by using bootstrapping procedures. We found that in most cases, the propensity-score estimate of the impact differed substantially from the "true impact" estimated by the experiment. We then attempted to assess how close the nonexperimental estimates were to the experimental ones. We suggested several different ways of attempting to assess "closeness." Most of them led to the conclusion, in our view, that the nonexperimental estimates were not very "close" and therefore were not reliable guides as to what the "true impact" was.

We put greatest emphasis on looking at the question of “how close is close enough?” in terms of a decision-maker trying to use the evaluation to determine whether to invest in wider application of the intervention being assessed—in this case, reduction in class size. We illustrate this in terms of a rough cost-benefit framework for small class size as applied to Project Star. We find that in 30 to 45 percent of the 11 cases, the propensity-score-matching nonexperimental estimators would have led to the “wrong” decision.

How Close Is Close Enough?
Testing Nonexperimental Estimates of Impact against Experimental Estimates of Impact with Education Test Scores as Outcomes

INTRODUCTION

The Nature of This Study and Its Motivation

In this study we attempt to assess the effectiveness of using nonexperimental methods in estimating the impact on test scores of a particular educational intervention. Our focus is on the *methods* of impact assessment rather than on the question of whether the educational intervention was effective. Though the findings here have potential bearing on the controversies surrounding the impact of this particular type of educational intervention (reduced class size), we do not emphasize those points.

Two major considerations motivated us to undertake this study. First, four important studies (Fraker and Maynard, 1987; LaLonde, 1986; Friedlander and Robins, 1995; and Dehejia and Wahba, 1999) have assessed the effectiveness of nonexperimental methods of impact assessment in a compelling fashion, but these studies have focused solely on social interventions related to work and their impact on the outcome variables of earnings, employment rates, and welfare utilization. We attempt to ascertain to what degree the conclusions from those studies carry over to situations in which the estimates of impacts are on *other types of outcomes*. Second, in recent years there has been a growing interest in the effectiveness of one particular type of nonexperimental method used for assessing the impacts of interventions on outcomes, namely, *propensity score matching*. We attempt to assess how well this nonexperimental method performs with respect to the impact of an educational intervention (reduction in class size) on standardized achievement test scores as outcome variables

The Basic Framework

The framework we used derives from Fraker and Maynard (1987) and LaLonde (1986).

One starts with data drawn from a “true random-assignment” study of a particular intervention. A group of individuals is randomly assigned to the intervention—the “treatment group” (or “treatments”)—and a group of individuals is assigned *not* to take part in the intervention—the “control group”(or “controls”). Estimates of what we will call the “true impact” are available by estimating the difference in the outcome of interest between the treatments and the *controls*.

The control group is critical because it provides the *counterfactual*: what would have happened to the treatments—and how it would have affected their outcome measures—had they *not* been given the opportunity to experience the intervention.

The nonexperimental methods are assessed by constructing a “comparison group” (or “comparisons”), which is substituted for the control group. The nonexperimental estimates of the impact are obtained by estimating the difference in the outcome of interest between the treatments and the *comparisons*. The key issue is how best to construct the comparison group and how reliable a counterfactual it will provide.

The great virtue of this framework is that it provides a clear standard against which to judge the reliability of the nonexperimental method in assessing impacts for the given sort of intervention for the chosen outcomes; that clear standard is the “true impact” obtained from the random assignment.

Because we are interested in testing nonexperimental methods on educational outcomes, we use Tennessee’s Project Star as the source of the “true random-assignment” data. We describe Project Star in detail below. We use the *treatment group data from a given school* for the treatments and then construct comparison groups in various nonexperimental ways with *data taken out of the control groups in other schools*.

Previous Research in This Genre

The first attempts to use the basic approach just outlined were carried out by Fraker and Maynard (1987) and LaLonde (1986) using estimates based on the National Supported Work Demonstration, a random-assignment experiment, as the experimental "true impact" estimates.

The National Supported Work Demonstration was carried out in 13 sites across the nation from 1975 to 1979. The intervention was to provide paid employment opportunities for up to 18 months to members of four disadvantaged groups: ex-addicts, ex-criminal offenders, women on welfare (AFDC), and teenage school dropouts. The outcomes examined were postprogram employment, earnings, and utilization of welfare. The researchers used the data for the treatment group members (those randomly assigned to participate in the employment program) and then constructed various comparison groups to use in generating nonexperimental estimates of the impact of the employment opportunity on the postprogram employment, earnings, and welfare outcomes. To obtain the nonexperimental comparison groups the researchers used data drawn from outside the demonstration program (the Current Population Survey in one case and the Panel Study of Income Dynamics in another). They tested a variety of methods to draw comparison samples which in some sense "matched" those in the Supported Work experimental groups. They also tried various types of "selection bias" corrections. As a generalization one could say that the nonexperimental estimators did not come very close to the "true impact" estimates in most cases.

The next effort using the basic framework was carried out by Friedlander and Robins (1995). These researchers used data from several welfare-to-work demonstration programs carried out in various sites during the 1980s. The outcomes measured were again employment, earnings, and welfare usage.

Several innovations were made in this study. First, rather than creating nonexperimental comparison groups from external data sources, Friedlander and Robins used the data for the control group outcomes drawn from the demonstrations themselves. For example, they used the treatment group from a demonstration in Baltimore with the control group from a demonstration in San Diego as the comparison group. Second, because in most sites the sample was enrolled over an extended period of time, they were able to use treatments enrolled in the early period with a comparison group from among the controls enrolled later. Third, for one site there were even different “offices” within the city that enrolled sample members, so they were able to construct a comparison group from the controls from one office to be compared with a treatment group drawn from another office, both enrolled at about the same time. They also tested different matching procedures and “selection bias” corrections.

The conclusions were quite similar to those of the previous study based on the Supported Work data. In a very high percentage of the cases the nonexperimental estimates of impacts did not come very close to the “true impact” estimates obtained from the direct experimental-control samples data drawn from the same site by taking experimental outcomes from an early cohort.

The final study of relevance is the recent work by Dehejia and Wahba (1998, 1999). It is this work that has given rise to the very great interest in the potential for using propensity score matching as a means of obtaining better nonexperimental impact estimates. These authors return to the National Supported Work Demonstration as a source of data for the “true experimental impacts” and the basic data for the treatment group and their outcomes. Using a carefully selected subgroup of the Supported Work sample for the treatment group members, they then construct comparison groups from the same two data sets LaLonde used (the Current Population Survey and the Panel Study of Income Dynamics). The propensity score method—which is quite elaborate, as will be described below—was used to focus attention on the small subset of the comparison units comparable to the treated units, hence alleviating the bias due to systematic differences between the treated and comparison units.

The Data Set: Project Star

The experimental data used in this paper is from Project Star, Tennessee’s large-scale class-size experiment (for further discussion, see Word et al., 1990; Krueger, 1999, 2000; and Mosteller, 1995).

