How Close Is Close Enough? Evaluating Propensity Score Matching Using Data from a Class Size Reduction Experiment

Elizabeth Ty Wilde
Robinson Hollister

Abstract

In recent years, propensity score matching (PSM) has gained attention as a potential method for estimating the impact of public policy programs in the absence of experimental evaluations. In this study, we evaluate the usefulness of PSM for estimating the impact of a program change in an educational context (Tennessee's Student Teacher Achievement Ratio Project [Project STAR]). Because Tennessee's Project STAR experiment involved an effective random assignment procedure, the experimental results from this policy intervention can be used as a benchmark, to which we compare the impact estimates produced using propensity score matching methods. We use several different methods to assess these nonexperimental estimates of the impact of the program. We try to determine "how close is close enough," putting greatest emphasis on the question: Would the nonexperimental estimate have led to the wrong decision when compared to the experimental estimate of the program? We find that propensity score methods perform poorly with respect to measuring the impact of a reduction in class size on achievement test scores. We conclude that further research is needed before policymakers rely on PSM as an evaluation tool.

INTRODUCTION

This study has several objectives. The overriding objective is to evaluate the usefulness of a particular nonexperimental method—propensity score matching (PSM)—for estimating the impact of a program change in an education context. To do so, we compare estimates of impact obtained using this method to impact estimates obtained from a random assignment experimental study of that program change. A second objective is to explore different criteria according to which a nonexperimental estimation of impact might be considered "close enough" to what the experimental estimate would have been.

We believe it is helpful to think about the issues addressed in this study in the context of a decision maker who has two levels of decisions to make. At the ultimate level, she must decide whether to put substantial resources into a particular type of program (for example, a new reading curriculum, professional development of teachers, or reduced class size). In order to make this decision intelligently, she seeks an estimate of the likely impact of the program on an outcome of interest. This leads to the first level of decision to be made: what method of evaluation to use to obtain a reliable estimate of the likely impact of the program on the outcome of interest. This

Journal of Policy Analysis and Management, Vol. 26, No. 3, 455–477 (2007)

Published by Wiley Periodicals, Inc. Published online in Wiley InterScience (www.interscience.wiley.com)
DOI: 10.1002/pam.20262

paper is devoted to assessing a particular method of estimating impacts, so we concentrate on this first level of decision making. Toward the end of the paper, however, we look at the ultimate decision level: given an evaluation outcome, the evaluator must decide whether to put substantial resources into a particular type of program.

The "gold standard" of policy evaluation approaches is the random assignment experimental evaluation. This approach uses randomly assigned treatment and control groups to produce a reliable counterfactual and estimate of the actual program impact. Experimental evaluations, however, may require additional resources for implementation and sometimes face resistance from the community. In addition, experimental evaluations require preparation and planning, and cannot be conducted post-hoc. Researchers, therefore, have long sought to develop nonexperimental methods for evaluating the impact of policy programs. Most of these approaches, however, have failed to match the rigor of experimental evaluations. Recently, a new approach, called propensity score matching, has gained attention as a possible solution to this problem. The primary goal of this study is to assess the success of PSM in measuring the impact of policies in the context of primary education.

The particular situation we focus on for our assessment is one of deciding whether to invest resources in reducing class size. The question the decision maker would (should) ask is whether the propensity score method will yield estimates that are "close enough" to the "true" impact of smaller class size to use in making the ultimate decision of whether to invest in a program of smaller class size. In an attempt to answer that question, we use as a measure of the true impact the estimate that is derived from experimental estimates of the impact of reduced class size on test scores—that is, estimates in which subjects are randomly assigned to small classes (the treatment group) or to regular size classes (the control group). The outcome of interest is a composite reading and math test score. The difference in the mean test scores at the end of a year for the treatment and the mean test scores of the comparison group is the nonexperimental estimate of the impact of the smaller class size. To assess the reliability of the propensity score matched (PSM) estimates of impact, we compare these estimates to the impact estimates derived from the random assignment experiment, using several different methods. We try to determine "how close is close enough," putting greatest emphasis on the question: Would the nonexperimental estimate have led to the wrong decision when compared to the experimental estimate of the impact of the program? We find that propensity score methods perform poorly with respect to measuring the impact of a reduction in class size on achievement test scores.

Three considerations motivated us to undertake this particular study. First, there is a small, but important, set of studies, briefly described here, which have attempted similar assessments of nonexperimental methods of estimating the impact of a program, but these have all focused on earnings, employment, or welfare utilization as the outcomes of interest. We wanted to see to what extent the conclusions from those studies would extend into the area of education in which outcomes, such as test scores, were the focus of interest. Second, in recent years there had been a surge of interest in one particular type of nonexperimental method used for assessing the impact of interventions on outcomes: namely, propensity score matching (PSM).[1] Finally, there was a dataset available that was particularly well suited for this type of study. Tennessee's Student Teachers Achievement Ratio Project (Project STAR) is a

[1] Such recent articles include Diaz and Handa (2006), Dehejia and Wahba (1999, 2002), Smith and Todd (2005), and Agodini and Dynarski (2004).

class size reduction experiment—one in which 79 schools were involved, each of which, for our purposes, can be treated as a separate experiment. Project STAR is described in more detail in a later section.

In the next section, we sketch out the basic framework that has been used to assess the performance of nonexperimental estimates of program impacts. In our third section, we describe the Project STAR experiment from which we draw the data used in this study. The fourth section is devoted to a brief description of how we developed the propensity scores and how they are used to create comparison group samples matched to the treatment group for a given school. In our fifth section, using the PSM-created comparison groups, we describe the results of our regression models. For each school, we first estimate the experimental impact on test scores of small class size using the difference in mean test scores between the experimental treatment and the experimental control groups. Second, we estimate nonexperimental impact estimates, the difference in mean test scores between the experimental treatment group and the PSM-created comparison group. Third, we estimate the difference in these two impact estimates, experimental and nonexperimental, and determine whether these estimates are statistically significantly different. In our next section, we begin our discussion of various ways in which one might decide "how close is close enough," that is, when are the nonexperimental estimates of impact close enough to the experimental estimates that one might decide they would provide a reasonable substitute for a full experimental evaluation design. In the next section, we further discuss one criterion we developed in the previous section, which we call the "wrong decision" criterion. The second to last section is devoted to a discussion of some of the weaknesses of this study and the final section provides some conclusions.

BASIC FRAMEWORK

The framework we use follows that of several previous studies that have tried to evaluate the performance of nonexperimental methods for estimating program impacts by using data from random assignment experiments. The experimental estimates are used as a benchmark against which to compare the nonexperimental estimates of the impact of the program.

This basic approach was first used by Fraker and Maynard (1987) and LaLonde (1986), who took estimates based on the National Supported Work Demonstration (a random assignment experiment involving an employment program) as experimental "true" impact estimates, and compared these experimental "true" impact estimates to nonexperimental estimates constructed using various types of comparison groups. Dehejia and Wahba (1999, 2002) revisited the National Supported Work Demonstration data and, using PSM as the nonexperimental method, appeared to get much closer to the experimental estimates than had Fraker and Maynard and LaLonde. However, Smith and Todd (2005) showed that Dehejia and Wahba's results were very sensitive to the selection of a particular subsample from the National Supported Work Demonstration data.[2]

Others who have used this approach to evaluate the performance of nonexperimental impact estimation methods include Friedlander and Robins (1995); Heckman and Hotz (1989); Heckman, Ichamura, and Todd (1997); Heckman, Ichamura, Smith, and Todd (1996); Heckman, Ichamura, Smith, and Todd (1998); and Diaz and Handa (2006). More recently, a special issue of the Review of Economics and

[2] More discussion of Smith and Todd's findings can be found below.

