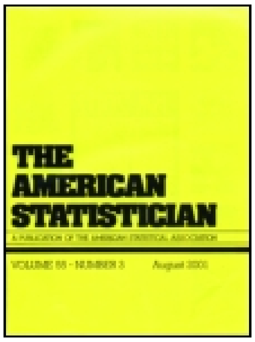

The American Statistician

ISSN: 0003-1305 (Print) 1537-2731 (Online) Journal homepage: http://amstat.tandfonline.com/loi/utas20

Propensity Score Matching

Deborah N Peikes, Lorenzo Moreno & Sean Michael Orzol

To cite this article: Deborah N Peikes, Lorenzo Moreno & Sean Michael Orzol (2008) Propensity Score Matching, The American Statistician, 62:3, 222-231, DOI: 10.1198/000313008X332016

To link to this article: http://dx.doi.org/10.1198/000313008X332016

Published online: 01 Jan 2012.

Propensity Score Matching: A Note of Caution for Evaluators of Social Programs

Deborah N. PEIKES, Lorenzo MORENO, and Sean Michael ORZOL

Over the past 25 years, evaluators of social programs have searched for nonexperimental methods that can substitute effectively for experimental ones. Recently, the spotlight has focused on one method, propensity score matching (PSM), as the suggested approach for evaluating employment and education programs. We present a case study of our experience using PSM, under seemingly ideal circumstances, for the evaluation of the State Partnership Initiative employment promotion program. Despite ideal conditions and the passing of statistical tests suggesting that the matching procedure had worked, we find that PSM produced incorrect impact estimates when compared with a randomized design. Based on this experience, we caution practitioners about the risks of implementing PSM-based designs.

KEY WORDS: Case study; Evaluation design; Nonexperimental methods; PSM; Randomized design; Social Security Disability Insurance; Supplemental Security Income.

Over the past 25 years, evaluators of social programs have devoted considerable attention to the use of nonexperimental methods to assess the impacts of interventions, most notably in the areas of employment and education (Lalonde 1986; Fraker and Maynard 1987; Dehejia and Wahba 1999). Distinguished econometricians and statisticians have argued that when an experimental design is infeasible, or when the evaluation questions are broader than assessing the effect of an intervention on participants, alternative designs such as those based on the identification of matched comparison groups may be effective (Heckman 1992; Heckman et al. 1998a). In particular, the selection of comparison groups using propensity score matching (PSM), originally proposed by Rosenbaum and Rubin (1983), has captured the interest of evaluators. In several instances this methodology has been endorsed, if certain conditions are met, as a viable alternative to experimental methods in the area of employment programs (Heckman et al. 1998b). PSM is an approach that allows researchers to match individuals in a treatment group to others who did not participate but have comparable characteristics. The innovation of PSM compared to other matching methods is that it develops a single (propensity) score that encapsulates multiple characteristics, instead of requiring a one-to-one match of each characteristic—simplifying matching by reducing dimensionality. While PSM has its supporters, opponents have warned of several limitations, most notably (1) requiring knowledge of the experimental impact estimates as the only means to assess whether PSM yields unbiased impact estimates, and (2) requiring data on the right variables that would accurately predict program participation (Smith and Todd 2005a). Despite the ongoing, healthy debate between the two camps, the popularity of PSM continues to grow. Indeed, some federal agencies now expect program evaluators to use this approach as a substitute for experimental designs but do not consider the perils and costs associated with its use.

In this article, we present evidence showing that the PSM does not replicate experimentally derived impact estimates. By *replicate*, we mean produce estimates that are statistically similar (of roughly the same magnitude and in the same direction), so that a policymaker or funder would draw the same policy conclusions about the effectiveness of the program. We do so by presenting a case study of our experience using PSM in the evaluation of the State Partnership Initiative—one of the first large-scale attempts by the Social Security Administration (SSA) to promote employment among beneficiaries with disabilities who receive Supplemental Security Income (SSI) or Social Security Disability Insurance (SSDI) benefits and who volunteered to participate in these employment-promotion projects. The monthly employment rate of SSI beneficiaries with disabilities has hovered below 7%, and four-fifths of beneficiaries who do work earn less than the amount that SSA designates as substantial gainful activity ($800 per month in 2003) (Pickett 2003). The goal of the intervention was to boost these low employment rates and earnings, reflecting a consensus that no one with a disability should be denied the right to participate fully in society, including in work, because of external barriers that can

Deborah Peikes (E-mail: *dpeikes@mathematica-mpr.com*) and Lorenzo Moreno (E-mail: *lmoreno@mathematica-mpr.com*) are Senior Researchers at Mathematica Policy Research, Inc. (MPR), P.O. Box 2393, Princeton, NJ 08543-2393. Sean Orzol (E-mail: *orzol@umich.edu*) was an analyst at MPR when the work was completed and is now a doctoral student in the Department of Health Management and Policy at the University of Michigan. The Social Security Administration (SSA) funded this research under Contracts 03-60161 600-96-27333/0440-99-38637; MPR also contributed to supporting the research. The authors acknowledge Paul O'Leary at SSA as well as Craig Thornton, Roberto Agodini, and Peter Schochet at MPR for invaluable assistance in conceptualizing and implementing this project. In addition, Paul O'Leary and Roberto Agodini, as well as Melissa Clark, Steve Glazerman, and David Wittenburg from MPR, gave us useful comments on previous versions of this article. Finally, thanks to Paul O'Leary for tirelessly running the matching and impact programs on the SSA mainframe; to Evan Schechter, our project officer, for advice; and to Nora Paxton, Miriam Loewenberg, and Lesley Hildebrand for computing assistance.