From 1985 to 1989, researchers collected observational data including sex, age, race, and free-lunch status from over 11,000 students (Word, 1990). The schools chosen for the experiment were broadly distributed throughout Tennessee. Originally, the project included eight schools from nonmetropolitan cities and large towns (for example, Manchester and Maryville), 38 schools from rural areas, and 17 inner-city and 16 suburban schools drawn from four metropolitan areas: Knoxville, Nashville, Memphis, and Chattanooga. Beginning in 1985–86, the kindergarten teachers and students within Project Star classes were randomly assigned within schools to either “small” (13–17 pupils), “regular” (22–25), or “regular-with-aide” classes. New students who entered a Project Star school in 1986, 1987, 1988, or 1989 were randomly assigned to classes. Because each school had “the same kinds of students, curriculum, principal, policy, schedule, expenditures, etc, for each class” and the randomization occurred within school, theoretically, the estimated within-school effect of small classes should have been unbiased. During the course of the project, however, there were several deviations from the original experimental design—for example, after kindergarten the students in the regular and regular-with-aide classes were randomly reassigned between regular and regular-with-aide classes, and a significant number of students switched class types between grades. However, Krueger found that, after adjusting for these and other problems, the main Project Star results were unaffected; in all four school types students in small classes scored significantly higher on standardized tests than students in regular-size classes.

In this study, following Krueger’s example, test score is used as the measure of student achievement and is the outcome variable. For all comparisons, test score is calculated as a percentile rank of the combined raw Stanford Achievement reading and math scores within the entire sample distribution for that grade.

The Project Star data set provides measures of a number of student, teacher, and school characteristics. The following are the variables available to use as measures prior to random assignment: student sex, student race, student free-lunch status, teacher race, teacher education, teacher career ladder, teacher experience, school type, and school system ID. In addition, the following variables measured contemporaneously can be considered exogenous: student age, assignment to small class size.

After presenting the results, we will return to review the strengths and weakness of the Project Star data with respect to their use to assess, as we do here, performance of nonexperimental methods.

EXPLORING PROPENSITY SCORE MATCHING FOR NONEXPERIMENTAL IMPACT ESTIMATES

Propensity Score Matching

We begin with the nonexperimental method that uses propensity score matching as the method for constructing the comparison group. One might have chosen to present this complicated method after presenting results with simpler and older methods. However, we lead off with this method because it is our perception that there is tremendous current interest in the research and policymaking communities about the efficacy of this nonexperimental method.

For the propensity score matching, we restricted ourselves to the data for kindergartners. There were two reasons for this. First, these data are not subject to complications that arose in later grades due to sample attrition, some switching in class types, and entry of new sample at a postkindergarten grade. Second, the complications of implementation of the full propensity score matching and subsequent analysis mean that running even a single regression example took 3 to 5 days of continuous computer run time. Therefore doing it for all the grades and different sample stratifications would have been inordinately time consuming.

We began by selecting kindergarten classes from the 11 schools that had over 100 kindergartners. Using both treatments and controls from these classes, we estimated the “true impact” for each of the 11 classes as the mean difference in test score percentile and obtained the standard errors to test for the statistical significance of that impact estimate. This is the standard against which we judged the efficacy estimates of impact obtained from the propensity score matching.

We next constructed estimates of impact separately for each of the 11 schools. To construct the comparison group we used as an initial pool all the data for all the control group students in kindergarten for all the schools in the sample *except for the one school being tested.* For example, call the school being tested X, then the comparison pool is all the students in kindergarten control groups in *non-X schools.* The propensity-score-matching procedures are then used to select from among this pool those students who are most closely matched to the treatment students in school X.

To do the actual propensity score matching, we followed as closely as we could the procedures prescribed in Dehejia and Wahba. Below is a brief description of our propensity-score-matching procedures—a somewhat fuller description is provided in Appendix A.

A sample is made by combining the treatments from the selected school with all of the pool of controls from other schools. Using a dummy variable equal to 1 if a treatment and 0 otherwise as a dependent variable, we run a logit regression with all of the characteristics (described above) of the students, the teachers, and the school types as independent variables. The coefficients from that logit regression are then used to generate a *propensity score* for each observation in both the treatment group and pool of controls from other schools, i.e., the predicted probability of being in the treatment group based on observed characteristics.

Next, the distribution of propensity scores, ranked from highest to lowest, is divided into a series of “bins,” and a test is performed to determine whether *within each bin* the mean propensity score value for the treatment group members is significantly different from the mean propensity score for those from the control group pool. If there are significant differences, the boundaries of the bins are shifted until within each bin there is “balance” between treatments and those from the pool of controls. Once “balance” is achieved, for each bin, a test is performed to determine whether within each bin all the observed characteristics are jointly insignificantly different between treatment group members and those from the pool of controls. If there remain significant difference in covariates, one goes back to the logit regression stage and changes some of the variables, e.g., adding higher-order terms in the variables or cross products, and redoes the whole procedure.

Once “balance” has been achieved within bins according to both tests, one proceeds to carry out the matching on propensity scores. The match is made by taking for each treatment group member the observation from the pool of controls which is the “nearest neighbor” in terms of propensity score. The result of this matching is a constructed comparison group made up of the selected observations from the pool of controls.

Now one is in a position to make nonexperimental estimates of the impact on test scores of being in a small class—the estimated difference in mean test percentile between the kindergarten treatment group (in small classes) in the selected school and the comparison group of propensity-score-matched kindergarten students (not in small classes) drawn from all the other schools. This is the nonexperimental estimate of the impact of being in a small class on the test scores of kindergartners.

We added two further steps. First, to obtain appropriate standard errors for the nonexperimental estimate of impact, we performed bootstrapping 500 times on the matching algorithm.[1] Second, given our framework for assessing nonexperimental estimates by comparing them to the “true impact” estimate obtained from comparing the randomly assigned treatments and controls within the same school, we wanted to test whether the *difference between the nonexperimentally estimated impact and the “true experimentally estimated impact” was statistically significant.* We performed this test for each round of

[1] Carrying out the bootstrapping procedure when the estimator involves propensity score matching is an extremely computer-intensive process. Our bootstrapping to produce standard errors with 500 repetitions took 3 to 5 days of continuous run time for just a single school.

the bootstrap, so we are able to report the relative percentage of the time that there was a statistically significant difference between the nonexperimental estimate and the experimental estimate.[2]

The regression run to generate the estimated impact of being in smaller classes was the same for the estimate based on the true experimental sample and for the estimate based on the nonexperimental sample used. The dependent variable in the regression is the ranked test score, as defined earlier. The independent variables include the demographic characteristics of the sample member prior to random assignment (sex, free-lunch status, and race), a dummy variable equal to 1 if the sample member is in the small-class-size group and 0 if the member is in the comparison (nonexperimental) or control (experimental) group. Table 1 provides a summary of the propensity score results.

Most readers will not be familiar with the forms in which many of the results are presented in this table. Each row represents the results for one of the 11 schools. We take the first row, which is school number 7, and walk across the table, explaining each column entry.

Column 1, labeled **School ID,** gives an identification number for the individual school, school 7. This is the school for which the estimates reported in the rest of the row apply.

Column 2, labeled **Small Exp,** gives the estimate of the effect on percentile test scores of small class size obtained using the random-assignment experimental and control sample, namely, mean test score of the treatments in small classes was 5.16 percentile points *less* than the mean test scores of those not in small classes in school 7.