Statistics (86[1], 2004) devoted several articles to the subject of evaluating nonexperimental methods of estimating program impacts.[3]

Existing literature evaluating nonexperimental methods, however, largely focuses on welfare to work and employment and job training programs, without questioning whether these results generalize to other policy areas, including education.[4] In contrast to this literature, we assess propensity score methods in the context of an educational intervention.

PROJECT STAR

We use Project STAR, the Tennessee class size experiment, as a source of true random assignment data. In 1985, kindergarteners within several Tennessee schools were randomly assigned to participate in an intervention (small classes of 13–17 students) or to not take part in the intervention (regular or regular with aide classes of 22–25 students). The schools chosen for the experiment were broadly distributed throughout Tennessee.

An analysis by an internal team of STAR researchers determined that students in small classes, on average, performed better than students in regular and regular with aide classes on standardized achievement tests (Word et al., 1990). In OLS estimates, which take into account teacher, student, and school characteristics, Krueger (1999) finds that students in small classes score five to seven percentage points higher than those assigned to regular size classes; this suggests, on average, that there was a positive impact from smaller classes on student achievement within individual schools, as we find in our experimental estimates. Following work by Krueger, which finds no significant difference in outcomes between regular and regular with aide classes in Project STAR schools, we combine regular and regular with aide students in this analysis (Krueger, 1999). Given that the randomization of teachers and students occurred within school, the estimated within-school effect of small classes should be unbiased. Because there were several deviations from the original experimental design following kindergarten (see Krueger, 1999), we use only data from kindergarteners.

The Project STAR data set provides measures of a number of student, teacher, and school characteristics, which are used to construct the propensity scores. The following variables are available as measures prior to random assignment: student sex, student race, student birth year and quarter, student free lunch status, teacher race, teacher education (bachelors, masters, masters plus, or specialist), teacher career ladder level (level 1, level 2, level 3, apprentice, probationary, or pending), teacher experience (years of teacher experience), and community type (rural, urban, inner city, or suburban). Other geographic characteristics are not available for the students or schools.

We use test score as the outcome variable. For all comparisons, we calculate test score as the average percentile rank within the sample distribution of the raw Stanford Achievement reading and math scores. We assign a respondent missing

[3] Also of note is work by Glazerman, Levy, and Myers (2003). They analyze a group of studies in which nonexperimental estimates of programs on earnings were assessed relative to experimental estimates.

[4] As we were developing this study, another group was developing a similar study of PSM with education outcomes as the focus (Agodini & Dynarski, 2004). Their study deals with outcomes derived from dropout prevention programs. We had extensive discussions with those authors as we were facing similar issues (Agodini & Dynarski, 2004). The recent article by Diaz and Handa (2006) also addresses some education outcomes.

Table 1. Descriptive statistics for all Project STAR schools and schools selected for propensity score evaluation.

Project STAR School ID Number	Students in Small Classes (*n*)	Students in Regular and Regular with Aide Classes (*n*)	Proportion of White Students	Proportion of Female Students	Proportion of Free Lunch Students
All STAR schools	1763	4111	.67	.49	.48
7	30	87	.88	.44	.40
9	38	82	.97	.47	.32
16	33	73	.00	.45	.98
22	28	103	.00	.52	.92
27	24	112	.00	.55	.87
28	56	75	.00	.49	.86
32	28	90	.00	.49	.97
33	28	75	.00	.52	.98
51	49	89	.86	.44	.15
63	29	83	.88	.48	.33
72	24	84	.96	.46	.60

either score the percentile rank of the subject for which a score exists. We exclude respondents without either score from the analysis.

We restrict our analysis to the Project STAR schools with more than 100 kindergarteners.[5] This criterion was selected because each school is being treated as a separate experiment, and this sample size was likely to provide sufficient statistical power. These schools were also selected because they are representative in size of typical policy interventions. The schools that were selected had, on average, a higher concentration of black students and a higher concentration of students receiving free lunches, were more likely to be located in the inner city, and were more likely to have non-white teachers than schools in the overall sample. The selected schools included schools from all four community types. These schools were located in six different school systems.

Table 1 lists the Project STAR school ID number for each of the selected schools, along with the number of students in small classes, the number of students in regular or regular with aide classes, the proportion of students who are white, the proportion of students who are female, and the proportion of free lunch students for these schools as well as for the overall Project STAR sample.

THE PROPENSITY SCORE METHOD

A propensity score is a conditional probability of participation in a treatment. For individuals who were geographically not eligible to participate in a treatment, for example, a propensity score provides a measure of how likely they were to have participated, had they been eligible for the program based on the characteristics of

[5] Although even 100 kindergarteners may seem to provide a small sample, as will be seen, these samples provided sufficient statistical power to detect statistically significant impacts for most of the schools. Further, we note that these sample sizes are not exceptional in educational evaluations. In fact, in their study of the impact of a computer-based program to improve reading skills, Rouse and Krueger (2004) measured four different types of test score outcomes with the final sample size for each outcome varying from 86 to 463, with an average of 127 participants.

those who were selected to participate. In this way, nonparticipants and participants with similar characteristics, as summarized by the propensity score, can be matched to create a comparison group. Propensity scores primarily simplify matching by reducing the dimensionality of the matching problem.

The recent work by Dehejia and Wahba (1999, 2002) gave rise to considerable interest in the potential for using PSM as a means of obtaining better nonexperimental impact estimates. These authors, like Fraker and Maynard (1987) and LaLonde (1986) use 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. They use the propensity score method to focus attention on a small subset of the comparison units most comparable in observable characteristics to the treated units, hence alleviating the bias due to systematic differences between the treated and comparison units. Dehejia and Wahba (1999, 2002) found that using propensity score methods, they could reasonably replicate experimental impact estimates.[6]

In the last few years, a number of computer programs for propensity score estimation have become available that make these methods much easier to implement. Because of this, and the interest generated by the initial Dehejia and Wahba articles, policymakers and practitioners became excited about the potential use of propensity scores to produce comparison groups and thereby provide a credible nonexperimental evaluation as an alternative to a full experimental design evaluation study.

IMPLEMENTATION OF THE PROPENSITY SCORE METHOD USING PROJECT STAR

Within each selected Project STAR school, the students in the regular or regular with aide classes provide the control group for each group of program participants. These students show what would have happened to the treatment students, had they not been given the opportunity to experience the intervention. To estimate the "true" impact of the program, we compare the test scores of students in small classes to those students in the *same* school who were in regular or regular with aide classes.

In order to implement tests of nonexperimental estimates of the Project STAR impacts on achievement test scores, using propensity score matching, we first construct for each of the 11 schools the nonexperimental comparison group using the control groups of *other* schools as the pool of potential nonexperimental control group members.[7] This approach replicates the situation that might occur in a nonexperimental evaluation when an entire school chooses to implement smaller class

[6] However, Smith and Todd (2005), in re-examining the National Support Work Demonstration data used by Dehejia and Wahba (1999, 2002), found that these estimates were "highly sensitive to both the set of variables included in the scores and the particular analysis sample used in estimation" (p. 305). They conclude: "Our evidence leads us to question recent claims in the literature by DW (1999, 2002) and others regarding the general effectiveness of matching estimators relative to more traditional econometric methods" (p. 347). See Journal of Econometrics, 125(1–2), pp. 355–375, for a full discussion, including a reply to Smith and Todd by Dehejia and a rejoinder by Smith and Todd.

