USING STATE ADMINISTRATIVE DATA TO MEASURE PROGRAM PERFORMANCE

Peter R. Mueser, Kenneth R. Troske, and Alexey Gorislavsky*

Abstract—We use administrative data from Missouri to examine the sensitivity of earnings impact estimates for a job training program based on alternative nonexperimental methods. We consider regression adjustment, Mahalanobis distance matching, and various methods using propensity-score matching, examining both cross-sectional estimates and difference-in-difference estimates. Specification tests suggest that the difference-in-difference estimator may provide a better measure of program impact. We find that propensity-score matching is most effective, but the detailed implementation is not of critical importance. Our analyses demonstrate that existing data can be used to obtain useful estimates of program impact.

I. Introduction

THERE has been growing interest on the part of governments in evaluating the efficacy of various programs designed to aid individuals and businesses. For example, state legislatures in California, Illinois, Massachusetts, Oregon, and Texas have all mandated that some type of evaluation of state welfare programs be undertaken. In addition, the federal government has required that federally funded training and employment programs administered at the state and local levels meet standards based on participant employment outcomes.

However, the best way for states to conduct evaluations remains an unanswered question. Early efforts to evaluate the effects of government-sponsored training programs such as the Manpower Development Training Act (MDTA) or the Comprehensive Employment Training Act (CETA) focused on choosing the appropriate specification of the model in the presence of nonrandom selection on unobservables by participants in the program (Ashenfelter, 1978; Bassi, 1984; Ashenfelter & Card, 1985; Barnow, 1987; Card & Sullivan, 1988). This research culminated in the papers by LaLonde (1986) and Fraker and Maynard (1987), which concluded that nonexperimental evaluations had the potential for severe specification error. This led both researchers and policymakers to argue that the only appropriate way to evaluate government-sponsored training and education programs is through the use of randomized social experiments.

However, recent critiques of social experiments (Heckman & Smith, 1995; Heckman, LaLonde, & Smith, 1999) argue that even randomized experiments have important shortcomings that limit their usefulness in policymaking. In both the planning and the implementation stages, random assignment designs are often resisted by program staff, due to ethical and administrative concerns with denying program access to eligible individuals, as well as fears that a poor evaluation could adversely affect program resources.[1] Even well-designed experiments require substantial efforts to assure that staff administer them as planned, producing participant and control groups that are truly randomly assigned. Finally, estimates of impact based on social experiments are not always directly relevant for policymakers in deciding whether to create new programs or to expand existing ones (see also Manski, 1996).

Based in part on these concerns, recent research has returned to earlier work on nonexperimental methods (Rosenbaum & Rubin, 1983; Heckman & Hotz, 1989; Friedlander & Robins, 1995). James Heckman and colleagues have published a set of papers (Heckman, Ichimura, & Todd, 1997, 1998; Heckman et al., 1998) using data from random assignment experiments to identify those strategies that may be successful in estimating program impacts from nonexperimental data. They stress that there is no magic methodology that will always produce unbiased and useful estimates of program impacts. Program evaluation requires researchers to first adopt a methodology that is suitable for the question they want to address, and second, to perform appropriate specification tests, and, finally, to use data that are appropriate for estimating the parameters of interest in a given institutional context. The results from these papers also suggest that, with suitable data on both program participants and nonparticipants, it may be possible to provide meaningful estimates of program impact.

Matching methods have been a primary focus of attention in recent years. Although matching in its simplest form involves pairing each participant or “treated” case with a comparison case that is similar, based on measured characteristics, more general matching methods apply a weight to each case in the comparison sample, estimating program effects as the difference in outcome between participants and the weighted comparison sample. Matching attempts to assure that the participants and weighted comparison samples are comparable in terms of all measured characteristics,

Received for publication January 16, 2004. Revision accepted for publication September 18, 2006.

* University of Missouri–Columbia and IZA; University of Kentucky–Lexington and IZA; and University of Missouri–Columbia, respectively.

Much of the work on this paper was done while Troske was a visiting research fellow at the Institute for the Study of Labor (IZA) in Bonn, Germany, and he would like to thank them for financial support. We would also like to thank seminar participants at the Centre for European Economic Research (ZEW) in Mannheim, European University Institute, Frisch Centre for Economic Research, Institute for Advanced Studies in Vienna, Oxford University, Tinbergen Institute, University College-Dublin, University of Illinois, University of Oklahoma, University of Zurich, and at the CERP/IZA-sponsored conference, “Improving Labour Market Performance: The Need for Evaluation,” Bonn, Germany, for comments. We are especially grateful to Jeff Smith for extensive and helpful comments on earlier drafts of this paper.