Column 3, labeled **P Small Exp,** gives the probability that an estimate of difference in mean test score due to class size reported in column 2 could have been obtained from a sample of this size if the true difference in mean test scores between those in small classes and those not in small classes was 0. In this case, there was a 37 percent probability, so by the usual statistical standards, a 5 percent chance, we

[2] In doing the test in each round we used the standard error calculated in the normal way for that round. We recognize that these are biased estimates of the actual standard errors for this estimator—which is why we carried out the bootstrapping.

TABLE 1
Project Star: Propensity Score Matching

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
School ID	Small Exp	P Small Exp	St. Error Exp	Small Nonexp	P Small Nonex	St. Error Non	P Diff - Exp/Nonex	St. Error Boot	Sig	Not sig
7	-5.160	0.371	5.739	20.912	0.013	8.139	0.003	10.224	193	307
9	13.039	0.005	4.551	30.687	0.000	5.766	0.000	7.798	178	322
16	24.191	0.000	4.247	9.890	0.090	5.750	0.000	6.005	500	0
22	33.090	0.000	5.756	36.115	0.000	7.337	0.843	9.819	387	113
27	-10.476	0.063	5.588	35.157	0.000	6.494	0.000	8.241	331	169
28	1.332	0.791	5.012	16.810	0.013	6.624	0.283	6.340	117	383
32	10.621	0.022	4.570	15.781	0.043	7.613	0.384	8.178	484	16
33	9.611	0.028	4.295	-6.797	0.224	5.524	0.001	7.593	479	21
51	14.731	0.002	4.568	24.450	0.000	5.680	0.026	8.234	229	271
63	16.153	0.003	5.406	30.334	0.000	6.029	0.002	8.378	182	318
72	19.264	0.003	6.391	45.645	0.000	8.174	0.005	9.413	105	395

would accept (not reject) a hypothesis that in fact the difference in mean test scores between small and large classes was 0.

Column 4, labeled **St. Error Exp,** gives the standard error of the estimate of the difference in mean test scores between large and small. This is the standard error used in conjunction with the mean test score difference in column 2 to calculate the probability given in column 3.

Column 5, labeled **Small Nonexp,** provides the estimate of the difference in mean test scores between those randomly assigned to be in small classes in the given school, here school 7, and the mean test scores of those in the comparison group who were selected by propensity score matching from the pool of students not in small classes; it gives the nonexperimental estimate of the effect of being in a small class in this school. In this case the pool of nonsmall-class students is made up of all the students from the control groups in all the schools *except* school 7. Using the propensity-score-matching comparison group, the nonexperimental impact estimate for school 7 is that those in small classes had mean test scores 20.9 percentile points *above* those in nonsmall classes.

Column 6, labeled **P Small Nonex,** gives the probability that an estimate of difference in mean test score due to class size reported in column 5 could have been obtained from a sample of this size if the *true* difference in mean test scores between those in small classes and those not in small classes was 0. So this is similar to column 3 but in this case applies to the *nonexperimental estimate*. Here the probability is 1.3 percent and thus well below the usual statistical cutoff of 5 percent, so we would *reject* the hypothesis that there was zero difference in mean test scores between those in small classes and those not in small classes.

Column 7, labeled **St. Error Non,** gives the standard error of the estimate of the difference in mean test scores between large and small. This is the standard error used in conjunction with the mean test score difference in column 5 to calculate the probability given in column 6.

Column 8, labeled **P Diff-Exp/Nonex,** gives the probability that a *difference* between the estimate of the *impact* of small class on test scores obtained from the experimental results given in column 2 and the estimate of the *impact* of small class on test scores obtained from the nonexperimental results given in column 5 could have been obtained if there was truly zero difference in the estimates of impact obtained from the two different procedures. In this case, the P level is 0.3 percent, well below the standard cutoff of 5 percent, and indeed below a 1 percent cutoff. We conclude that for this school, the experimental estimate of impact and the nonexperimental estimate of impact are indeed statistically significantly different.

Column 9, labeled **St. Error Boot,** is the estimate of the standard error for the nonexperimental estimate of the impact of small class size on test scores as obtained by the bootstrapping procedure estimated with 500 repetitions of the bootstrap. In this case, the bootstrap standard error is considerably larger than the standard error in column 7, which was calculated in the usual way from the nonexperimental sample.

Column 10, labeled **Sig,** gives the number of times in the 500 bootstrap repetitions that the difference in the experimental impact and the nonexperimental impact was statistically significantly different from 0. In this case, 193 of the 500 repetitions yielded nonexperimental estimates of impact that were significantly different from the experimental impact estimates.

Column 11, labeled **Not Sig,** gives the number of times in the 500 bootstrap repetitions that the difference in the experimental impact and the nonexperimental impact was *not* statistically significantly different from 0. In this case, 307 of the 500 repetitions yielded nonexperimental estimates of impact that were *not* significantly different from the experimental impact estimates.

With this explanation of the table entries in hand, we can give our overview of the results.

How well do the propensity-score-matched estimates of the impact of class size approximate the experimental impact estimates? Our general conclusion would be that across this set of 11 schools tested, they don't do very well, by which we mean the that nonexperimental estimates of impact are not close to the experimental estimates of impact. Now the question "how close is close enough?" requires a more careful answer than the gross generalization just stated. We suggest several ways of addressing this question by discussing various aspects of the data provided in Table 1.

One very important and stringent measure of closeness is whether there are many cases in which the nonexperimental impact estimates are *opposite in sign* from the experimental impact estimates *and both sets of impact estimates are statistically significantly different from 0*, e.g., the experimental estimates said that the mean test scores of those in smaller classes were significantly negative while the nonexperimental estimates indicated they were significantly positive. There is only one case in these 11 which comes close to this situation. For school 27, the experimental impact estimate is −10.5 and significant at the 6 percent level, just above the usual significance cutoff of 5 percent. The nonexperimental impact estimate is 35.2 and significant at better than the 1 percent level. In other cases (school 7 and school 33), the impact estimates are of opposite sign, but one or the other of them fails the test for being significantly different from 0.

If we weaken the stringency of the criterion a bit, we can consider cases in which the experimental impact estimates were significantly different from 0 but the nonexperimental estimates were not (school 16 and school 33), or vice versa (schools 7, 16, and 28).

Another, perhaps better, way of assessing the differences in the impact estimates is to look at column 8, which presents a test for whether the impact estimate from the nonexperimental procedure is significantly different from the impact estimate from the experimental procedure. For eight of the 11 schools, the two impact estimates were statistically significantly different from each other.

Still another way to look at this issue is through inspection of column 10 or column 11. We are less confident that this is a reasonable way to assess the performance of the nonexperimental estimators, but comment on it anyway. If we take the entries in column 10 and divide them by 500, we get the percentage of the bootstrap repetitions in which the nonexperimental estimate of impact was found to be significantly different from the experimental estimate of the impact. These percentages range from a low of 21 percent for school 72 to a high of 100 percent for school 16; at best only one-fifth of the repetitions for a given school yield statistically significant differences between nonexperimental and experimental impact estimates, and at worst all of the repetitions yield statistically significant differences between the two types of impact estimates.[3]

These are some of the various ways of looking at Table 1 that led us to our generalization that with these data the propensity-score-matching nonexperimental estimates do not do very well in approximating the experimental estimates of the impacts of small class size on test score performance.