[7] In many of the previous studies cited, the dataset from which the comparison group pool is constructed is entirely different from the experimental sample (although some of the studies did use controls in other sites within the same experiment; for example, Friedlander and Robins, 1995). We used the controls at other sites because all the variables, outcomes, and covariates are measured in precisely the same way as they are for the treatment group. This is particularly important for the outcome variable, achievement test scores. In a recent article on PSM, the importance of comparability in measurement was strongly emphasized: "We find that PSM performs well for outcomes that are measured comparably across survey instruments and when a rich set of control variables is available. However, even small differences in the way outcomes are measured can lead to bias in the technique" (Diaz & Handa, 2006, p. 319).

sizes. One would look to students in other schools as the comparison group. In constructing this comparison group, one would want to find children at other schools who match as closely as possible the characteristics of the children and teachers in the "treatment" classrooms. We use propensity score methods to identify the comparison observations empirically most similar to the program (assigned to small class) participants in that school.

To estimate the propensity score, we first regressed an indicator for small class participation on a vector of covariates. This regression identified any characteristics that might differentiate the students in the treatment classrooms at the school from the general population. For the purposes of being clear about the sample used in estimating the individual propensity scores, we write the equation as a linear probability model:

$$Y_i = \boldsymbol{B_j}\, \mathbf{X}_{ji} + \varepsilon_i$$

Where:

- i = 1… m for members of the small class group in the target school and m + 1… n for members of regular size classes in the other Project STAR schools.
- Y_i = 1 for those in small classes in the target school and 0 for those in regular classes in the other Project STAR schools.
- X_{ji} is a vector of j covariates.

In fact, the propensity score equation for each school was estimated as a logit regression using maximum likelihood with the n sample members as just described.[8]

The propensity score for each individual in the sample is then generated using that individual's values for the $\mathbf{X}_{ji}$ and the estimated $\boldsymbol{B_j}$ coefficients. Such propensity scores are estimated separately for each of the target schools in the sample. The sample for each is made up of the 1… m members of small classes in that school and the m + 1 … n members of regular size classes (control groups) from the other Project STAR schools. With these propensity score estimates for each individual, the remainder of the PSM procedure is carried out to develop a comparison group for the given school.

The remainder of the PSM procedure is designed to create a comparison group that best mirrors the full range of characteristics present in the treatment group, as well as finding the best matching control for each individual within the treatment group. First, the pool of potential comparison observations is restricted to those with propensity scores above the minimum and below the maximum propensity scores for those in the treatment group in the selected school. Then the procedure "balances" the distribution of propensity scores between the treatment (small class) group and the potential pool of comparisons.

To test the balancing criterion, the distribution of propensity scores, ranked from highest to lowest, is divided into a series of bins. A + test is performed to test whether within each bin the mean propensity score value for the treatment group members is significantly different from the mean propensity score for the comparison

[8] The log likelihood function that is maximized is $£(B_j) = \Sigma Y_i \log(G(\mathbf{X}_{ji} B_j)) + (1 - Y_i) \log\{1 - G(\mathbf{X}_{ji} B_j)\}$ where G is the cumulative distribution function for a standard logistic variable. The predicted probability for each observation is generated by inserting the individual values for the $\mathbf{X}_{ji}$ times the estimated B_j into the expression $Y_i \log\{G(\mathbf{X}_{ji} B_j)\} + \log\{1 - G(\mathbf{X}_{ji} B_j)\}$ and exponentiating. In practice, we estimate the log likelihood function using STATA's *logit* command and obtain the estimated probabilities of $Y_i = 1$ (which, in this case, are the estimated propensity score values for each individual) using STATA's *predict* command.

group pool. If there are significant differences within bins, the boundaries of the bins are shifted until within each bin there are no significant differences between treatments and controls in propensity score values. Five percent was chosen as the level of significance for the balancing criterion.

Once the first propensity score balance test is passed, the next test is to determine whether all the observed characteristics are jointly insignificantly different between treatment group members and the pool of potential comparisons within each bin. To determine this, a Hotelling test is performed for each bin. This test is very similar to an F test for joint significance.[9] In our analysis, the particular means tested by the Hotelling test depended on the particular school being analyzed. For each school, a subset of the following covariates was tested: student sex, student race, student birth year and quarter, student free lunch status, teacher race, teacher education, teacher career ladder level, teacher experience, and community type. For cases where there were significant differences in covariates (differences at the 5 percent level of significance), the original covariates were adjusted, the logistic regression re-estimated, and the procedure repeated.[10,11]

Once the balance criterion is satisfied, the program participants are matched to the observation from the potential comparison pool (of controls from all of the other schools) that is closest in absolute value in propensity score. This study used matching with replacement methods (that is, a comparison observation from the pool that was matched to one treatment observation was eligible to be matched again to another treatment observation). Therefore, in the calculations of means and in the regression analysis, each comparison observation is weighted in accordance with its number of matches to treatment participants.[12] The set of matches for each of the children in small classes, selected using PSM methods from the entire set of all kindergarteners in regular and regular with aide classes outside of the selected school, formed the nonexperimental comparison group.[13] We repeated this process separately for each of the target schools in the sample.

Table 2 provides one way to assess the similarity of the control and comparison groups. For each selected Project STAR school, the table provides the probability values for the Hotelling test of the joint mean between the observed characteristics of the treatment and experimental control group. Table 2 also shows the probability value of the same test between the observed characteristics of the treatment and the nonexperimental constructed comparison group. In none of the 22 cases are the comparison or control groups significantly different from the treatment group.

[9] The Hotelling t^2 tests the hypothesis that the sample mean vectors $\mathbf{x}_1$ and $\mathbf{x}_2$, from two samples, assuming common population covariance matrices, are equal. That is, for the variables, $X_1, X_2, X_3, \ldots X_p$, the sample mean vectors include the mean from each sample for each variable. In practice, Hotelling's t^2 statistic is calculated as: $T^2 = n_1 n_2 (\mathbf{x}_1 - \mathbf{x}_2)' \mathbf{C}^{-1} (\mathbf{x}_1 - \mathbf{x}_2)/(n_1 + n_2)$, where n_1 and n_2 are the number of observations for each sample (1 and 2), $\mathbf{C}$ is the pooled variance covariance matrix of $\mathbf{x}_1$ and $\mathbf{x}_2$, and the transformed statistic, $F = (n_1 + n_2 - p - 1)\, T^2/[(n_1 + n_2 - 2)p]$, is distributed according to the F distribution with p and $(n_1 + n_2 - p - 1)$ degrees of freedom (Manly, 1986, p. 28).

[10] In a few cases with five or fewer observations within a bin, the test of joint means could not be carried out.

[11] STATA automatically dropped some covariates due to collinearity. Appendix A details the final bin sizes for each propensity score estimating procedure. (All appendices are available at the end of this article as it appears in JPAM online. Go to publisher's website and use search engine to locate article at: http://www3.interscience.wiley.com/cgi-bin/jhome/34787).

[12] The results were not qualitatively affected when matching without replacement was used.

[13] An alternative procedure would have involved matching to control group members instead of program participants. However, in a real nonexperimental evaluation, there would be no experimental control group available to which to match. Therefore, the appropriate way of testing propensity score methods is by matching to the treatment group.

Table 2. Similarity of covariates between treatment and control and comparison groups in 11 Project STAR schools, using Hotelling test of joint means.

Project STAR School ID Number	P Value of Test of Characteristics: Treatment versus Experimental Control*	P Value of Test of Characteristics: Treatment versus Nonexperimental Control*
7	.81	1.0
9	.61	1.0
16	.22	1.0
22	.45	1.0
27	.93	1.0
28	.22	.98
32	.50	1.0
33	.85	1.0
51	.20	.99
63	.45	1.0
72	.17	1.0

*Hotelling t test

Table 3 shows, for each selected STAR school, the Project STAR School ID number and the proportion of students in the treatment, control, and constructed comparison groups who are white, female, and of free lunch status. As is clear in Table 3, there are several schools (for example, 27 and 72) for which the treatment and comparison groups have identical descriptive statistics, suggesting that the propensity score method and matching mechanism worked to choose observations that were similar in observable covariates to the treatment observations.