©American Statistical Association DOI: 10.1198/000313008X332016

reasonably be removed.

As we will discuss, the voluntary nature of participation in the intervention raised numerous challenges for the evaluation. Nevertheless, our evaluation offered the opportunity to test PSM under seemingly ideal circumstances that included the availability of comprehensive administrative data on a key predictor of both participation and subsequent employment outcomes—employment and earnings for five years before the beginning of the intervention; large pools of potential comparison group members (hereafter candidates); detailed data on program participation; a rigorous protocol for deciding the specification of the propensity score models; and impact estimates derived from experimental methods to validate the performance of PSM. Despite these seemingly ideal conditions, and the passing of tests that, according to the literature, indicate PSM had worked, PSM produced impact estimates that differed considerably from the gold standard experimental estimates in terms of statistical significance, magnitude, and most important, sign. Specifically, the PSM approach would have led policymakers to conclude incorrectly that the interventions *increased* earnings, when they actually *decreased* or had *no effects* on earnings. Based on this experience, our goal is to caution practitioners that PSM can generate incorrect estimates, even under seemingly ideal circumstances.

This article is divided into five sections. First, we provide the context for how the evaluation of the State Partnership Initiative demonstration was designed. Next, we explore the rationale for using PSM in this evaluation and summarize the recent experience of practitioners using PSM in the evaluation of employment programs. In the third section, we describe in detail how PSM was implemented in the evaluation, including a description of the comparison groups selected from the matching process. In the fourth section, we report findings from the validation of the impact estimates derived from the PSM method, and in the fifth section we discuss the lessons learned from the implementation of PSM in the evaluation of the State Partnership Initiative.

1. DESIGN OF THE STATE PARTNERSHIP INITIATIVE EVALUATION

The demonstration's evaluation was designed to estimate impacts by comparing outcomes for the beneficiaries with disabilities who participated in each of 11 projects (participants) with outcomes for a comparison group that was selected using PSM to match the participants before enrollment in terms of demographic characteristics, previous benefit receipt, employment, earnings, and economic and service environments. The 11 projects were operated in California, Iowa, Minnesota, New Hampshire, New Mexico, New York, North Carolina, Ohio, Oklahoma, Vermont, and Wisconsin. Funded in fall 1998, the first project began enrollment in January 1999, and most projects provided services through September 2004. SSA promoted innovative initiatives by ceding responsibility to the state partners for designing the interventions and evaluation designs, and for identifying and recruiting the target populations. The projects all tested benefits counseling (which explains to beneficiaries how working affects receipt of multiple types of public benefits). In addition, some tested employment services, consumer direction, and waivers to SSI regulations (primarily waivers that allowed beneficiaries to retain more of their earnings). A detailed description of the demonstration and the context in which it was implemented is in Peikes et al. (2005). A beneficiary's decision to volunteer for a project reflected two factors: (1) the decision made by the beneficiary to volunteer, and (2) any targeting and outreach done to particular types of beneficiaries by each project. All but three projects chose to use nonexperimental designs to evaluate their interventions. Consequently, when Mathematica Policy Research, Inc. (MPR) was asked to use a uniform design to estimate impacts for each of the projects for a national evaluation, random assignment had already been ruled out because of the projects' individual decisions. Among the possible nonexperimental approaches to evaluating the projects, selecting comparison groups using PSM was one method available.

2. RATIONALE FOR THE PROPENSITY SCORE MATCHING METHOD

In their original article, Rosenbaum and Rubin (1983) argued that when many characteristics are used in the matching process, statistical matching using propensity scores can be used to select comparison groups that are similar, on average, to participants along those characteristics. The propensity score is a single summary number that can be used to determine the extent to which one person is similar to another along observed characteristics. Rosenbaum and Rubin showed that, in situations where the outcome is independent of participant status, *conditional on the observed characteristics,* the potential outcome is also independent of participant status, conditional on the propensity score.

The interest in PSM accelerated after Heckman and colleagues (1998a,b) assessed the validity of using propensity matching to characterize selection bias using experimental data, and Dehejia and Wahba (1999) used PSM to approximate the true (experimental) results from the National Supported Work Demonstration. They attributed the success of the nonexperimental PSM method to having multiple measures of the preenrollment outcome in their data to use in selecting comparison group members. In addition, they touted the ability of PSM to "balance" the treatment and comparison groups on their preenrollment outcomes and other important characteristics that were available in the data and that they believed were also correlated with important unobserved characteristics affecting postenrollment outcomes.

At the time, the demonstration appeared to be an ideal context in which to use PSM to estimate impacts. First and foremost, we had excellent employment and earnings data, and because participants were statistically more likely to be employed and to earn more than eligible beneficiaries who lived in the same areas but chose not to participate, we believed we could use these important predictors of participation. Second, SSA administrative data enabled us to match participants to eligible beneficiaries who lived in nondemonstration comparison areas with similar economic and service environments. Third, in these comparison areas we had large pools of candidate comparison group members from which to select. In the three states with random assignment, the ratio of candidate comparison group members to participants ranged from 8:1 to over 90:1 (Table 1). Fourth, we had more than 250 variables summarizing demographics; disability type; participation in SSI, SSDI, and Medicaid; use of work incentives; and employment history. These four features are the conditions that the literature says are key to obtaining results from PSM that successfully mimic those based on random assignment. The evaluation's inclusion of three projects that used random assignment offered the study team the unique possibility to test whether the nonexperimental estimates replicated the direction and magnitude of the experimental estimates. (To take into account sampling variability, we tested whether the estimates were statistically similar with a significance level of 5%.) Reflecting promising early results about PSM's potential, a Technical Evaluation Support Group made up of leading labor economists, policymakers, and advocates agreed that PSM was the best available alternative to random assignment. The design team considered comparing all eligible beneficiaries living in the catchment areas with all eligible beneficiaries living in matched comparison areas. However, the small sample sizes targeted by the states would not have been adequate to detect policy-relevant effects.