[1] The random assignment design for the Job Training Partnership Act (JTPA) experiment carried out in the 1980s was rejected by managers of the majority of sites asked to participate. The sites ultimately included in the study were in large part self-selected. For a discussion of the issues raised by this refusal, see Heckman and Smith (1995).

The Review of Economics and Statistics, November 2007, 89(4): 761–783

an approach conceptually distinct from methods, like regression, that are based on fitting models that use individual characteristics to predict outcomes.

Recent work by Dehejia and Wahba (1999, 2002) is supportive of the view that matching methods can be used to obtain unbiased program estimates. They reanalyze LaLonde's (1986) data using matching methods, arguing that despite limited information on individuals and use of a comparison group that differs in important ways from the treatment group, unbiased estimates of program impact can be obtained. However, their methods are questioned by Smith and Todd (2005a, 2005b), who take strong exception to these conclusions (see also Zhao, 2003).

A growing literature now applies matching methods in various evaluation environments, but few papers provide a comprehensive comparison of alternative approaches.[2] Our work provides such comparisons, investigating the possibility of evaluating government-sponsored training programs using existing data sources. If such methods are reliable, there are tremendous opportunities for evaluating programs because most states already possess rich data sets on participants in various state programs that are used to administer these programs, as well as data on earnings for almost all workers in the state. Thus, it may be possible to evaluate government training programs without resorting to expensive experimental evaluations (or searching for instruments), producing estimates of program impacts that are useful for policymakers.

The goal of this paper is to use administrative data from one state, Missouri, to examine the sensitivity of estimates of program impacts across alternative evaluation methods and alternative outcome variables. We also examine the sensitivity of our results to the quality of the data available for analysis. We assess the estimates from different methods by comparing them to each other, and also by comparing them to estimates of program impacts based on experimental methods that have been reported in the literature. In addition, we conduct a number of specification checks of our evaluation methods. The methods we consider are simple difference, regression analysis, matching based on the Mahalanobis distance, and matching based on the propensity score. For propensity-score matching we also consider a number of alternative ways to match participants with nonparticipants such as pairwise matching, pairwise matching with various calipers, matching with and without replacement, matching using propensity-score categories, and kernel density matching. Finally, for each method we present both cross-sectional and difference-in-difference estimates.

The program we examine is Missouri's implementation of job training programs under the federal Job Training Partnership Act (JTPA). Our data on participants come from information collected by the state of Missouri to administer this program. Our comparison group consists of individuals registered with the state's Division of Employment Security (ES) for job exchange services. Our data on earnings and employment history come from the Unemployment Insurance (UI) program in the state. These data have a number of features that make them ideal for use in evaluating government programs. First, they contain very detailed location information allowing us to compare individuals in the same local labor market. Second, they allow us to identify individuals in our comparison group who are currently participating or who have recently participated in the JTPA program. Thus, we can avoid the problem of contamination bias, which occurs when individuals in the comparison group are participants in the program being evaluated.[3] Finally, the data on wages and employment history are being generated by the same process for both participants and nonparticipants. Results in Heckman et al. (1998) indicate that these factors are critical in constructing an appropriate nonrandom comparison group.

Although federal legislation passed in 1998 replaced the JTPA program with the Workforce Investment Act (WIA), the new program has much in common with the JTPA. For that reason, results obtained using the JTPA—both impact estimates and conclusions about appropriate evaluation methods—have important implications for WIA. In addition, the data we have from Missouri are similar to administrative data collected by other states in implementing various workforce development and UI programs, so it should be possible to use the results from our study when conducting evaluations of other states' programs. A study supported by the U.S. Department of Labor undertakes evaluation of the Workforce Investment Act in seven states using methods closely related to those examined here.[4]

Our specification tests suggest that, when we use the difference-in-difference estimator, we are constructing comparison groups that are very similar to our participant group in terms of earnings growth, meaning we are comparing individuals who are comparable on relevant dimensions. In addition, we find that our estimates are insensitive to the method used for constructing comparison groups, providing some confidence in the robustness of our results. Finally, we find that our estimates of the impact of the JTPA program on earnings are similar to previous estimates of the effect of JTPA based on data from randomized experiments (Orr et al., 1996). While certainly not definitive, these results do suggest that it is possible to evaluate government job training programs using administrative data that are currently being collected by most state governments.