It is interesting to note from this table that in almost every case, the average comparison group characteristics are closer to the treatment group (small class) than are the control group members for that school. As we will suggest later, this gives even greater support to the view that the weaknesses of nonexperimental estimates have primarily to do with the impact of unobserved or unobservable characteristics.[14]

RESULTS

We first obtain estimates for each of the 11 schools of the experimental impact (which we refer to as the "true" impact).[15] These are calculated as the difference in the mean combined reading and math test score percentiles between those in small (treatment group) and regular classes (control group), controlling for sex, race, and free lunch status. This is the standard against which the estimates of impact obtained from the PSM are judged below.

[14] It is not so surprising that the propensity score matched samples are closer in observable characteristics to the experimental treatments than are the true controls. The problem is that those dimensions upon which the nonexperimental samples are more similar are not necessarily the dimensions that reliably predict outcomes.

[15] We recognized that because of sampling variability, the estimated impact will deviate, in repeated samples, from the mean impact for this school.

Table 3. Project STAR descriptive statistics for kindergarten students in treatment, control, and comparison groups by school ID.

	Proportion White			Proportion Female			Proportion Free Lunch Status		
Project STAR School ID Number	Students in Small Classes	Students in Control Groups (Regular Classes)	Students in Comparison Groups	Students in Small Classes	Students in Control Groups (Regular Classes)	Students in Comparison Groups	Students in Small Classes	Students in Control Groups (Regular Classes)	Students in Comparison Groups
7	.93	.86	.93	.33	.47	.33	.37	.41	.37
9	.97	.96	.97	.42	.49	.42	.37	.30	.37
16	.00	.00	.00	.58	.40	.58	1.00	.97	1.00
22	.00	.00	.00	.46	.53	.46	.82	.94	.82
27	.00	.00	.00	.50	.56	.50	.93	.86	.92
28	.00	.00	.00	.50	.48	.54	.76	.93	.73
32	.00	.00	.00	.46	.50	.46	.96	.98	.96
33	.00	.00	.00	.54	.52	.52	.96	.99	.99
51	.82	.89	.82	.49	.42	.41	.08	.18	.06
63	.93	.86	.93	.62	.43	.62	.28	.35	.35
72	.92	.98	.92	.46	.46	.46	.50	.63	.50

Journal of Policy Analysis and Management DOI: 10.1002/pam
Published on behalf of the Association for Public Policy Analysis and Management

The general equation used to estimate the impact of smaller class on achievement scores is of the following form, where standard errors are robust and clustered at the classroom level:

$$Z^q_{ik} = B^q_j X^q_{jik} + B^q_t T^q_{ik} + \varepsilon^q_k + \theta^q_{ik}$$

Where q indexes the school for which the equation is estimated;[16]

i indexes individuals;
j indexes covariates;
k indexes groups (classrooms).
Z^q_{ik} is the percentile rank on the combined reading and math test score for individual i in group k.
$\mathbf{X}^q_{jik}$ is a set of j covariates, including sex, race, and free lunch status, for individual i in group (classroom) k.[17]
T^q_{ik} is a treatment indicator, indicating whether individual i in group k was in a small class ($T_{ik} = 1$) or regular size class ($T_{ik} = 0$).
ε^q_k is an error that is iid between groups (classrooms) with mean 0 and variance ρ^2.
θ^q_{ik} is an error that is iid between individuals within groups with mean 0 and variance ϕ^2.

For the experimental estimates for each school of the impact of the treatment (smaller class size), B^q_t, the sample includes all the students for that school randomly assigned to the treatment group (small size classes) or the control group (regular size classes).[18]

Next, for the nonexperimental estimates, we use the equivalent model, but instead of including the control group (regular size class) members in the *same* school, we use the propensity score matched comparison group members developed for that school. For this sample, there is one comparison group member (drawn from the controls for schools other than q) for each member of the treatment group in school q. Then the definition for T^q_{ik} is a treatment indicator, indicating whether individual i in group (classroom) k was in the small class in school q ($T^q_{ik} = 1$) or in the propensity score matched comparison group from a regular size class ($T^q_{ik} = 0$).[19]

The nonexperimental estimate of the impact on test scores of being in a small class is calculated as 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, while controlling for demographic characteristics.

Mean percentile test score ranks for the treatment group, experimental control group, and then the comparison group for each selected Project STAR school are

[16] We include the q superscript to emphasize that each equation is estimated separately for each of the 11 schools.
[17] These covariates are included to increase the precision of the program effect estimates.
[18] We estimated the equations using the SAS procedure Gen Mod, allowing for clustering at the classroom level.
[19] Suppose for the treatment group in school q there were m individuals. Then, in the nonexperimental estimating equation, there will be 2 m individuals, m in the treatment group and m in the comparison group, one match for each treatment group member. For the experimental estimating equation, where in school q there are m individuals in the treatment group, there may be more than m individuals in the control group; first, because the regular size classes are larger than the treatment small size classes; and second, because there may be more than one regular size class in school q.

Table 4. Tennessee Project STAR mean percentile combined math and reading scores within school for students in treatment, experimental control, and nonexperimental PSM comparison group.

1	2	3	4
Project STAR School ID Number	Mean Percentile Combined Test Score Small Class	Mean Percentile Combined Test Score Control Group (Regular and Regular with Aide Classes)	Mean Percentile Combined Test Score Nonexperimental PSM Comparison Group
7	62.02	65.35	50.34
	(32.50)	(27.60)	(19.51)
9	70.04	60.20	48.07
	(21.79)	(24.61)	(23.71)
16	43.66	20.16	61.00
	(25.28)	(15.51)	(27.83)
22	68.94	44.73	55.07
	(29.30)	(24.33)	(29.85)
27	58.78	69.54	25.57
	(20.07)	(23.32)	(15.23)
28	44.53	41.71	51.37
	(29.41)	(24.46)	(29.43)
32	35.14	23.19	57.02
	(24.38)	(19.19)	(33.92)
33	40.46	29.08	60.26
	(16.30)	(20.35)	(29.47)
51	74.23	60.04	59.32
	(20.64)	(24.29)	(24.01)
63	80.25	62.08	50.67
	(17.77)	(25.66)	(23.96)
72	78.92	57.62	45.24
	(22.70)	(27.85)	(20.73)

Standard deviations in parentheses.

reported in Table 4. Each row contains the results for one of the 11 selected schools. Column 2 provides the mean percentile rank of math and reading scores for students in small classes in the selected school. The third column shows the mean percentile rank of math and reading scores for students in regular classes in the selected school. The fourth column shows the mean percentile rank of math and reading scores for students in the comparison group, constructed using propensity scores. Standard deviations are in parentheses.

Table 5 provides the major results for this study. It gives a summary of the estimates of the impact of smaller classes on combined reading and math test scores, both for the experimental estimates and for the nonexperimental (propensity score matched) estimates. The dependent variable in the regressions that underlie the impact estimates in the table is the percentile test score, as defined earlier.[20] The independent variables include the demographic characteristics of the sample member prior to random assignment (sex, free lunch status, and race). The estimated coefficients for these variables are not shown.

[20] We also ran regressions analogous to those in Table 5, using normal curve equivalents as the dependent variable. The results were unaffected.

Table 5. Project STAR regression adjusted estimates of program effect using experimental controls and nonexperimental comparison groups.