How Close Is Close Enough for Decision-Making?

Here we take a slightly different approach to the question of closeness of the nonexperimental and experimental estimates.

When we first began considering this issue, we thought that a criterion might be based on the percentage difference in the point estimates of the impact. For example, for school 9 the nonexperimental estimate of the impact is 135 percent larger than the experimental impact. But for school 22 the nonexperimental impact estimate is 9 percent larger. Indeed, all but two of the nonexperimental estimates are more than 50 percent different from the experimental impact estimates.

Whereas in this case the percentage difference in impact estimates seems to indicate quite conclusively that the nonexperimental estimates are not generally close to the experimental ones, in some cases such a percentage difference criterion might be misleading. The criterion which seems to us the most definitive is whether distance between the nonexperimental and the experimental impact estimates

[3] We do not discuss average or median values of these percentages of significant differences between the two estimates as we don’t know what the distribution really represents. There are notable differences among the schools in how well the nonexperimental estimates approximate the experimental ones, but we have not tried to search systematically for what features of the schools might explain such differences.

would have been sufficient to cause an observer to make a different decision from one based on the true experimental results. For example, suppose that the experimental impact estimate had been .02 and the nonexperimental impact estimate had been .04, a 100 percent difference in impact estimate. But further suppose that the decision about whether to take an action, e.g., invest in the type of activity which the treatment intervention represents, would have been a *yes* if the difference between the treatments and comparisons had been .05 or greater and a *no* if the impact estimate had been less than .05. Then even though the nonexperimental estimate was 100 percent larger than the experimental estimate, one would still have decided *not* to invest in this type of intervention whether one had the true experimental estimate or the nonexperimental estimate.

This perspective suggests that in order to decide how close nonexperimental estimates have to be to the experimental ones to be considered "close enough," the point beyond which a decision of "no" would change to a decision of "yes" depends on the specific decision context. Let's call this approach to "close enough" the "wrong decision" criterion.

We can give a rough example of applying this sort of criterion in the case from which these data were drawn, the Project Star experiment with class-size reduction in Tennessee. In a couple of his articles presenting aspects of his research using Project Star data, Krueger (1999, 2000) has developed some rough benefit-cost calculations related to reduction in class size. In Appendix B we sketch in a few elements of his calculations which provide the background for the summary measures derived from his calculations that we use to illustrate our "close enough." The benefits Krueger focuses on are increases in future earnings that could be associated with test score gains. He carefully develops estimates—based on other literature—of what increase in future earnings might be associated with a gain in test scores in the early years of elementary school. With appropriate discounting to present values, and other adjustments, he uses these values as estimates of benefits and then compares them to the estimated cost of reducing class size from 22 to 15, taken from the experience in Project Star and appropriately adjusted.

For our purposes, what is most interesting is the way he uses these benefit-cost calculations to answer a slightly different question: How big an effect on test scores due to reduction of class size from 22 to 15 would have been necessary to just justify the expenditures it took to reduce the class size by that much? He states the answer in terms of "effect size," that is the impact divided by the estimated standard deviation of the impact. This is a measure that is increasingly used to compare across outcomes that are measured in somewhat different metrics. His answer is that an effect size of 0.2 of a standard deviation of tests scores would have been just large enough to generate estimated future earnings gains sufficient to justify the costs.[4] Krueger indicates that the estimated effect for kindergarten was a 5.37 percentile increase in achievement test scores due to smaller class size and that this was equivalent to 0.2 of a standard deviation in test scores. Therefore we use 5.4 percentile points as the critical value for a decision of whether the reduction in class size from 22 to 15 would have been cost-effective.

In Table 2 we use the results from Table 1 to apply the cost-effectiveness criterion to determine the extent to which the nonexperimental estimates might have led to the wrong decision. To create the entries in this table, we look at the Table 1 entry for a given school. If the impact estimate is greater than 5.4 percentile points and statistically significantly different from 0, we enter a **Yes**, indicating the impact estimate would have led to a conclusion that reducing class size from 22 to 15 was cost-effective. If the impact estimate is less than 5.4 or statistically not significantly different from 0, we enter a **No** to indicate a conclusion that the class-size reduction was not cost-effective. Column 1 is the school ID, column 2 gives the conclusion on the basis of the experimental impact estimate, column 3 gives the conclusion on the basis of the nonexperimental impact estimate, and column 4 contains an **x** if the nonexperimental estimate would have led to a "wrong" cost-effectiveness conclusion, i.e., a different conclusion from the experimental impact conclusion about cost-effectiveness.

[4] Actually, Krueger gives several different critical values which vary with different combinations of key assumptions, but from among these we chose the 0.2 of a standard deviation for the purpose of our illustration.

TABLE 2
Cost-Effectiveness Decisions for Project Star Based on 5.4 Percentile Impact Critical Criterion

[1]	[2]	[3]	[4]
School ID	Experimental Estimate	Nonexperimental Estimate	Wrong Decision
7	No	Yes	x
9	Yes	Yes	
16	Yes	Maybe	?
22	Yes	Yes	
27	No	Yes	x
28	No	Yes	x
32	Yes	Yes	
33	Yes	No	x
51	Yes	Yes	
63	Yes	Yes	
72	Yes	Yes	

It is easy to see in Table 2 that the nonexperimental estimate would have led to the wrong conclusion in four of the 11 cases. For a fifth case, school 16, we entered a **Maybe** in Column 4 because, as shown in Table 1, for that school the nonexperimental estimate was significantly different from 0 at only the 9 percent level, whereas the usual significance criterion is 5 percent. Even though the nonexperimental point estimate of impact was greater than 5.4 percentile points, strict use of the 5 percent significance criterion would have led to the conclusion that the reduction in class size was not cost-effective. On the other hand, analysts sometimes use a 10 percent significance criterion, so it could be argued that they might have used that level to conclude the program was cost-effective—thus the **Maybe** entry for this school.

Overall, then, in four or five of the 11 cases considered, the nonexperimental estimates would have led to the wrong cost-effectiveness criterion. Anyone can decide what to make of this indication of the likelihood that the nonexperimental estimates are "close enough." Our own conclusion is that it would be too risky to rely on the propensity-score-matching estimates in this situation.

Would the "Wrong Decision Criterion" Be Operationally Useful in Deciding on Whether to Use a Nonexperimental Method, and If So, Which One?

The interest in the performance of nonexperimental methods arises in the context of making decisions about the approach to take in designing evaluations of the impact of a given policy or program intervention. It seems natural to ask: Would the "wrong decision criterion" be operationally useful in deciding on whether to use a nonexperimental method, and if so, which one? In other words, could this "wrong decision" approach be operationally applied?

Two elements are required to operationalize the "wrong decision criterion." First, one needs a value for the critical impact threshold, an impact magnitude that would determine that the intervention was justified on benefit-cost, or similar, grounds. In practice, in carefully designed impact evaluations, experimental or nonexperimental, we often develop such a "critical impact" value as part of developing our estimates of the required sample size for the study; we want the sample size to be large enough to assure the statistical power to have a high probability of detecting an impact of that magnitude or greater if it actually occurs.