[2] The February 2004 issue of *The Review of Economics and Statistics* published papers in a symposium on matching methods. Of particular relevance to our work here are papers by Frölich, Imbens, and Zhao, which we discuss in detail in the next section.

[3] We do not know whether individuals in our comparison sample are participating in other government-sponsored training programs or private training programs. Therefore, there could be other sources of contamination bias.

[4] The work is part of the Administrative Data Research and Evaluation (ADARE) project. Evaluation results appear in Hollenbeck et al. (2004).

The remainder of the paper is as follows. In the next section we discuss the various methods we use to construct our nonexperimental comparison groups. Section III contains a discussion of our data. Section IV presents our main results. In section V we examine the sensitivity of our results to the quality of the data used in the analysis. Section VI concludes.

II. Estimating Program Effects Based on Conditional Independence

Our goal is to estimate the effect of participating in the JTPA program on program participants. Let Y_1 be earnings for an individual following participation in the program and Y_0 be earnings for that individual in the absence of participation. It is impossible to observe both measures for a single individual. If we define $D = 1$ for those who participate and $D = 0$ for those who do not participate, the outcome we observe for an individual is

$$Y = (1 - D)Y_0 + DY_1.$$

Experimental evaluations employ random assignment to the program, assuring that the treatment is independent of Y_0 and Y_1 and the factors influencing them. The average program effect for individuals subject to random assignment may be estimated as the simple difference in outcomes for those assigned to treatment and those assigned to the control group. Where D is not independent of factors influencing Y, participants may differ from nonparticipants in many ways, including the effect of the program, so the simple difference in outcomes between participants and nonparticipants need not identify program impact for any definable group.

If we assume that, conditional on measured characteristics, X, participation is independent of the outcome that would occur in the absence of participation,

$$Y_0 \perp\!\!\!\perp D | X, \tag{1}$$

the effect of the program on participants conditional on X can be written as

$$\begin{aligned} E(Y_1 - Y_0 | D = 1, X) &= E(\Delta Y | D = 1, X) \\ &= E(Y_1 | D = 1, X) - E(Y_0 | D = 0, X), \end{aligned} \tag{2}$$

where $Y_1 - Y_0 = \Delta Y$ is understood to be the program effect for a given individual and the expectation is across all participants with given characteristics. Matching and regression adjustment methods are all based on some version of assumption (1). They differ in the methods used to obtain estimates of $E(Y_1 | D = 1, X)$ and $E(Y_0 | D = 0, X)$.[5,6]

A. Simple Regression Adjustment

Until recently, the most common approach (for example, Barnow, Cain, & Goldberger, 1980) to estimating program impact was based on a simple linear specification assuming that the earnings function was the same for participants and the comparison group. Program impact δ was estimated, along with a vector of parameters of the linear earnings function, β, by fitting the equation

$$Y = X\beta + \delta D + e$$

with e, an error term independent of X and D. Although this approach can be pursued using more flexible functional forms, estimates of program impact rely on a parametric structure in order to compare participants and nonparticipants.

The critical question for regression adjustment is whether the functional form properly predicts what post-program wages would be for participants if they had not participated. Even under the maintained assumption in equation (1) that outcomes for participants and the comparison group do not differ once observed characteristics are controlled, if most of the comparison sample has characteristics that are quite distinct from those of the participants, regression adjustment will be predicting outcomes for participants by extrapolation. If the functional relationships differ by values of X, the regression function may be poorly estimated, resulting in a potential bias.

B. Matching Methods[7]

Methods that focus more explicitly on matching by X are designed to ensure that estimates are based on outcome differences between comparable individuals. Where the set of relevant X variables is small and each has a very limited number of observed values, it may be possible to estimate the terms on the right side of equation (2) for each distinct combination of characteristics. In most cases, there are too many observed values of X to make such an approach feasible.

[5] Where concern focuses on program impact for nonparticipants or other subgroups, a stronger assumption than equation (1) is required. Normally, it is assumed that, conditional on X, both Y_0 and Y_1 are independent of participation. This assumption will generally be violated if individuals select the treatment partly on the basis of unmeasured factors associated with expected benefits.