1	2	3	4	5
Project STAR School ID Number	Regression Adjusted Estimate of Program Effect with Experimental Controls	Regression Adjusted Estimate of Program Effect with Nonexperimental Comparisons	Are the Effects Opposite in Sign?	Is the Difference Between Experimental and Nonexperimental Estimates Significant?
7	−4.59	11.63		
	(7.07)	(6.86)	Yes	Yes
9	11.89	21.97		
	(3.87)*	(5.54)*	No	No
16	22.73	−17.35		
	(4.56)*	(11.23)	Yes	Yes
22	24.14	13.87		
	(10.06)*	(13.32)	No	No
27	−10.01	33.20		
	(8.07)	(5.68)*	Yes	Yes
28	0.79	−6.01		
	(8.87)	(12.38)	Yes	No
32	12.10	−21.89		
	(5.31)*	(15.26)	Yes	Yes
33	11.53	−19.80		
	(6.82)	(11.64)	Yes	Yes
51	14.00	15.20		
	(4.36)*	(2.53)*	No	No
63	15.17	29.59		
	(6.75)*	(5.68)*	No	No
72	18.50	33.69		
	(4.98)*	(5.87)*	No	Yes
Across schools weighted average	12.72	17.79		
	(1.70)*	(1.73)*	No	Yes

*Robust standard errors clustered at the classroom level in parentheses. Asterisk indicates significance at the 5 percent level. See Appendix B for discussion of the standard errors. (All appendices are available at the end of this article as it appears in JPAM online. Go to publisher's website and use search engine to locate article at: http://www3.interscience.wiley.com/cgi-bin/jhome/34787).

The impact estimate is the coefficient B_t^q on the dummy variable T_{ik}^q which is 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.

Again, each row in Table 5 shows the results for one of the 11 schools.[21] Column 1 provides an identification number for each school. Column 2 gives the estimate of

[21] See Appendix B, where there is a discussion of the serious technical issues involved in estimating standard errors for these regressions. (All appendices are available at the end of this article as it appears in JPAM online. Go to publisher's website and use search engine to locate article at: http://www3.interscience.wiley.com/cgi-bin/jhome/34787).

the effect on percentile test scores of small class size obtained using the random assignment experimental and control sample, controlling for race, sex, and free lunch status. Column 3 provides the estimate of the difference in mean test scores between those randomly assigned to be in small classes in the given school 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, after controlling for race, sex, and free lunch status. Column 4 shows whether the signs of 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 3 are different. Column 5 shows whether, for the given school, the experimental impact estimates and nonexperimental impact estimates are statistically significantly different.[22] The significance of the difference between the experimental and nonexperimental control group was assessed by comparing the mean of the experimental control group with the mean of the comparison group, controlling for demographic characteristics, at the 5 percent level of significance. For all values, standard errors are in parentheses, and significant results, as implied by the 5 percent threshold, are indicated with an asterisk.

How well do the propensity score matched estimates of the impact of class size approximate the experimental impact estimates? Table 5 gives us the opportunity to look at 11 different cases for which there are two sets of estimates of the impact of smaller class size on combined reading and math test scores, one generated by random assignment of students and teachers to a treatment group or control group and the other using a comparison group created by PSM.

Looking first at the experimental estimates, we can see that they vary considerably across the schools. For 7 of the 11 schools, the impact on the test scores of the smaller class size is positive and statistically significant. The impact estimates range from minus 10 percentile points to plus 24.[23] Neither of the two cases where the estimate of the impact was negative in sign was statistically significantly different from zero.

For the nonexperimental (propensity score matched) estimates, there is also substantial variability across schools. For 5 of the 11 schools, the impact on test scores of the smaller class size is positive and statistically significant. The impact estimates range from minus 22 to plus 33 percentile points. None of the four cases where the estimate of the impact was negative in sign were statistically significantly different from zero.

Now we move to a discussion of the differences between the experimental and nonexperimental impact estimates. The magnitudes and signs of the differences between the experimental and the nonexperimental impacts are, in most cases, substantial. For only two of the schools (28 and 51) are the experimental and nonexperimental

[22] The difference between the two impact estimates is actually tested by testing the difference in the mean test scores between the experimental control group and the comparison group. This can be shown algebraically to be equivalent to a test for the significance of the difference between the experimental impact estimate and the nonexperimental impact estimate.

[23] In his evaluation of the Project STAR experiment, Alan Krueger noted this extensive variation across schools. For each school, he pooled the data across all the grade levels (kindergarten, 1st, 2nd, and 3rd grades) and then generated school-level impacts of smaller class size on the test scores. He reports: "Two-thirds of the school-specific small-class effects are positive, while one-third are negative. Furthermore, 2.5 percent of the 80 coefficients had t-ratios less than –2, while 30 percent had t-ratios exceeding +2. … Thus, some schools are more adept at translating smaller classes into student achievement than are other schools" (Krueger, 1999, pp. 524–526).

impacts less than 10 percentile points apart, and for one of those (school 28), neither impact estimate was statistically significantly different from zero. On the basis of such a casual examination, we would conclude that the nonexperimental propensity score matched estimates are not likely to give a reliable estimate of the "true" impact of smaller class sizes (of the magnitude in Project STAR) on achievement test scores.

We feel a need to push harder, however, in trying to reach a judgment about the degree of difference between these two types of estimates. First, in column 4 of Table 5, we noted whether the signs of the impact estimates for the experimental and the nonexperimental estimates were different. In 6 of the 11 cases, they are different. However, when one of the estimates is not significantly different from zero, this test cannot be taken as a strong indicator of misleading estimates. Second, in column 5 of Table 5, we test whether the impact estimates themselves are statistically significantly different; that is, is the experimental impact estimate statistically significant different from that of the nonexperimental (propensity score matched) estimate? The answer is that in 6 of the 11 estimates, there is a statistically significant difference between the two estimates of the impact of smaller class size on achievement test scores.

Averages of Impacts

We here address a couple of issues that might arise in examining these two types of estimates. First, it has been pointed out by some analysts that if experimental and nonexperimental impacts are averaged across many separate estimates, the averages seem to be closer to each other than the individual estimates are.[24] And, some argue, for many large evaluation projects, the intervention (treatment) is carried out at multiple sites, and the analysis is carried out by pooling data across sites and then estimating impacts, so such averaging of impacts seems reasonable.

Our basic stance is that we want to take advantage of the fact that Project STAR was an experiment carried out across many sites in exactly the same way, with variables measured in exactly the same fashion. In light of this, we can regard the individual schools as separate experiments and think of the variation across schools as an indication of the possible distribution of impact estimates in repeated samples. At least with respect to education outcomes, as far as we are aware, there has rarely, or never, been a chance to analyze multiple implementations of exactly the same treatment (in this case, smaller class size) and obtain multiple estimates of the impact of that treatment. Then, as we seek to compare experimental (random assignment) estimates of impacts to nonexperimental estimates (in this case, propensity score matched estimates), we have 11 replications.