The second element of the "wrong decision criterion" is a reasonable guess as to how far off the "true impact" (the bias of the nonexperimental estimate), and in which direction, the nonexperimental impact might be in a given case. Looking across our 11 cases, we see tremendous variation in "how close" this particular nonexperimental estimator was to the experimental. Unless we accumulate enough studies similar to this one so that we have reasonable grounds for judging how far off the nonexperimental estimates are likely to be in a given situation, we cannot operationally implement this "wrong decision criterion."

Weaknesses of This Study as an Assessment of Propensity Score Matching and Their Relevance to an Evaluation Design Decision Context.

We want to explicitly recognize the limitations of these data as a means of assessing the performance of propensity score matching.

On the positive side, as noted above, the Project Star data provide an unusual opportunity because they come from a "true random-assignment" intervention and they represent not just one experiment but as many experiments as there were schools (79). In addition, the sample sizes are quite large even at the individual school level. All the variables are defined and measured the same way across all the sites. Thus we were able to use treatment groups from one school, or set of schools, and then the controls from another set of schools as the pool from which a nonexperimental comparison group could be constructed.

For the purposes of testing propensity score matching, and some other nonexperimental methods, these data have some weaknesses. For any matching method, the larger the number of covariates the better. In the Project Star data there are only a limited number of covariates that can be used in the matching process. For our matching model we could use student age, student sex, student race, student free-lunch status, teacher race, teacher education, teacher career ladder, teacher experience, school type (rural, urban, inner-city, suburb), and school system ID. The Project Star data have no preintervention measure for the outcome variable, i.e., there are no achievement tests scores for students before they were randomly assigned to small classes or to regular-size classes. For many nonexperimental estimators, such preintervention measures can improve the matching and can be used for pre-post estimates (or difference in differences) of impact. Finally, many researchers who attempt to use nonexperimental estimators of impacts argue that it is important that the pool from which comparison group members will be drawn be located as close to the site where the treatment group members are located—e.g., for earnings and employment outcome studies, that comparisons and treatments be from the same labor market. Some have argued that the propensity score match pool of potential comparisons be from the same school. This is not feasible in this study because we are using control group members as the potential comparison group pool, and if we took comparison group members from the controls in the same school we would essentially be replicating the experiment.

Recognizing these limitations, we argue, however, that in many situations in which decisions are to be made about the design of an impact evaluation, the decision-makers would face similar limitations to the data they have or could generate. For example, if one were testing an intervention in kindergarten, it would be difficult to get preintervention measures on test scores. Even for older youth, nonexperimental estimators that depend on several periods of preintervention measurement on the outcome variable in order to estimate individual trajectories in the outcome change over time and then try to match on trajectories may find these methods far less powerful for youth than they are for older persons where trajectories are more stable. Finally, while it is legitimate to argue that one probably will get a better match when the members of the potential comparison pool are located close to the treatment group members, e.g., in the same school, if they are really that near then it seems it should be feasible to do a random-assignment design and obtain the far greater reliability of the experimental design estimate of the impact.

COMPARING PROPENSITY-SCORE-MATCHING PERFORMANCE WITH THAT OF A SIMPLER NONEXPERIMENTAL METHOD

We actually began this project by testing comparison groups drawn from the pool of control group members and using a simpler nonexperimental estimator that uses only multiple regression with covariates included to attempt to deal with bias. We will refer to this as the multiple-regression-correction estimator. We will present some of these results later in the paper. Here we use the multiple-regression-correction procedure for the kindergartners in the same 11 schools for which we estimated propensity-score-matched impact estimates. We then make a rough comparison of how well the multiple-regression-correction impact estimates compare with the propensity-score-matched impact estimates.

Using the treatment group for each of the 11 school kindergartens reported in Table 1, we created the comparison group from the kindergarten control group members for all the schools except the one whose treatment group members we were using, i.e., the same pool from which we had drawn comparisons using the propensity scores. We drew the comparison group from this pool in three different ways. First, we simply used everyone in the pool as a member of the comparison group (about 3,900 cases). We refer to this in Table 3 as Type 1. Second we took a random sample from that pool of size approximately equal to the size of the treatment group for that school (ranging from about 100 to 135). Third, we used only school of the same type as the treatment school for the comparison pool, e.g., when the treatment school was rural, all other rural school controls made up the comparison group (comparison sample ranged from 290 to 1,830).[5]

[5] Results using the other two comparison group pools yielded qualitatively the same results.

For each type of comparison sample drawn, we ran the nonexperimental regressions using the comparison group selected and a dummy for small class and the covariates as right-hand-side variables. Thus, we obtained a simpler nonexperimental impact estimate for the same 11 kindergarten classes.

We will not reproduce all the estimates for the three different methods of drawing the comparison group but simply select the first—all the controls from schools other than the one we are using for treatments—and then compare them to the propensity-score-matching estimates and the true experimental estimates for that school. Table 3 provides this comparison.

The first column of Table 3 gives the school ID. The second column reports the nonexperimental estimate of the class-size effect using a comparison group of all the controls in kindergartens except the one for the given school. The third column gives the propensity-score-matching estimate of the class-size effect for that school. The fourth column gives the experimental estimate of the class-size effect for that school. The fifth column simply indicates which nonexperimental estimate was closest to the experimental estimate.[6] We call the regression with covariates Type 1 to indicate this was the type in which we used the entire comparison pool.

This crude test shows that the propensity-score-matching estimate performed better than the simpler multiple regression estimate only for two of the 11 school estimates. It is not clear to us why the propensity score estimates should have been worse, since the propensity score comparisons are drawn from the pool that makes up Type 1 for the simple regressions, but that is what we find at this stage, so we would not put great emphasis on this result. We felt, however, that we wanted to see if the more complex propensity-score-matching procedures clearly performed better than the simpler nonexperimental regression-adjusted estimates, as some would argue it should. We do not see any better performance by the propensity-score-matched estimators on the basis of these data.

[6] We are just using the crude test of how close the point estimates are. We could do more with this by taking into account standard errors and repeating some of the alternative ways of judging closeness but decided not to present all of that analysis.

TABLE 3
Simple Regression versus Propensity Score Matching

	Multiple-Reg Type 1	Pscore	Experimental	Closest
School ID	Small Nonexp	Small Nonexp	Small Exp	
7	9.639	20.912	-5.160	**Type 1**
9	15.300	30.687	13.039	**Type 1**
16	1.861	9.890	24.091	**Pscore**
22	35.171	36.115	33.090	**Type 1**
27	20.363	35.157	-10.476	**Type 1**
28	0.200	16.810	1.332	**Type1**
32	-8.616	15.781	10.621	**Pscore**
33	-5.602	-6.797	9.611	**Type 1**
51	16.510	24.450	14.731	**Type 1**
63	24.320	30.334	16.153	**Type1**
72	27.774	45.645	19.264	**Type 1**

MULTIPLE-REGRESSION-CORRECTION ESTIMATORS WITH COVARIATES TO DEAL WITH BIAS IN COMPARISON GROUP ESTIMATES OF IMPACT

In this section we describe the much larger set of results we obtained using the multiple regression including the covariates for correction for bias, multiple-regression-correction nonexperimental estimators. We produced results using this type of nonexperimental estimator for a much wider set of cases. We estimate them separately for kindergarten through third grade and for different groupings according to school type and district type.