[6] Although it is convenient to explicate estimation techniques in terms of a single population from which a subgroup receives the treatment, in practice treatment and comparison groups are often separately selected. The combined sample is therefore "choice based," and conditional probabilities calculated from the combined sample do not reflect the actual probabilities that individuals with given characteristics face the treatment in the original universe. However, the methods used here can apply under choice-based sampling. In particular, if assumption (1) applies in the population from which the treatment and comparison groups are drawn, assumption (1) will also apply (in the probability limit) in the choice-based sample where the probability of inclusion differs for treated and untreated individuals but is otherwise unrelated to individual characteristics. The methods outlined here can be shown to be consistent for such a sample.

[7] See Rosenbaum (2002) and Imbens (2004) for general discussions of matching methods.

A natural alternative is to compare cases that are "close" in terms of *X*. Several matching approaches are possible. In the analysis here, we will first consider nearest-neighbor pair matching, in which each participant is matched with one individual in the comparison group, and where no comparison case is used for more than one match. We also consider variations on this basic matching technique. We then turn to methods based on grouping cases with similar measured characteristics. All these methods can be interpreted as schemes to weight the comparison sample so it will replicate the distribution of participants on the *X*.

C. *Mahalanobis Distance Matching*

We first undertake pair matching according to Mahalanobis distance. If we specify X' as the vector of observed values for a participant and X'' for a comparison individual, the distance between them is calculated as

$$M(X',X'') = (X' - X'')^T V^{-1}(X' - X''),$$

where V is the covariance matrix for X. The Mahalanobis distance has the property that matching will reduce differences between groups by an equal percentage for each variable in X, assuming that V is the same for the two groups.[8] This ensures that the difference between the two groups in any linear function will be reduced (Rosenbaum & Rubin, 1985). Friedlander and Robins (1995) illustrate the use of Mahalanobis distance in program evaluation.

The simplest pair matching approach begins by ordering participants and the comparison group members randomly. The first participant is matched to the comparison group member that minimizes $M(X', X'')$. The matched comparison group member is then eliminated from the set, and the second participant is matched to the remaining comparison group member that minimizes $M(X', X'')$. The process continues through all participants until the participant or comparison group is exhausted.[9]

Matches produced by such an approach are not invariant to the order in which the data are sorted prior to matching. An alternative approach, optimal full matching, requires searching across all possible sets of matches to find the set that minimizes the *sum* of all distances (Rosenbaum, 2002) or a related systemwide criterion (Hansen, 2004). Rather than perform such optimal matching, we present results of a matching approach based on one-by-one comparisons across matched pairs. In this "modified" matching procedure, we not only compare the distance between the participant and all comparison group members but also compare the distance for all members of the comparison group who were previously matched to participants. A prior match is broken and a new match formed if $M(X', X'')$ from the new match is smaller than that of the previous match. The participant in the broken match is then rematched, in accord with the same procedure. Under this procedure, the results are invariant to the ordering of the data.[10]

Of course, if the comparison group contains sufficient numbers of cases with very similar values on all X to those among participants, the matching procedure will produce directly comparable groups. In most cases, however, there remain substantial differences between cases for some matched pairs. We try to account for this in two ways. First, we examine the impact of additional regression adjustment on estimates of program impact. Second, we drop the 1% of the matches with the largest distance.

D. *Propensity-Score Matching*

In the combined sample of participants and comparison group members, let $P(X)$ be the probability that an individual with characteristics X is a participant. Rosenbaum and Rubin (1983) show that

$$Y_0 \perp\!\!\!\perp D|X \Rightarrow Y_0 \perp\!\!\!\perp D|P(X).$$

This means that if we consider participant and comparison group members with the same $P(X)$, the distribution of X across these groups will be the same. Based on this "propensity score," the matching problem is reduced to a single dimension. Rather than attempting to match on all values of X, we can compare cases on the basis of propensity scores alone. In particular,

$$E(\Delta Y|P) = E_P(E(\Delta Y|X)),$$

where E_P indicates the expectation across values of X for which $P(X) = P$ in the combined sample. This implies that

$$E(\Delta Y|D = 1) = E_{X|D=1}(\Delta Y|P(X)),$$

where $E_{X|D=1}$ is the expectation across all values of X for participants. The propensity score is thus a balancing score for X, assuring that for a given value of the propensity score, the distribution of X will be the same for participants and comparison cases. Although other balancing scores could serve the same function, estimating the propensity score is particularly convenient.