However, we also provide some results in which we average both the experimental and the nonexperimental estimates of impact across the 11 sites. In the final row

[24] A complete discussion of these issues would take us far from our central theme. However, we cite the most thorough discussion to date ,of which we are aware, concerning these issues, which is in Glazerman, Levy and Myers (2002, pp. 83–85). They conclude: "The within-study evidence . . . suggests that the average bias (their term for the difference between experimental and nonexperimental estimates of impacts) across all methods, subgroups, and time periods is sometimes positive and sometimes negative and often still in hundreds of dollars [they were studying earnings impacts]. This suggests that a mechanistic application of a large number of NX (nonexperimental) estimators might improve the inference one could draw from such evidence, but not in a predictable way. Whether the average bias, properly weighted within and between studies, is really close enough to zero for policymakers, and whether the bias cancels out within a narrower domain of research are questions that we plan to address as more design replication studies are completed."

of Table 5, we present a set of averages across the 11 sites and their standard errors.[25] In the first column, the weighted average of the experimental impacts is 12.7 percentile points. In the second column, the weighted average for the nonexperimental (propensity score matched) impacts is 17.8 percentile points. The weighted average difference between the nonexperimental and experimental impacts is 5 percentile points and that difference is statistically significantly different from zero. These results are consistent with the results reported earlier; propensity score methods do not do very well, at least in this situation. The nonexperimental estimates of impact are not close to the experimental estimates of impact.

Other Assessments of Closeness

The question of how close is close enough requires a more careful answer than the gross generalization just stated. One way of assessing the differences in the impact estimates is to test whether the impact estimate from the nonexperimental procedure is significantly different from the impact estimate from the experimental procedure. As indicated in column 5 of Table 5, for 6 of the 11 schools, the two impact estimates were statistically significantly different from each other. These estimates suggest that propensity score methods are not likely to be reliable.

A second way to assess closeness is to see whether the experimental impact estimate is statistically significantly different from zero, but the nonexperimental impact estimate is not statistically significantly different from zero, or vice-versa. This occurs in two cases, schools 16 and 32, where the experimental estimates of impact were positive and statistically significant but the nonexperimental impact estimates were not statistically significantly different from zero.

A third possible criterion to use to evaluate the closeness of the estimates would be the percentage difference in the point estimates of the impact. For example, for school 16, the nonexperimental estimate of the impact is different in sign and 24 percent smaller in absolute value than the experimental impact. But for school 51, 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.

These results led to the conclusion that with these data the propensity score matching nonexperimental estimates do not approximate the experimental estimates of the impacts of small class size on test score performance well. As aforementioned with respect to Table 4, the propensity score matched comparison groups seem, by observable characteristics, to more closely resemble the treatment groups than do the actual control group for the given school; yet they do not produce treatment estimates that are close to the experimental estimates. *This strongly suggests that the effect of unobservable characteristics (which propensity score methods do not adequately address) on impacts is quite significant.*

[25] The reported averages are weighted averages and the standard deviations are the standard deviations of these weighted averages. The weights are the inverse variance weights. This type of weighting scheme is often used by those who do meta-analysis (see Lipsey & Wilson, 2000, pp. 129–133). The weights take into account both the sample size for each of the individual site estimates and the variances for each site.

HOW CLOSE IS CLOSE ENOUGH FOR DECISION MAKING?

Here we wish to return to the consideration of the decision context that we outlined in the introduction. Using this context we wish to make two somewhat distinct points.

For a decision maker who is considering the design of a quantitative evaluation, considering sample design and different methods of evaluating the impact of a potential program change, the criterion that requires first attention is the nature and size of the impact estimate that would be determinative for their decision. This study is devoted to the examination of an alternative method, but the first consideration that should drive both the method choice and the design of the sample is this: What is the threshold value of the impact that would make the decision positive, for example, to adopt the new curriculum or reduce class sizes? In the sample design stage, this threshold value should determine the minimum detectable effect for which the sample should be designed, a priori, to provide.[26]

Beyond constituting a plea for more rational evaluation design procedures, this leads to our second point about the decision making context. We suggest that to judge whether the nonexperimental estimating procedure yields estimates that are "close enough" to those which would be obtained from experimental estimates, we should ask whether the distance between the nonexperimental and the experimental impact estimates would have been sufficient to cause an observer to make a different decision from one based on the 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, for example, invest in the type of activity that 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.

A rough example of applying this sort of criterion can be demonstrated in the case from which the data in this study were drawn, the Project STAR experiment with class size reduction in Tennessee.

In articles presenting aspects of his research using the Project STAR data, Krueger (1999; Krueger & Hanushek, 2000) has developed some rough cost-benefit calculations related to reduction in class size. In Appendix C[27] a few elements of his calculations are described to provide the background for the summary measures derived from his calculations that we use to illustrate our "close enough." The benefits that Krueger focuses on are increases in future earnings associated with test score gains. He carefully develops estimates, based on other literature, of what increase in future

[26] The minimum detectable effect is the smallest true program effect that would have a probability at a given power level of producing an impact estimate that would be significantly different from zero at some specified significance level (for example, a 5 percent significance level with an 80 percent [power] probability).

[27] All appendices are available at the end of this article as it appears in JPAM online. Go to publisher's website and use search engine to locate article at: http://www3.interscience.wiley.com/cgi-bin/jhome/34787.

Table 6. Cost effectiveness decisions for Project STAR using a 5.4 percentile impact criterion.

1	2	3	4
Project STAR School ID Number	Would Decision Makers Invest in the Program Based on Experimental Estimate?	Would Decision Makers Invest Based on Nonexperimental Estimate?	Would the Nonexperimental Estimate Lead to the Wrong Decision?
7	No	No	No
9	Yes	Yes	No
16	Yes	No	Yes
22	Yes	No	Yes
27	No	Yes	Yes
28	No	No	No
32	Yes	No	Yes
33	No	No	No
51	Yes	Yes	No
63	Yes	Yes	No
72	Yes	Yes	No
School average	Yes	Yes	No

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 Project STAR experience and appropriately adjusted.

For this study's purposes, what is most interesting is the way he uses these cost-benefit 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 justify, right at the margin, 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.[28] 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.[29] Krueger indicates that critical effect size of 0.2 of a standard deviation in test scores translates into a 5.37 percentile increase in achievement test scores due to smaller class size (the measure used here for the impact estimates). 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.

Table 6 uses the results from Table 5 to apply the cost effectiveness criterion to determine the extent to which the nonexperimental estimates would, in the cases of

[28] Effect size has increasingly become the way in which impact outcomes are stated. Suppose an impact estimate (treatment group test score mean minus comparison group test score mean) is 15 points. If, for example, the standard deviation of the test score is 75 points, then the effect size is 15/75, which equals .2. That is, the impact of the program is .2 standard deviations. This procedure is useful for comparing impacts where the metric of the measure (say test scores) is different across studies and has been promoted for meta-analytic studies.

[29] Krueger gives several critical values, which vary with different key assumptions. For this illustration, 0.2 of a standard deviation is chosen.

these 11 schools, have led to the wrong decision. If the impact estimate is greater than 5.4 percentile points and statistically significantly different from 0, in Table 5, a yes is entered, inferring that 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, a no is entered, to indicate that the impact estimate would have led to the conclusion that the class size reduction was not cost effective. In Table 6, column 1 is the school number; column 2 provides the conclusion a policymaker would come to on the basis of the experimental impact estimate; column 3 gives the conclusion a policymaker would have made if relying only upon the nonexperimental impact estimate; and column 4 contains a yes if the nonexperimental estimate would have led to a "wrong" cost effectiveness conclusion, that is, a different conclusion from the experimental impact conclusion about cost effectiveness.