Since publication of the article by Dehejia and Wahba (1999) and the design of the demonstration's evaluation, however, several studies have faulted PSM because, while it sometimes does come close to replicating experimentally derived results, it more often produces estimates that differ by policy-relevant margins. These studies found that PSM could not reliably generate impact estimates in the same direction as well-implemented randomized designs across a range of social programs and policies: reductions in classroom size (Wilde and Hollister 2007); job training and employment services (Glazerman et al. 2003); and school dropout prevention programs (Agodini and Dynarski 2004). Furthermore, Smith and Todd (2005a) directly refuted the findings of Dehejia and Wahba by showing their results to be sensitive to sample and model specification. These various studies demonstrate that nonexperimental methods can *sometimes* give accurate estimates of program impacts but that researchers cannot know in advance when and under which circumstances they will do so. This is not surprising, given the longstanding difficulty of identifying nonexperimental methods that can consistently measure accurate impacts (see, e.g., LaLonde 1986; Fraker and Maynard 1987; Friedlander and Robins 1995). Reflecting this difficulty, the medical field requires randomized controlled trials to support evidence-based medicine.

3. HOW PSM WAS IMPLEMENTED IN THE DEMONSTRATION'S EVALUATION

3.1 Overall Approach

Our matching process selected comparison beneficiaries for participants from the pool of candidate comparison group members using propensity scores to ensure comparability along 250 important variables measured before enrollment, summarizing the beneficiaries' demographics, diagnosis, and Social Security benefit type and amount from Social Security administrative records; and employment history from Social Security and Internal Revenue Service (IRS) federal income tax records. We followed three steps:

1. Identify pools of candidate comparison group members in areas where the demonstration was not offered. The first step was to identify a set of comparison counties with economic and service environments similar to those of the counties in which the projects operated. We completed this step in 2002 (Agodini et al. 2002), generally using characteristics from June 1999 (approximately the first month of project enrollment). State project staff reviewed the initial selections to identify recent labor market changes or policy or environmental difference that cannot be measured well with available data. When project staff provided a clear reason that an area selected by the quantitative process was not a good match, the county was dropped. We next limited the candidate comparison beneficiaries eligible for PSM selection to those who lived in each project's selected comparison areas and met each project's disability, benefit coverage, and in one state, unemployment eligibility criteria. For most projects, there were thousands, sometimes tens of thousands, of candidate comparison group members.

For each project, we selected comparison groups separately for as many as four groups of participants. (All participants were included, whether or not they later disenrolled from services.) Participants were first divided into those who at enrollment were receiving either SSI benefits only or both SSI and SSDI benefits concurrently (hereafter, "SSI-concurrent" participants), and those who at enrollment were receiving only SSDI benefits (hereafter, "SSDI-only" participants). We did this because SSA tracks additional data for SSI and concurrent beneficiaries, and because the projects were likely to have affected the employment behaviors of the two groups in different ways (because the SSI and SSDI programs tend to serve different populations and use different rules and work incentives). For the New York project, we further divided participants into those living in the New York City site and those living in the Buffalo site to ensure a comparable level of economic activity and employment support services. For the three projects that had used an experimental design, we selected five sets of comparison groups (New York SSI-concurrent in New York City, New York SSI-concurrent in Buffalo, New Hampshire SSI-concurrent, New Hampshire SSDI, and Oklahoma SSI-concurrent). (The New York project excluded SSDI-only beneficiaries, and the sample size of SSDI-only beneficiaries in Oklahoma was too small to allow estimation of impacts.)

2. Estimate a probability model of participant status. Our goal when selecting comparison group members was to find people who were at the same point in the employment decision process as participants but did not have the opportunity to enroll in the State Partnership Initiative. This is important, because people tend to enter employment programs when they want to find a job or increase their earnings—not at a random point in their lives (Ashenfelter 1978). To address this issue, we first determined the months during which each candidate comparison group member lived in the comparison area and met select

Table 1. Baseline Characteristics of Participants Randomized Through December 2001 And Comparison Group Members

	New York SSI-concurrent (benefits counseling and waivers)			New York SSI-concurrent (benefits counseling, waivers and employment services)			New Hampshire SSI-concurrent			New Hampshire SSDI			Oklahoma SSI-concurrent		
		Comparison group			Comparison group			Comparison group			Comparison group			Comparison group	
	Participants	Selected	Candidate	Participants	Selected	Candidate	Participants	Selected	Candidate	Participants	Selected	Candidate	Participants	Selected	Candidate
Earnings in the year before enrollment ($)	$1,665	$1,421	$1,210*	$1,812	$1,600	$1,210**	$2,140	$3,330	$2,706	$2,723	$2,545	$2,683	$1,030	$922	$645**
Employed in the year before enrollment (%)	41.5	31.8**	23.0***	38.2	36.2	23.0***	54.5	50.0	48.2	65.7	57.1	30.4***	28.3	34.4	20.4***
Age at enrollment	41.0	40.7	41.6	41.1	41.1	41.6	34.7	40.5	41.1**	46.7	42.9*	49.8*	41.1	41.6	41.5
White (%)	42.7	38.3	58.6***	36.9	37.5	58.6***	90.9	81.8	93.1	99.9	99.9	97.6	51.6	56.4	77.5***
Disability: mental disorder (%)	92.8	100***	100***	94.0	100***	100***	59.1	68.2	52.1	80.0	71.4	37.1***	96.5	100***	100***
Average SSI benefit per month in 2 years before enrollment	$461	$469	$389***	$468	$452	$389***	$321	$174**	$133***	n.a.	n.a.	n.a.	$391	$389	$285***
Average SSDI benefit per month in 2 years before enrollment	$176	$171	$194	$165	$160	$194*	$176	$228	$416***	$779	$744	$839	$96	$89	$150***
Used work incentives in the year before enrollment (%)	18.8	15.5	16.6	18.9	16.6	16.6	27.3	22.7	20.7	n.a.	n.a.	n.a.	13.7	12.4	14.4
Sample Size[a]	277	253	13,183	301	271	13,183	22	19	851	35	34	3,190	314	247	2,603