The first set of 28 nonexperimental control groups helped us to compare estimates made using *within* school comparison groups, i.e, experimental estimates (children in small classes in school X to children in regular classes in school X), to estimates made by comparing *across* schools (children in small classes in school X to children in regular classes in school Y). Schools chosen were of the *same demographic type*: inner-city, suburban, rural, or urban. For example, to construct the first nonexperimental control group of inner-city kindergartners, the total sample of inner-city kindergarten students was divided into two cohorts, by student ID number (cohort 1 included students in A, B, C, D, E, F, G schools, whereas cohort 2 included students in T, U, V, W, X, Y, Z schools).

To obtain the “true” experimental estimate of the effect on inner-city kindergartners of being in a small class, only children in the first group or first half of all inner-city kindergartners (including those in the small, regular, and regular-with-aide classes) were used. For the experimental estimate, test scores of children in small classes were compared with test scores of other children within their schools who were in regular and regular-with-aide classes (children in small classes in A, B, C, D, E, F, G were compared with children in regular and regular-with-aide classes in A, B, C, D, E, F, G). To obtain the nonexperimental, constructed estimate, however, the children in the small classes from the *first* cohort were compared with the children in regular and regular-with-aide classes in the *second* cohort (children in small classes in A, B, C, D, E, F, G were compared with children in regular and regular-with-aide classes in T, U, V, W, X, Y, Z).

For each demographic type (inner-city, suburban, rural, or urban), this process was repeated for kindergarten, first grade, second grade, third grade, cumulative first grade (cumulative—defined as excluding those children who joined Project Star in that particular year of analysis), cumulative second grade, and cumulative third grade.

The second type of comparison group, involving 21 pairs of regressions, compared estimates obtained within and across *districts*. The *district* was the comparison unit. A district type was determined by the percentage of students of a given school type within that district. For example, if more than 50 percent of the students in a district attended rural schools, the district was classified as rural. By this definition, there were seven urban systems, one inner-city system, five suburban systems, and 29 rural systems.

Because the majority of schools in Tennessee are rural, this analysis focused on rural districts. For the experimental comparison across district type, program and control groups were constructed for rural districts. For the experimental estimate, the test scores of students in small classes were compared with students in regular classes in rural districts (i.e., children in small classes in rural districts A, B, C, D were compared with children in regular classes in those same districts). For the nonexperimental control groups, however, groups defined from other district types (either inner-city, suburban, or urban) were used. For example, for the first nonexperimental control group, children in regular classes in inner-city kindergartens were used (children in small classes in rural districts A, B, C, D were compared with children in regular classes in inner-city districts O, P, Q, R) Again, these evaluations were repeated for kindergarten, first grade, second grade, third grade, cumulative first grade (see definition above), cumulative second grade, and cumulative third grade.

THE ECONOMETRIC SPECIFICATION

For all the comparison groups, a linear regression model was used to produce the experimental and nonexperimental estimates of program effect. The same specification was used for each method.

The dependant variable in the regression is the ranked test score, as defined earlier. The independent variables include the demographic characteristics of the sample member prior to random assignment (sex, free-lunch status, and race), a dummy variable equal to 1 if the sample member is in the small-class-size group and 0 if the member is in the comparison (nonexperimental) or control (experimental) group, along with four fixed-effects dummy variables that reflect the year in which the student entered Project Star.[7]

The estimate of the effect of being in small classes is the coefficient on the small class dummy. Given that the randomization within Project Star has been implemented correctly, the experimental estimate of the effect of small class size should be unbiased. When the control groups have been *constructed*, as in the nonexperimental comparison groups, however, the estimate of program effect may or may not be biased.

To test the significance of the difference in the program effect between the experimental and nonexperimental estimates, we end up essentially testing for the significance of the difference between the mean test scores for the actual experimental control group and the mean test scores for the nonexperimentally constructed comparison group. This is because in the estimate of the program impact the treatment group (those in small classes) *is the same* for both the experimental and the nonexperimental impact estimates. Thus any difference between the two impact estimates is generated by

[7] Work by Krueger and Whitmore (1999) inspired this model specification. Their results suggest that the year of entry into Project Star greatly influences the significance of the effect of small classes on student achievement.

the differences in outcomes for the experimental control and the constructed nonexperimental comparison groups.

To implement this test with the two (across schools and across district type estimates nonexperimental methods), we ran another regression, including only members of the experimental control group and constructed comparison groups, and regressed the four demographic characteristics and the entry year dummy variables, plus a dummy variable equal to 1 for the experimental control group, on the ranked test score. The coefficient on the "control" dummy variable represented the mean difference between the experimental control and nonexperimental comparison groups.

Empirical Results

Table 4 presents a number of selected experimental and nonexperimental estimates of effects for the school and district type comparisons. In each row of the table, for a given estimate of program effect four items are listed: (1) the estimated program effect (2) the standard error of the estimated program effect, (3) the probability value for the test of significant difference of the coefficient on small, the program coefficient, and finally (4) the sample size. In the second and third columns of Table 4, sample sets of comparison experimental and nonexperimental estimates are given, with one from each of the school demographic types (inner-city, suburban, rural, urban) and each of the experimental/nonexperimental district type estimations (rural-city, rural-suburban, rural-urban).

In all seven selected cases, the experimental and nonexperimental estimates differ considerably from each other. One of the nonexperimental estimates is of the wrong sign, while in the other estimates, the signs are the same but all the estimates differ by at least 1.8 percentage points, ranging up to as much as 12 percentage points (rural-city). Statistical inferences about the significance of these program effects also vary (five of the seven pairs had differing inferences—i.e., only one estimate of the program effect in a pair is statistically significant at the 10 percent level). All of the differences between the experimental and nonexperimental estimates (the test of difference between the outcomes for the

TABLE 4
Selected Representation of Various Estimates of the Effect of Small Class Size on Student Achievement

Program (year)	Experimental	Nonexperimental
Inner City (K)		
Program effect	3.258	9.556
Standard error of small coefficient	2.433	2.567
P value of small coefficient	.168	.000
Sample size	N=684	N=603
Suburban (1)		
Program effect	13.158	6.188
Standard error of small coefficient	2.047	2.148
P value of small coefficient	.000	.004
Sample size	N=735	N=777
Rural (1-cum)		
Program effect	.570	-5.805
Standard error of small coefficient	1.757	1.644
P value of small coefficient	.746	.000
Sample size	N=1098	N=1144
Urban (1)		
Program effect	4.177	11.518
Standard error of small coefficient	3.915	3.408
P value of small coefficient	.287	.001
Sample size	N=246	N=313
Rural-City (1)		
Program effect	3.126	15.976
Standard error of small coefficient	1.097	2.011
P value of small coefficient	.004	.000
Sample size	N=3014	N=2063
Rural-Suburban (1 cum)		
Program effect	3.384	1.433
Standard error of small coefficient	1.220	1.866
P value of small coefficient	.006	.443
Sample size	N=2249	N=1034
Rural-Urban (2)		
Program effect	1.069	4.174
Standard error of small coefficient	1.110	1.813
P value of small coefficient	.336	.022
Sample size	N=2859	N=1159

experimental control group and the nonexperimental comparison group) in this subset were statistically significant.