Table 6 shows that the nonexperimental estimate would have led to the wrong conclusion in 4 of the 11 cases. Overall, then, in 4 of the 11 cases considered, the nonexperimental estimates would have led to the wrong cost effectiveness conclusion. Anyone can decide what to make of this indication of the likelihood that the nonexperimental estimates are "close enough." But our own conclusion is that it would be too risky to rely on the propensity score matching estimates in this situation.[30]

WOULD THE WRONG DECISION CRITERION BE OPERATIONALLY USEFUL IN DECIDING 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 this question: Would the wrong decision criterion be operationally useful in deciding on whether to use a nonexperimental method, and if so, which nonexperimental method? 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 cost-benefit, 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 nonexperimental estimate could be expected to be (the bias of the nonexperimental estimate) and in which direction the nonexperimental impact might be in a given case. Looking at the selected schools, there is tremendous variation in how close this particular nonexperimental estimator was to the experimental. Unless enough studies similar to this one are accumulated so that we have reasonable grounds for judging how far off the nonexperimental estimates are likely to be in a

[30] Additional estimates from other nonexperimental methods (including using schools of similar community [or district] types as comparison groups; using schools of different community types or schools from districts of different types as comparison groups; using all controls from all other schools rather than selecting among the pool to match; or a random selection of controls from all other schools or schools of the same type) are available upon request from the authors. None of these alternatives altered qualitatively the conclusions previously presented.

given situation, this wrong decision criterion cannot be operationally implemented. Initially, when evaluating propensity score methods, our hope was that the direction of the bias in the method would be clear (and also the magnitude of the bias), such that this bias could be predicted and corrected for. Unfortunately, as is clear in Table 5, neither the direction nor the magnitude of the bias is consistent amongst the 11 example estimates provided by this Project STAR sample.

WEAKNESSES OF THIS STUDY AS AN ASSESSMENT OF PSM AND THEIR RELEVANCE TO AN EVALUATION DESIGN DECISION CONTEXT

The limitations of these data as a means of assessing the performance of PSM should be explicitly recognized.

The Project STAR data provide an unusual opportunity because they come from a random assignment intervention and they represent not just one experiment, but as many experiments as there were schools. In addition, the sample sizes are reasonable even at the individual school level. As has been emphasized here, all the variables are defined and measured the same way across all the sites. Thus, this study was able to use treatment groups from one school, and then the controls from another set of schools, as the pool from which a nonexperimental comparison group could be constructed. This provides, as stressed previously, a test that is *more favorable to finding propensity score methods are effective than other studies* that rely on data drawn from different sources, with different variable definitions.

For the purposes of testing PSM and some other nonexperimental methods, these data have some weaknesses. The Project STAR data has only a limited number of covariates that can be used in the matching process.[31] Of course, it is not necessarily the case that having more covariates would improve the matching process. What matters is whether the number of covariates suffice to satisfy the conditional independence assumption; that is, that the observed covariates are the covariates that matter for predicting the outcome.

It seems probable, however, that there are some covariates that would be likely to influence outcomes but are unavailable in the Project STAR data. In particular, the Project STAR data have no pre-intervention measures for the outcome variable.[32] That is, there are no achievement test scores for students before they were randomly assigned to small classes or to regular classes. For many nonexperimental estimators, such pre-intervention measures might improve the matching and can be used for pre-post (or difference-in-differences) estimations of impact. It would be nice to be able to test directly the degree to which having pretest achievement test scores would significantly improve the propensity score match. It should not be presumed, however, that pretest measures will be a panacea for propensity score matching problems. In two studies carried out at about the same time as this one, extensive pre-intervention outcome measures were available, and the performance of the propensity score match impact estimates compared to the experimental impact measures was equivalent to ours. One study (Michalopolous, Bloom, & Hill, 2004) looked at welfare-to-work programs and had as principle outcome variables employment and welfare

[31] As described earlier, for the matching model, student age, student sex, student race, student free-lunch status, teacher race, teacher education, teacher career ladder, teacher experience, and community type were available.

[32] The lack of a pretest score was true for all the grades in Project STAR, not only for the kindergarten grades. This is not a serious problem for experimental estimates that arise from a random assignment process with sufficient sample size, because that process leads with high probability to equivalence between treatment and control groups on average for all variables, observed and unobserved.

utilization. For both outcomes, they had pre-random assignment measures for as long as three years. The other study (Agodini & Dynarski, 2004) looked at the School Dropout Demonstration Assistance Program. The major outcome variables were measures of dropping out, educational aspirations, self-esteem, and absenteeism. These authors had pre-random assignment measures on all these measures, plus measures on 32 other covariates. Of course, the correlation of pre- and post-measures for the outcomes in both of these studies may be lower than those for achievement test scores, so these findings do not definitively establish that pretest measures for the Project STAR data, were they available, would have the same effect on the quality of the nonexperimental estimates produced using propensity score matching.

Finally, many researchers who attempt to use nonexperimental estimators of impacts argue that it is important that the pool from which comparison group members is drawn be located as close as possible geographically to the site where the treatment group members are located. This recommendation comes from studies of the employment and training data where the cogency of the local market may make some sense. For education as an outcome, however, we would argue there is no analog to a "local market."

Some have also argued that the propensity score match pool of potential comparisons should be from the same school. This approach, however, would not constitute a true test of PSM. Treatment and control groups within schools are randomly assigned and therefore control students within the same school are probabilistically identical to the treatment group in both measured and unmeasured characteristics. Using PSM to create a control group in this situation would lead to results very close to the experimental results, but it would not be a true test of a nonexperimental methods. A nonexperimental method, by definition, should be able to account for differences between groups that are not randomly divided and may have unmeasured differences. Therefore, a true test of PSM must draw on a pool that is in some way different from the treatment group, in this case students at other schools.

If comparison group members were taken from the controls in the same school, we would essentially be replicating the original experiment.

Recognizing these limitations, however, there are many situations in which decisions are to be made about the design of an impact evaluation and decision makers would face similar limitations. For example, if one were testing an intervention in kindergarten, it would be difficult to get pre-kindergarten test scores. Even for older youth, nonexperimental estimators that depend on several periods of pre-intervention measurement on the outcome variable in order to estimate individual trajectories in the outcome changing 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, although it might be 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, for example, in the same school, if they are really that near, then it seems it should be feasible to use a random assignment design and obtain the far greater reliability of the experimental design estimate of the impact.

CONCLUSION

In this work, we tested the performance of a nonexperimental estimator of impact applied to an educational intervention—reduction in class size—where achievement test scores were the outcome. We compared the nonexperimental estimates of

impacts to "true" impact estimates provided by a random assignment design. We focused on constructing comparison groups using propensity score matching methods. In this context, our conclusion is that propensity score 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 short, in the search for a second-best impact estimator—that is, second best to a random assignment experiment—propensity score methods do not provide a golden panacea. We hope that this study raises a flag of caution for decision makers: Do not rush to adopt a propensity score matching estimator thinking it will be an adequate substitute for one derived from a true experimental design. Unfortunately, we cannot provide any guidance as to what nonexperimental methods may be the "second best" when the experimental methods are not feasible. We found that propensity score matching methods not only did not provide impact estimates that were close to the "true impact" estimates, but also that they performed no better in that regard than simpler nonexperimental methods, for example, simple regression adjustments. To our knowledge, some other popular nonexperimental methods, such as instrumental variables or difference-in-difference estimation, have not been assessed in the framework we used here, that is, compared to estimates of impacts derived from a random assignment experiment.

ELIZABETH Ty WILDE is a graduate student in the Department of Economics, Princeton University.

ROBINSON HOLLISTER is Professor of Economics at Swarthmore College.

ACKNOWLEDGMENTS

The research presented here was supported in part by a graduate research fellowship from the National Science Foundation and by the Industrial Relations Section at Princeton University. The authors wish to thank Roberto Agodini, Tom Dee, Mark Dynarski, Henry Farber, Matissa Hollister, Alan Krueger, Guido Imbens, Diane Whitmore Schanzenbach, Peter Schochet, the Princeton Development Working Group, and three anonymous reviewers for helpful criticisms and suggestions. Any remaining errors are the responsibility of the authors.

REFERENCES

Abadie, A., Drukker, D., Leber Herr, J., & Imbens, G. (2004). Implementing matching estimators for average treatment effects in Stata. Stata Journal, 4, 290–311.