Source: All characteristics, except annual earnings and employment, are based on SSA administrative data. The Summary Earnings Record (SER) provides annual IRS earnings and employment.

Notes: All dollar amounts are inflation adjusted to 2004 dollars using the Consumer Price Index for Urban Wage Earners and Clerical Workers (CPI-W).

Statistics are computed using weights, where each participant received a weight of one and each selected comparison group member received a weight equal to the number of times he or she was matched to a participant.

The candidate comparison group members for all three states were limited to beneficiaries aged 18 to 65. New York reported having mental illness as their only targeting criterion, so we limited the pool of candidate comparison group members to those with a diagnosis of mental illness. We limited the pool for Oklahoma using their criteria of having a diagnosis of mental illness and currently being unemployed. New Hampshire reported targeting people with a diagnosis of mental illness in one site and all diagnoses in the other site. More than 10% of participants in the site that targeted people with mental illness did not have a diagnosis of mental illness recorded in the SSA data; therefore, we did not limit our pool to those with a diagnosis of mental illness.

NA = not available; SSA = Social Security Administration; SSI = Supplemental Security Income.

[a] The PSM comparison group was drawn for *participants only.* For New York and Oklahoma, participants were limited to beneficiaries randomized into the treatment groups through December 2001. For New Hampshire, the participants include beneficiaries randomized into both the treatment and the control groups through December 2001. Control group members, and their selected comparison group members, were later dropped from the PSM sample when we estimated the impacts of the New Hampshire program using PSM. For some characteristics, the sample size may differ because of missing data.

*Significantly different from participants at the 0.10 level, two-tailed test.

**Significantly different from participants at the 0.05 level, two-tailed test.

***Significantly different from participants at the 0.01 level, two-tailed test.

eligibility criteria that we could simulate using data (i.e., age, disability type, beneficiary type, and, in one state, unemployment). We then picked one of those months at random so that the distribution of simulated enrollment months for the comparison group matched the distribution of the actual enrollment dates for participants. We measured all time-varying characteristics relative to the real (for participants) or simulated (for candidate comparison group members) enrollment date. Because our matching process focused on each comparison group member's characteristics at a random point in time, our approach might not have identified another point at which that beneficiary could have been a good (or even better) comparison group member for a specific participant. However, the large pool of candidate comparison group members made this unnecessary.

We estimated a logit model in which a binary dependent variable that equaled one for participants and zero for candidate comparison group members was regressed on available characteristics. We estimated the model for each project grouping using participants and all candidate comparison group members from their respective comparison counties. Because the number of available variables often exceeded the number of beneficiaries in each participant group, or the variables were highly correlated and would have led to collinearity problems, we used a subset of these variables in the first iteration of the matching process. This subset consisted of those variables that statistical tests indicated were different across participants and candidate comparison group members: about 50 variables for the SSI-only or concurrent group and 26 for the SSDI-only group (available upon request).

3. Use propensity scores to select comparison group member. The propensity score equals the predicted probability of participating, given the individual's values for the variables included in the regression model. For each participant, we followed the PSM procedures used in the literature to select the candidate comparison group member with the closest absolute propensity score—the "nearest neighbor." We conducted the selection process with replacement so that a candidate comparison group member could be matched to several participants (Dehejia and Wahba 2002; Smith and Todd 2005a). About 90% of SSI-concurrent participants matched to a unique comparison group member. This one-to-one match percentage was higher (97%) for SSDI-only beneficiaries.

3.2 Tests Used to Assess the Comparability of Participants and Comparison Groups

To determine whether the participants and selected comparison group members were well matched on observable characteristics available to us, we first assigned participants and comparison group members to strata, where each stratum included participants and selected comparison group members whose average propensity scores were not significantly different.

We defined the strata in a way that has often been used in other studies (see, e.g., Dehejia and Wahba 1999). We ranked the participants and comparison group members according to their propensity scores. Beneficiaries were then divided into strata, with an equal number in each stratum. Each stratum contained enough beneficiaries to ensure that statistical tests conducted within it would have enough power to detect any meaningful differences in the variables of participants and comparison group members. A stratum that contains about 80 beneficiaries (equally split between participants and comparison group members) is typically sufficient. Given this definition, we determined that three strata would be feasible given the sample size for most projects. Two groups (New Hampshire SSI-concurrent and Oklahoma SSDI-only) did not have a sufficient sample size to divide into three strata; we used two strata for those groups. Within each stratum, we conducted two-tailed t-tests of the similarity of each of the full set of nearly 250 variables between participants and selected comparison group members. We considered a comparison group to be well matched to its respective group of participants if it passed "the 95% test": for each stratum, 95% of the statistical tests of the similarity of each of the 250 variables (whether or not they were included in the matching regression model) failed to detect a statistically significant difference (at the 5% level). Note that the 95% test described by Agodini et al. (2002) tested the total number of two-tailed t-tests of similarity summed across strata. Our test was more conservative, because it tested this within each stratum.