Table 5 shows the results for the complete set of the first 49 pairs of estimates. Each column shows a different type of comparison (either school type or district type). The top row in each column provides the number of pairs of experimental and nonexperimental estimates in the column. The second row shows the mean estimate of program effect from the (unbiased) experimental estimates. The third row has the mean absolute differences between these estimates, providing some indication of the size of our nonexperimental bias. The fourth row provides the percentage of pairs in which the experimental and nonexperimental estimates led to different inferences about the significance of the program effect. The fifth row indicates the percentage of pairs in which the difference between the two estimated values was significant (again the test of difference between control and comparison group).

Looking at the summarized results for comparisons across school type, these results suggest that constructing nonexperimental groups based on similar demographic school types leads to nonexperimental estimates that do not perform very well when compared with the experimental estimates for the same group. In 50 percent of the pairs, experimental and nonexperimental estimates had different statistical inferences, with a mean absolute difference in effect estimate of 4.65. Over 75 percent of these differences were statistically significant. About half of the estimated pairs in comparisons across school type differ by more than 5 percentage points.

Turning next to the summarized results for the comparisons across district types (using urban, suburban, and inner-city districts as control groups for the rural districts), these results also indicate that district-level comparisons may not be trustworthy. In particular, in 52 percent of the pairs, the experimental and nonexperimental estimates had different statistical inferences, with a mean absolute difference in estimate of 2.6 percentage points. Over 80 percent of these differences were statistically significant. Approximately one-half of the experimental and nonexperimental estimates differ by more

TABLE 5
Summary of Experimental and Nonexperimental Estimates of Effect of Small Class Size

	Comparison Group Specification	
Statistic	Across-School Type	Across-District Type
Number of pairs	28	21
Mean experimental estimate	5.536	2.621
Mean absolute experimental-nonexperimental differences	4.649	4.925
Percentage with different inference	.500	.523
Percentage with statistically significant difference (10% level)	.796	.809

than 5 percentage points. It appears from Table 5 that using same school type to draw the comparison group does somewhat better than using same district type. However, for neither set do the nonexperimental multiple-regression-correction estimators perform very well. We have not gone into the same type of detail about measures of closeness and how close is close enough as we did for the propensity-score-matched nonexperimental estimators because we felt there was far greater interest in the performance of the more complex propensity score method.

REPRISE AND CONCLUSIONS

We set out to test the performance of some nonexperimental estimators of impacts applied to an educational intervention—reduction in class size—where achievement test scores were the outcome. We tested the performance by comparing the nonexperimental estimates of the impacts to "true impact" estimates provided by a random-assignment design used to assess the effects of that intervention.

Previous important studies that have used this approach of testing nonexperimental estimators against experimental ones have all had employment or welfare use as the outcome variable. We felt it important to try to determine whether the conclusions from those studies about the performance of nonexperimental estimators carried over into the educational domain.

Our primary focus in this study was on a nonexperimental estimator based on a complex procedure called propensity score matching. In recent years, social policy analysts have become increasingly interested in finding out to what degree this method might provide an adequate alternative to estimators derived from random-assignment designs. We report, also, some tests of much simpler estimators in which multiple regression including covariates is used to attempt to correct for bias in the nonexperimental estimators.

We carried out the propensity-score-matching estimating procedure and used it to derive nonexperimental estimates of the impact of smaller class size. We also developed proper standard errors for the propensity-score-matched estimators by using bootstrapping procedures.

We found that in most cases, the propensity-score estimate of the impact differed substantially from the “true impact” estimated by the experiment.

We then attempted to address the question, “How close are the nonexperimental estimates to the experimental ones?” We suggested several different ways of attempting to assess “closeness.” Most of them led to the conclusion, in our view, that the nonexperimental estimates were not very “close” and therefore were not reliable guides as to the “true impact.”

We put greatest emphasis on looking at this question of “how close is close enough?” in terms of a decision-maker trying to use the evaluation to determine whether to invest in wider application of the intervention being assessed—in this case reduction in class size. We illustrated this in terms of a rough cost-benefit framework for small class size as applied to Project Star. We found that in 35 to 45 percent of the 11 cases where we had used propensity score matching for the nonexperimental estimate, it would have led to the “wrong decision,” i.e., a decision about whether to invest which was different from the decision based on the experimental estimates.

We next noted the strengths and weaknesses of the Project Star data set as a source for testing the performance of nonexperimental estimators when compared with the experimental estimates. The primary strength is that the variables were measured in exactly the same way, and that there were a large number of schools with each school and each of three grades in that school being capable of serving as a separate experiment. The primary weaknesses are that there were no measures of the outcome variable for individual students in the period prior to random assignment and that the number of measures of the characteristics of the students and their teachers (potential covariates) was relatively small.

Recognizing that these data are not ideal, we argue that they still may reflect realistically a type of context which a decision-maker would face in judging, before the evaluation began, whether the nonexperimental design is likely to provide adequate, reliable information.

We attempted to establish, in a very rough way, how well the propensity-score-matching estimators do relative to a much simpler nonexperimental estimator. This question might be relevant for some cases in which resort to an adequate random-assignment design is just not feasible; if one's only choice is among nonexperimental estimators, which appear to perform better?

We did this test by contrasting the propensity-score-matching estimates of impacts with nonexperimental estimates that use a multiple regression, based on the covariates available, to attempt to correct for selection bias. The crude test is simply to see whether the propensity-score-matched estimate of the impact was closer to the experimental estimate of impact than the multiple-regression-correction estimate of the impact. We carried out this test for each of the 11 school-kindergartens for which we had developed propensity-score-matched estimates. Somewhat to our surprise, the propensity-score-matched estimator did not perform notably better than the multiple-regression-corrected estimators for most of the 11 cases.

Finally, we briefly reported some results obtained from using the multiple-regression-correction nonexperimental estimation method. We were able to do this for several grade levels and with several permutations on what pool the comparison sample was drawn from.

None of these multiple-regression-correction nonexperimental estimation methods appeared to perform very well where the performance criterion was how close their impact estimate was to the "true impact" experimental estimate obtained from the random-assignment design.

All of this work was carried out in a context of attempting to estimate the impact of an educational intervention—smaller class size—on one type of educational outcome—achievement test scores. In that context, our conclusion is that nonexperimental estimators do not perform very well when judged by standards of “how close” they are to the “true impacts” estimated from experimental estimators based on a random-assignment design.

In addition, one might wonder whether this type of study could give much guidance for a choice *among nonexperimental estimators*—a search for a second-best estimator. Unfortunately, in this study, which tested both a very complex nonexperimental estimator—propensity score matching—and a relatively simple nonexperimental estimator—multiple regression correction—no clear second-best could be identified.