Abadie, A., & Imbens, G. (2006). On the failure of the bootstrap for matching estimators. NBER Technical Working Paper No. 325. Cambridge, MA: National Bureau of Economic Research.

Agodini, R., & Dynarksi, M. (2004). Are experiments the only option? A look at dropout prevention programs. Review of Economics and Statistics, 86, 180–194.

Dehejia, R., & Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. Journal of the American Statistical Association, 94, 1053–1062.

Dehejia, R., & Wahba, S. (2002). Propensity score matching methods for nonexperimental causal studies. The Review of Economics and Statistics, 84, 151–161.

Diaz, J., & Handa, S. (2006). An assessment of PSM as a nonexperimental impact estimator: Evidence from Mexico's PROGRESSA program. Journal of Human Resources, 41, 319–345.

Fraker, T., & Maynard, R. (1987). Evaluating comparison group designs with employment-related programs. Journal of Human Resources, 22, 194–227.

Friedlander, D., & Robins, P. (1995). Evaluating program evaluations: New evidence on commonly used nonexperimental methods. American Economic Review, 85, 923–937.

Glazerman, S., Levy, D., & Myers, D. (2003). Nonexperimental vs. experimental estimates of earnings impacts. Annals of the American Academy of Political and Social Science, 55, 63–93.

Heckman, J., & Hotz, V. J. (1989).Choosing among alternative nonexperimental methods for estimating the impact of social programs: The case of Manpower training. Journal of the American Statistical Association, 84, 862–874.

Heckman, J., Ichamura, H., Smith, J., & Todd, P. (1996). Sources of selection bias in evaluating social programs: An interpretation of conventional measures and evidence on the effectiveness of matching as a program evaluation method. Proceedings of the National Academy of Sciences, 93, 13416–13420.

Heckman, J., Ichamura, H., Smith, J., & Todd, P. (1998). Characterizing selection bias using experimental data. Econometrica, 66, 1017–1098.

Heckman, J., Ichamura, H., & Todd, P. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. Review of Economic Studies, 64, 605–654.

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

Krueger, A., & Hanushek, E. (2000). The class size policy debate. EPI Working Paper No. 121. Washington, DC: Economic Policy Institute.

LaLonde, R. (1986). Evaluating the econometric evaluations of training programs. American Economic Review, 76, 604–620.

Lipsey, M., & Wilson, D. (2000). Practical meta-analysis. Thousand Oaks, CA: Sage Publications.

Manly, B. F. (1986). Multivariate statistical methods: A primer. New York, NY: Chapman and Hall.

Michalopolous, C., Bloom, H., & Hill, C. J. (2004). Can propensity score methods match the findings from a random assignment evaluation of mandatory welfare-to-work programs? The Review of Economics and Statistics, 86, 156–179.

Rouse, C., & Krueger, K. (2004). Putting computerized instruction to the test: A randomized evaluation of a "scientifically based" reading program. Economics of Education Review, 23, 323–338.

Smith, J., & Todd, P. (2005). Does matching overcome LaLonde's critique of nonexperimental estimators? Journal of Econometrics, 125, 305–353.

Word, E., Johnston, J., Bain, H., Fulton, B., Zaharias, J., Achilles, C., et al. (1990). The state of Tennessee Student/Teacher Achievement Ratio (STAR) Project final summary report (1985–1990). Tennessee State Department of Education.

Appendix A

Table A1. Project STAR results from within bin statistical tests.

	Sample Size		Statistical Test	
School ID Number/Bin	Treatment Group (n)	Comparison Group (n)	Estimated p-value (Ho: Pt = Pc)	Estimated p-value (Ho: Xt = Xc)
School 7				
Bin 1	30	1745	.06	.84
School 9				
Bin 1	16	74	.06	.73
Bin 2	4	47	.23	.05
Bin 3	4	18	.06	.41
Bin 4	4	31	.07	.60
Bin 5	9	58	.08	.11
Bin 6	1	4	.22	—
School 16				
Bin 1	33	665	.06	.71
School 22				
Bin 1	16	414	.08	.71
Bin 2	2	5	.06	.64
Bin 3	9	316	.06	.05
Bin 4	1	44	.42	.82
School 27				
Bin 1	24	641	.93	.95
School 28				
Bin 1	14	199	.08	.24
Bin 2	5	45	.55	.052
Bin 3	14	138	.29	.54
Bin 4	1	2	.09	—
Bin 5	17	253	.18	.92
Bin 6	2	15	.05	.41
Bin 7	3	58	.07	.22
School 32				
Bin 1	18	504	.06	.35
Bin 2	7	82	.05	.61
Bin 3	3	54	.05	.16
School 33				
Bin 1	24	598	.06	.78
Bin 2	1	2	.06	—
Bin 3	3	118	.61	.61
School 51				
Bin 1	49	58	.29	.74
School 63				
Bin 1	12	708	.12	.40
Bin 2	7	205	.05	.73
Bin 3	5	360	.11	.43
Bin 4	1	3	.07	—
Bin 5	4	94	.10	.17
School 72				
Bin 1	2	122	.10	.90
Bin 2	4	244	.06	.12
Bin 3	6	376	.10	.40
Bin 4	12	807	.06	.72

Appendix B. Standard Error Estimates.

There are two central problems to be faced in trying to generate unbiased (or at least consistent) estimates of the standard errors for these equations. First, because students are clustered in classrooms, we need to take account of clustering (or design effects) in our estimating equations for both experimental and nonexperimental impact estimates. Second, in the nonexperimental impact estimation, we need to take account of the constructed nature of the propensity score matched comparison groups.

In an early draft of this study, we did not address this first issue in our estimation procedures and, as a result, our standard errors for both sets of estimates were downwardly biased. In this version, we run estimation programs that cluster within classrooms (procedure GEN MOD in SAS).

In the earliest draft of this study, we attempted to take account of the propensity score matched comparison groups by using bootstrapped estimates of standard errors. These estimates did not differ greatly from the conventionally estimated standard errors.[33] However, very recently, Abadie and Imbens (2006) demonstrated that with propensity score matched estimates, bootstrapped estimates of standard errors may be biased and inconsistent. Fortunately, Abadie, Imbens, and others developed a program to generate estimates of standard errors for propensity score matched estimates that are not subject to the bias suffered with common bootstrap estimates (Abadie, Drukker, Leber Herr, & Imbens, 2004). This program is available for STATA (nnmatch). We have generated estimates using this program. In general, the standard errors generated by the nnmatch program are larger than the conventionally generated standard error estimates. These nnmatch estimates are available on request from the authors.

Unfortunately, however, at present there is to our knowledge no estimation program available that will deal both with clustering and with the problems of correct standard errors for propensity score matched estimates.

We decided, therefore, to present in the text the estimates that account for the clustering problem only and which are in general larger than the standard errors that do not account for clustering and, also, on average, larger than the standard errors that account for the propensity score matching. We estimate the standard errors for both the experimental and the nonexperimental impact estimates this way. The experimental standard errors should be unbiased and consistent. The nonexperimental estimates of standard errors, however, may be underestimated, because they do not account for the repeated sampling in the propensity score matched estimates.

Appendix C. Krueger's Cost-Benefit Calculations for Project STAR.

Krueger (1999; Krueger & Hanushek, 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 1 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-earnings 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 three 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 almost 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, is designed 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: 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 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 the Krueger papers, we can determine that 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 6 for deciding whether smaller classes in kindergarten pay off.

[33] This was a finding similar to those for other propensity score matched studies at that time. See Michalopolous, Bloom, & Hill (2004) and Agodini & Dynarski (2004).