In addition to ensuring that the average propensity scores across participants and comparison group members were similar, and that the 95% test was passed, we were more conservative than most studies, requiring that the two groups not show specific patterns of differences among any of the fewer than 5% of variables that were significantly different within the stratum. We paid closest attention to eliminating preenrollment differences in average earnings and employment, because these are the strongest predictors of postenrollment values of earnings and employment, the outcomes of interest.

If a comparison group failed any of these stratum-specific tests, following the literature, we respecified the regression model and reselected the comparison group until it passed. For variables that were already included in the model but differed across participants and comparison group members, respecifying the model meant adding higher-order or different specifications for those variables. Examples of the additional variables we included are age cubed, average earnings in the past two years, and log of earnings in the past year squared. Decisions about how to respecify the model followed a Delphi-type consensus approach among the researchers and computer programmers involved in the evaluation and were carefully documented (documentation available from the authors).

3.3 Comparison Groups Selected From the Matching Process

The selected comparison groups passed all the tests used by the PSM approach. For each of the participant groups, we were able to draw a comparison group that passed the stratum-specific 95% test. In addition, we were satisfied that, among those variables that were significantly different within the stratum, there were no patterns that remained dissimilar between the participant and comparison groups.

While the participants differed markedly from the candidate comparison group members, PSM successfully selected comparison groups that were remarkably similar along the entire spectrum of available characteristics, including diagnosis, education, benefit type and amount, and work history. Table 1 shows this information for the most important variables in the three states that used randomized designs. With the exception of a few variables which were significantly different at the $p = 0.05$ or 0.10 level, the only differences between participants and selected comparison groups were small and reflected our decision to restrict the pool of comparison group members to beneficiaries who met the projects' individual eligibility criteria. For example, we limited the comparison groups in both New York and Oklahoma to beneficiaries with psychiatric disabilities, because that eligibility criterion was used by the projects. However, only 94% of participants in New York and 97% in Oklahoma had such a diagnosis in the SSA records, and, by design, all of the selected comparison group members did. The Oklahoma project was also limited to beneficiaries who were not working at intake, but 3.5% of participants were.

4. ASSESSING THE VALIDITY OF PSM

4.1 Approach

While most studies would conclude that PSM worked and proceed to estimate impacts, we were able to analyze the validity of PSM by comparing estimated impacts based on experimental methods with those based on the nonexperimental comparison groups selected through PSM. The first step in the analysis consisted of comparing the preenrollment characteristics of the treatment and randomly assigned control groups to ensure that the random assignment approach, designed and supervised by MPR, was correctly implemented (it was) and computing experimental results for each of the three projects in which a randomized design was used (New Hampshire; New York, which tested two separate interventions; and Oklahoma). We use the term *control group* to refer to the randomly assigned groups, and the term *comparison group* to indicate the nonexperimental groups chosen through PSM. We computed experimental results as the difference in average outcomes between the randomly assigned treatment groups and the control groups using an *intent-to-treat* design (i.e., including all sample members regardless of whether they participated or the length of their participation). For the nonexperimental PSM design, we estimated impacts in a similar way, except that rather than use randomly assigned control groups, we used the comparison groups selected using PSM. To minimize the effect of any chance preenrollment differences between groups in the outcome variables, we used the *difference-in-differences* method (comparing the treatment-control difference in the changes in the outcomes over time) and regression adjustment (which increases precision of the impact estimates) to estimate those impacts.

If PSM worked, we expected the impact estimates from PSM and the randomized designs to be statistically comparable 95% of the time. As noted, the results of the validity analysis provide a strong indication of whether PSM can be used to estimate the true effect of the projects.

We estimated demonstration impacts on three key employment and earnings outcomes during the year after the calendar year of enrollment for enrollees through December 2001. The treatment group sample for the PSM results contained *participants* enrolled through December 2001. The sample for the random assignment results contained *all beneficiaries* randomized to the treatment group through December 2001, regardless of whether they participated in the demonstration or dropped out. The cutoff date was dictated by the availability of IRS earnings data in full calendar years and by the 14-month lag in the availability of those data. We divided estimates for the random assignment results by the participation rate to obtain comparable per-participant estimates to those from the PSM results.

The three measures we examined are (1) the change in the proportion employed (i.e., having reported earnings) at all relative to the year before the calendar year of enrollment, (2) the change in earnings relative to the year before the calendar year of enrollment, and (3) the change in earnings relative to the average over the two years before the calendar year of enrollment.

4.2 Findings

To assess the validity of the comparison groups selected using PSM, we compared per-participant treatment-*control group* differences in outcomes (based on the randomized designs) with treatment-*comparison group* differences (based on the nonexperimental PSM design) in New York, New Hamsphire, and Oklahoma.

New York. Estimates of the impacts on employment from the comparison group were similar to those from the experimental design; impacts on earnings, however, were a poor approximation of the impacts from the experimental design. As Table 2 shows, the estimates of the effect on employment rates based on the PSM comparison group were larger than under the randomized study for the benefits counseling and waivers group, but would lead to the same conclusion: 14 percentage points ($p = 0.001$), compared with a smaller 9 percentage points and not quite statistically significant ($p = 0.186$) for the experimental estimates. For the benefits counseling, waivers, and employment services group, the other intervention New York tested, both methods estimated impacts to be statistically significant and similar to each other: 26 percentage points ($p < 0.001$), versus 17 percentage points ($p < 0.008$) for the experimental estimates.