APPENDIX A
Details of Propensity Score Matching and Bootstrapping Procedures

The first type of comparison group was constructed using propensity-score-matching methods. The sample was restricted to only kindergartners, and selected from all of the Project Star schools the 11 schools with over 100 kindergartners. Using a logit model, we regressed treatment group status on all of the reported exogenous observable characteristics, including teacher education, teacher career ladder, student race, student free-lunch status, student sex, student age, teacher race, teacher experience, school type, and system ID number. Separate logit models were estimated for each of the 11 schools. We used these logit results to generate a propensity score, or predicted probability of inclusion in the program group, based on observable characteristics.

Following Dehejia and Wahba's (1998) suggested method, we ranked all of the propensity scores. We then designed an algorithm to generate bins across the set of propensity scores, for which the differences between propensity scores for program and control groups were insignificant. Next, we designed and implemented another algorithm to generate bins across the set of propensity scores, for which all of the covariates, or observable characteristics between program and control groups, were jointly insignificant.[8] Initially, having estimated the propensity score using only the original characteristics (listed above), the covariates across bins did not balance. Consequently, as Dehejia and Wahba suggest, we began adding higher-order terms. Ultimately, after adding these higher-order terms[9] and reimplementing the bin-balancing algorithms, all of the covariates balanced. Having passed Dehejia's specification test, we used this set of variables and covariates to estimate the propensity score used in creating all of the propensity score nonexperimental comparison groups.

[8] We used a Hotelling's T-squared test.

[9] The higher-order terms, (teacher highest degree)^3 and (total experience)^3, replaced the original terms, teacher highest degree and total experience, in the propensity-score-estimating algorithm.

For the 11 schools with over 100 kindergartners, we estimated the experimental effect of small classes. After estimating and ranking the propensity scores for the complete set of kindergartners, we used a nearest-neighbor matching without replacement multiloop algorithm to select, for each program group propensity score within a chosen school, the nonexperimental control (nonsmall) propensity score outside the school nearest in absolute value to the program propensity scores. This set of scores, matched to the program groups and selected from the entire set of all kindergartners in regular and regular-with-aide classes outside of the selected school, together became the constructed comparison group, used to yield estimates of program effect. To obtain accurate estimates of the program effect and standard errors for each of the 11 schools, the matching without replacement algorithm was bootstrapped 500 times. The same logit model variables (determined by the “balance testing” for the initial estimate) were used in each bootstrap replication, but the coefficients in the logit were reestimated each time, so the estimated propensity scores for each observation could differ from one replication to another.

APPENDIX B
Krueger's Benefit-Cost Calculations for Project Star

Krueger (1999, 2000) limits himself to assessing the benefits of class-size reduction in terms of future earnings gains that may be associated with increases in achievement test scores generated by the smaller class size. To obtain an estimate of the link between test scores in elementary school and future earnings, Krueger relies on several other studies. After carefully assessing them, he concludes that a reasonable estimate is that a one standard deviation increase in either math or reading scores can be translated into an 8 percent increase in average real earnings throughout the earnings lifetime of an individual. Using the 8 percent earnings increase per standard deviation increase in test scores and age-earning profile for workers in 1998, he is able to translate the increase in test scores into the estimated increase in real earnings at each age. Using the Project Star data, he estimates that reducing class size from 22 students to 15 students was associated with a 0.20 standard deviation increase in math and reading scores.

Krueger then notes that productivity growth over time should lead to increase in real earnings levels and presents three possible assumptions about annual productivity growth in the future—0 percent,1 percent, and 2 percent. Then it is necessary to choose a discount rate to convert the stream of lifetime earnings into a present value. He presents five different discount rates—.02, .03, .04, .05, and .06. The combination of productivity growth and discount rate pairs gives him 15 different values for the discounted present value of future earnings generated by the reduction in class size from 22 to 15.

He estimates costs of class-size reduction from the Project Star experience. These costs incurred over 3 possible years in smaller classes should then be discounted to obtain present value of costs. He gives the five different values of the discounted present value of costs, one associated with each of the five discount rates.

The two sets of discounted present values, one for benefits and one for costs, can be combined to yield 15 benefit-to-cost ratios each dependent on the choice of assumption about the future productivity growth and the appropriate discount rate. An annual productivity growth rate of 1 percent and a discount rate of 6 percent would make the discounted present value of benefits just about equal to the discounted present value of costs for class-size reduction from 22 to 15.

A related calculation, which Krueger makes on the basis of these sorts of data and assumptions, attempts to answer the question, “What is the minimum increase in test scores from a reduction in class size of seven students in grades K-3 that is required to justify the added cost?” In much of the evaluation literature, impacts are translated into standard deviation units in what is called the “effect size.” An “effect size” is the estimated impact on the outcome variable divided by the estimated standard deviation of the outcome variable. So the question can be restated as, “What is the critical effect size for impact on test scores that would justify the added cost of the class-size reduction?” He can obtain the answer to this using the assumptions and data just outlined by solving the benefit-cost expression for the value of the number of standard deviations in test scores that would just cause the present value of benefits to equal the present value of costs. This estimate will be different for each pair of assumptions about the annual growth rate of future productivity and the discount rate. Thus he gives a table with 15 different possible values. If, out of these 15, one focuses, as we do in the text, on the 1 percent productivity growth and 6 percent discount rate, the “critical effect size” is 0.19 for the math and reading test scores. From the data provided in Krueger’s papers, we can determine the a 0.2 standard deviation in the combined math and reading test scores translates into an effect amounting to a 5.4 percentile point increase, and therefore that is the critical criterion value we use in constructing Table 2.

References

Dehejia, Rajeev, and Sadek Wahba. 1998. “Propensity Score Matching Methods for Non-Experimental Causal Studies.” National Bureau of Economic Research Working Paper # 6829.

Dehejia, Rajeev H., and Sadek Wahba. 1999. “Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs” *Journal of the American Statistical Association* 94 (448): 1053–1062.

Fraker, Thomas, and Rebecca Maynard. 1987. “Evaluating Comparison Group Designs with Employment-Related Programs.” *Journal of Human Resources* 22: 194–227.

Friedlander, Daniel, and Philip Robins. 1995. “Evaluating Program Evaluations: New Evidence on Commonly Used Nonexperimental Methods.” *American Economic Review* 85 (4): 923–937.

Krueger, Alan. 1999. “Experimental Estimates of Education Production Functions.” *Quarterly Journal of Economics* 114 (2): 497–532.

Krueger, Alan. 2000. “The Class Size Policy Debate.” Economic Policy Institute Working Paper No.121.

Krueger, Alan, and Diane Whitmore. 1999. “The Effects of Attending a Small Class in the Early Grades on College-Test Taking and Middle School Test Results: Evidence from Project Star.” Institute for Industrial Relations Working Paper #427, Princeton University.

LaLonde, R. 1986, “Evaluating the Econometric Evaluations of Training Programs.” *American Economic Review* 76: 604–620.

Mosteller, Frederick. 1995. The Tennessee Study of Class Size in the Early School Grades.” *The Future of Children* 5: 113–127.

Word, Elizabeth, John Johnston, Helen Bain, et al. 1990. “The State of Tennessee’s Student/Teacher Achievement Ratio (STAR) Project.” Tennessee Board of Education.