Notably, the comparison group approach incorrectly estimated impacts on earnings. For example, the PSM approach estimated statistically significant *positive* impacts of between \$1,000 and \$1,200, whereas the experimental estimates were statistically significant and *negative* for the benefits counseling and waivers group (−\$1,080 or −\$1,161, depending on the model specification) and −\$367 or −\$455 and not statistically significant for the benefits counseling, waivers, and employment services group.

For those curious about how the intervention could have reduced earnings, though the treatment groups' earnings during the year after randomization were greater than they were during both the first year and first two years before randomization (data

Table 2. Per-Participant Impact Estimates Among Enrollees Through December 2001, by Comparison Group Selection Method

	PSM		Random assignment	
	Impact	*p*-value	Impact	*p*-value
Change in proportion employed in the year after the randomization year relative to the year before (percentage points)				
New York—SSI/concurrent				
Benefits counseling and waivers	14.3***	0.001	8.8	0.186
Benefits counseling, waivers, and employment services	25.9***	<0.001	17.0***	0.008
New Hampshire (*note small sample sizes)*				
SSI/concurrent	−4.1	0.818	−29.5*	0.065
SSDI only	16.7	0.249	−29.6**	0.018
Oklahoma				
SSI/concurrent	10.6***	0.005	17.0	0.152
Change in earnings in the year after the randomization year relative to the year before (dollars)				
New York—SSI/concurrent				
Benefits counseling and waivers	1,214***	<0.001	−1,080*	0.059
Benefits counseling, waivers, and employment services	1,209***	0.002	−455	0.401
New Hampshire				
SSI/concurrent	3,942***	0.009	−709	0.511
SSDI only	339	0.694	−1,633**	0.045
Oklahoma				
SSI/concurrent	−75	0.715	451	0.448
Change in earnings in the year after the randomization year relative to the two years before (dollars)				
New York—SSI/concurrent				
Benefits counseling and waivers	977***	0.001	−1,161**	0.045
Benefits counseling, waivers, and employment services	1,188***	0.002	−367	0.504
New Hampshire				
SSI/concurrent	5,620***	0.001	−597	0.530
SSDI only	2,166**	0.047	−512	0.670
Oklahoma				
SSI/concurrent	59	0.764	43	0.941

Source: Calculations conducted by Mathematica Policy Research, Inc., on SSA Administrative and Summary Earnings Record (SER) data.

C = comparison/control; PSM = Propensity Score Matching; RA = random assignment; SSA = Social Security Administration; SSDI = Social Security Disability Insurance; SSI = Supplemental Security Income; T = treatment.

Notes: The December 2001 cutoff date for defining the sample was dictated by the availability of IRS earnings data in full calendar years and by the 14-month lag in the availability of those data.

The research sample for the PSM results contains *participants* enrolled through December 2001 and their selected comparison group members (T; C): New York—SSI: benefits counseling and waivers (277; 253); New York—SSI: benefits counseling, waivers, and employment services (301; 271); New Hampshire SSI (22; 19); New Hampshire SSDI (35; 34); Oklahoma SSI (314; 244).

The sample for the random-assignment results contains *all* beneficiaries randomized to the treatment and control groups through December 2001 (T; C): New York—SSI: benefits counseling and waivers (937; 914); New York—SSI: benefits counseling, waivers, and employment services (932; 914); New Hampshire SSI (22; 27); New Hampshire SSDI (35; 29); Oklahoma SSI (1,440; 256).

Because the experimental treatment group includes *both participants and nonparticipants* (to preserve the intent-to-treat analysis), the sample sizes used in the experimental and nonexperimental designs are different. We divided the random assignment impact estimates for the random assignment design by the participation rate to obtain per-participant estimates that are comparable to the PSM estimates.

*Significantly different from participants at the 0.10 level, two-tailed test.

**Significantly different from participants at the 0.05 level, two-tailed test.

***Significantly different from participants at the 0.01 level, two-tailed test.

not shown), the control group's annual earnings rose by an even larger amount, resulting in the negative impact associated with the intervention.

New Hampshire. Although the sample sizes were small for New Hampshire, with only 22 participants in the SSI-concurrent group and 35 in the SSDI-only group, we compare the PSM estimates with the random assignment outcomes here for completeness. The results for the New Hampshire project suggest that the comparison groups did a poor job approximating impacts from the experimental design for both employment and earnings outcomes. In particular, the random assignment impact estimates suggest that the project *decreased* employment rates for both SSI-concurrent, and SSDI-only participants: −30 percentage points for each group ($p = 0.065$ and $p = 0.018$, respectively), but the comparison group estimate would indicate there was no effect: −4 percentage points ($p = 0.818$) and 17 percentage points ($p = 0.249$), respectively, for SSI-concurrent, and SSDI-only participants.

Results for the change in earnings between the year after randomization and the average over the two years before randomization were especially worrisome. We found no effect on the change in earnings when using random assignment: −$597 ($p = 0.530$) for SSI-concurrent participants and −$512 ($p = 0.670$) for SSDI-only participants, but we found a very large positive and statistically significant effect when using the propensity score comparison groups: $5,620 ($p = 0.001$) and $2,166 ($p = 0.047$), respectively, for SSI-concurrent and SSDI-only participants. As with the New York example, the inferences we would make about the State Partnership Initiative in New Hampshire would be incorrect if we relied on PSM.

Oklahoma. The estimates using PSM came closer to approximating random assignment impacts for Oklahoma: the estimated effect on employment of 11 percentage points ($p = 0.005$) for the comparison group is fairly close to the effect for the random assignment design of 17 percentage points ($p = 0.152$). However, the PSM comparison group approach indicated a statistically significant effect, whereas random assignment finds no such effect. Turning to impacts on earnings, the PSM comparison group approach generates an estimate with the wrong sign, but because both the PSM and random assignment estimates are not statistically significant, both methods lead to a similar conclusion that there was no impact.

5. LESSONS LEARNED

Our results indicate that despite matching on a propensity score that summarized hundreds of powerful predictor variables—most notably prior employment and earnings, having large samples from which to choose the comparison groups, having comparison areas with comparable economic and social services environments as the demonstration areas, and passing multiple tests of the matching process—the impact estimates based on the PSM-selected comparison groups did not replicate the experimentally derived estimates. The estimates based on PSM overestimated the impacts of the projects on employment and typically overstated the impacts on earnings relative to the estimates derived from an experimental design. In two of the three projects, the PSM estimates would have supported the wrong policy conclusions by showing a positive impact when random assignment estimates showed a negative one. We learned several important lessons from this process:

In the case of the State Partnership Initiative, PSM may have failed because the study sample volunteered to participate in the program based on factors that either cannot be observed or are not typically collected in administrative data. In the State Partnership Initiative demonstration (as in most social programs evaluated in the United States), beneficiaries enrolled after they were recruited and volunteered to participate. The decisions to volunteer might have been related to observable and unobservable demographic or human capital characteristics that were present at that point in time (such as motivation, locus of control, and health status) and are related to the employment and earnings outcomes. Not surprisingly, participants differed significantly from the average candidate comparison group members across most observable characteristics before enrollment, such as education, use of SSA work incentives during the year before randomization, and employment. However, PSM selected comparison groups that were similar to participants along the entire set of variables available in the administrative data, including prior employment and earnings.

Unfortunately, even with comparison groups that are similar on average to participants along the available variables, which comprise most of the variables that the literature has found to be related to the outcomes, PSM was not able to replicate the experimentally derived impact estimates and tended to overestimate the effects. This positive bias indicates that participants who volunteered for the randomized experiments were more likely to have positive outcomes—regardless of whether they received intervention services—than the comparison groups selected through PSM. The findings are consistent with the recent literature reviewing PSM, because they suggest that PSM may not produce valid results when people self-select into a program such as the State Partnership Initiative. Rather than reflect the true effects of the projects, the PSM-based results reflect both the true effects and unmeasured or unmeasurable preenrollment differences that exist between participants and the comparison group members that affect outcomes.

The method is labor-intensive and time-consuming. While random assignment requires substantial up-front costs to design and implement, PSM required even more. PSM required obtaining data on the candidate comparison group, examining results from hundreds of statistical tests, and modifying the selection process when the tests indicated that the comparison group was not suitable. Furthermore, the nature of statistical matching means that the evaluators have to use a trial-and-error process to respecify models to produce well-matched comparison groups. It is not possible to predict how long this process will take to produce good matches. For the State Partnership Initiative evaluation, the number of iterations needed to select acceptable comparison samples for the 11 projects averaged 4.5 and ranged from 2 to 10. There may be cases when no amount of model respecification will produce comparison groups that pass the statistical tests (Agodini and Dynarski 2004). Our experience indicates that use of PSM in lieu of random assignment is costly in terms of researcher time and resources and, because the method produced incorrect estimates, the benefits were limited in this evaluation.

Larger samples than are typically available in evaluations of social programs may be required. Both the Glazerman et al. (2003) study and the Smith and Todd (2005b) rejoinder to Dehejia and Wahba (2002) warned that using PSM with a small sample may result in incorrect estimates. This may be due to experimental impacts that are not precisely estimated or from PSM estimates that are sensitive to small decisions in specifying the regression model that might otherwise have little effect in a larger sample. While there were only 35 or fewer treatment group members for the New Hampshire groups, each of the New York groups and Oklahoma contained about 300 participants for whom we attempted to find a nearest-neighbor match. Even with relatively large samples of participants and candidate comparison group members and a thoughtful matching protocol, the incorrect estimates indicate that PSM was unable to identify people who were well matched to treatment group participants. PSM may be feasible only with extremely large numbers of participants, say, over 10,000. Such large sample sizes—many times that of the State Partnership Initiative—are often unavailable in evaluations of social programs. Unfortunately, at this point, there is no evidence of success even with such large samples.

Validation of the method requires knowledge of the true impact estimates. The demonstration's evaluation had two relevant components: (1) a national evaluation, which MPR conducted using PSM; and (2) internal evaluations, conducted by each project, of which three used random assignment. This convenient overlap of approaches in three projects allowed us to use the experimental findings to conduct validity analyses and determine that the comparison groups selected through PSM incorrectly estimated impacts on employment and earnings. In practice, however, researchers will not usually have parallel randomized studies with which to perform such validity tests.

There is no way to know in advance whether the method will work. One major drawback to using PSM to evaluate social policy programs is that typically there is no way researchers can determine whether PSM will succeed in a particular evaluation *before* the decision not to use random assignment is made. While the literature is useful in describing the conditions that should be present for PSM to be successful, meeting these conditions is not enough to ensure success, as the State Partnership Initiative demonstrates. For example, the Dehejia and Wahba papers suggest that observing more than one year of pretreatment earnings information is necessary. The State Partnership Initiative evaluation had five years of preenrollment earnings information and monthly observations going back two years for a variety of other pretreatment variables. Furthermore, both Michalopoulos et al. (2004) and Glazerman et al. (2003) outlined conditions that suggest that bias can be reduced by drawing comparison group members from the same labor market. For all three projects in our evaluation, we chose from within comparison areas we selected on the basis of having similar labor market and service characteristics. Even with a wealth of preprogram data and use of similar labor markets, PSM was unable to replicate the experimentally derived impact estimates, which offers a solid counterexample to the conditions Dehejia and Wahba claimed were needed for PSM to work in practice.

Additional analyses can aid in validating a comparison group selected using PSM, but these are not designed to be used before the evaluation, at least until comparison groups are selected and pass statistical tests, nor are they enough to ascertain with confidence that the process selected a comparison group that will estimate the correct impact. One approach, pioneered by Heckman and Hotz (1989), is to select a comparison group for participants several periods before enrollment and then examine the outcomes during the period after this earlier data but before enrollment into the program. If the comparison group is well matched, the impact estimate during the period before enrollment should equal zero, because neither participants nor comparison group members received any program services.

While tools such as these—reselecting comparison groups using an earlier enrollment date—may help explain any under- or overestimate in the impact estimates, a different specification is typically required for each treatment group-comparison group combination. Furthermore, without conducting an experiment in parallel, one can never know if the PSM process was ultimately successful.

Summary. Our experience evaluating the State Partnership Initiative evaluation showed that anyone interested in applying PSM to evaluate other real-life programs should proceed with caution. PSM implemented under seemingly ideal conditions produced estimates that differed considerably from the experimentally derived estimates. If researchers have to use PSM, we would advise them to at least consider an experimental design in one or more sites to test the method's validity. In our case, we benefited from having experimental results available from three states to conduct this assessment, but this was a rare confluence of events.

[Received April 2006. Revised March 2008.]

REFERENCES

Agodini, R., and Dynarski, M. (2004), "Are Experiments the Only Option? A Look at Dropout Prevention Programs," *Review of Economics and Statistics*, 86, 1, 180–194.

Agodini, R., Thornton, C., Khan, N., and Peikes, D. (2002), "Design for Estimating the Net Outcomes of the State Partnership Initiative," report submitted to the Social Security Administration, Princeton, NJ: Mathematica Policy Research, Inc.

Ashenfelter, O. (1978), "Estimating the Effect of Training Programs on Earnings," *Review of Economics and Statistics,* 60, 47-57.

——— (1987), "The Case for Evaluating Training Programs with Randomized Trials," *Economics of Education Review*, 6, 4, 333–338.

Dehejia, R. H., and Wahba, S. (1999), "Causal Effects in Nonexperimental Studies: Re-Evaluating the Evaluation of Training Programs," *Journal of the American Statistical Association*, 94, 1053–1063.

——— (2002), “Propensity Score-Matching Methods for Nonexperimental Causal Studies,” *Review of Economics and Statistics*, 84, 1, 151–161.

Fraker, T., and Maynard, R. (1987), “The Adequacy of Comparison Group Designs for Evaluations of Employment-Related Programs,” *Journal of Human Resources*, 22, 2, 194–227.

Friedlander, D., and Robins, P. K. (1995), “Estimating the Effects of Employment and Training Programs: An Assessment of Some Nonexperimental Techniques,” New York: Manpower Demonstration Research Corporation.

Glazerman, S., Levy, D. M., and Myers, D. (2003), “Nonexperimental Versus Experimental Estimates of Earnings Impacts,” *Annals of the American Academy of Political and Social Science*, 589, 63-93.

Heckman, J. (1992), “Randomization and Social Policy Evaluation,” in *Evaluating Welfare and Training Programs*, eds. C. Manski and I. Garfinkel, Cambridge, MA: Harvard University.

Heckman, J., and Hotz, V. J. (1989), “Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case for Manpower Training,” *Journal of the American Statistical Association*, 84, 862–880.

Heckman, J., Ichimura, H., and Todd, P. (1998a), “Matching as an Econometric Evaluation Estimator,” *Review of Economic Studies*, 65, 261–294.

Heckman, J., Ichimura, H., Smith, J., and Todd, P. (1998b), “Characterizing Selection Bias Using Experimental Data,” *Econometrica*, 66, 2, 5, 1017–1099.

LaLonde, R. (1986), “Evaluating the Econometric Evaluations of Training Programs with Experimental Data,” *American Economic Review*, 76, 4, 604–620.

Michalopoulos, C., Bloom, H. S., and Hill, C. (2004), “Can Propensity Score Methods Match the Findings from a Random Assignment Evaluation of Mandatory Welfare-to-Work Programs?” *Review of Economics and Statistics,* 86, 1, 156–175.

Orr, L. L. (1999), *Social Experiments: Evaluating Public Programs with Experimental Methods*, Thousand Oaks, CA: Sage.

Peikes, D., Orzol, S., Moreno, L., and Paxton, N. (2005), “State Partnership Initiative: Selection of Comparison Groups for the Evaluation and Selected Impact Estimates,” Princeton, NJ: Mathematica Policy Research, Inc.

Pickett, C. (2003), *SSI Disabled Recipients Who Work*, Washington, DC: Social Security Administration.

Rosenbaum, P. R., and Rubin, D. B. (1983), “The Central Role of the Propensity Score in Observational Studies for Causal Effects,” *Biometrika,* 70, 41–55.

Smith, J., and Todd, P. (2005a), “Does Matching Overcome LaLonde’s Critique of Nonexperimental Estimators?” *Journal of Econometrics*, 125, 305–353.

——— (2005b), “Rejoinder,” *Journal of Econometrics*, 125, 365–375.

Wilde, E. T., and Hollister, R. (2007), “How Close Is Close Enough? Evaluating Propensity Score Matching Using Data from a Class Size Reduction Experiment,” *Journal of Policy Analysis and Management*, 26, 3, 455–477.