Propensity-score matching is now the dominant approach in analyses using matching. In addition to reducing the matching problem to a single dimension, the propensity score facilitates investigation of whether the treatment and

[8] In practice one must estimate V using the sample of either participants or nonparticipants or using a weighted average of the covariance matrices from the two groups. We follow most of the previous literature in estimating V as a weighted average of the covariance matrices for participants and nonparticipants with the weights being the proportion of each group in the data. Calculating V in this manner minimizes sampling error.

[9] See Rosenbaum and Rubin (1985) and Rosenbaum (2002), who refer to this as "greedy matching."

[10] We implemented the conventional and modified matching for Mahalanobis distance and propensity scores with programs we wrote in C+ and OX.

comparison groups have sufficient overlap to allow meaningful comparison. We report such comparisons in our analyses.

We estimate $P(X)$ using a logit specification with a highly flexible functional form allowing for nonlinear effects and interactions. We also test to assure that the score in fact balances our independent variables by testing for statistically significant differences between variable means within propensity-score bands. We first undertake one-to-one matching based on the propensity score using the methods described in the previous subsection. We also use calipier matching, where we remove matches for which the difference in propensity scores between matched pairs exceeds some threshold or caliper. In the analysis we report results based on calipers ranging from 0.01 to 0.2.[11]

These simple matching estimators can be conceptualized as based on a weighting function where each comparison case receives a weight of 1 if chosen and 0 otherwise. Recent attention has focused on alternative weighting schemes that may be viewed as generalizations of one-to-one matching. In addition to considering many-to-one matching, in which participants are matched with more than one comparison case, we consider two general matching or weighting functions, matching by propensity-score category, and kernel density matching.

First consider matching by propensity-score category or stratum. Let the kth stratum be defined to include all cases with values of X such that $P(X) \in [P_1{}^k, P_2{}^k)$. Let $N_k{}^1$ be the number of participants within the kth stratum, $N_k{}^0$ the number of individuals in the comparison group within the kth stratum, and N the total number of participants in our sample. Our estimate of the treatment effect within stratum k is given by

$$E_k(\Delta Y) \;=\; E(\Delta Y | P(X) \in [P_1^K, P_2^K)) = \sum_{i=1}^{N_k^1} \frac{1}{N_k^1} Y_{i1} - \sum_{j=1}^{N_k^0} \frac{1}{N_k^0} Y_{j0}. \tag{3}$$

Our estimated average treatment effect across all strata is then given by

$$E(\Delta Y) = \sum_k \frac{N_k^1}{N} \times E_k(\Delta Y). \tag{4}$$

In choosing P_1^k and P_2^k we follow the algorithm outlined in Dehejia and Wahba (2002).[12] In particular, we choose P_1^k and P_2^k such that remaining differences in X between participants and nonparticipants within each stratum are likely due to chance.[13]

Our second approach is the kernel matching procedure described in Heckman, Ichimura, and Todd (1997) and Heckman et al. (1998). The kernel matching estimator is given by

$$E(\Delta Y) = \frac{1}{N} \sum_{i \in T} \left[Y_{i,1} - \frac{\displaystyle\sum_{j=1}^{N_1^C} Y_{j,0}^i K\left(\frac{P(X_{j,0}^i) - P(X_{i,1})}{b_w}\right)}{\displaystyle\sum_{j=1}^{N_i^C} K\left(\frac{P(X_{j,0}^i) - P(X_{i,1})}{b_w}\right)} \right],$$

where T is the set of cases receiving the treatment and N is the number of treated cases; $Y_{i,1}$ and $X_{i,1}$ are dependent and independent variables for the ith treated case; $Y_{j,0}^i$ and $X_{j,0}^i$ are dependent and independent variables for the jth comparison case that is within the neighborhood of treatment case i, that is, for which $|P(X_{j,0}^i) - P(X_{i,1})| < b_w/2$; N_i^C is the number of comparison cases within the neighborhood of i; $K(\bullet)$ is a kernel function; and b_w is a bandwidth parameter. In general, a kernel is simply some density function. In practice, the choice of $K(\bullet)$ and b_w is somewhat arbitrary. In our analysis we use cross-validation methods to choose the bandwidth and kernel.

Whatever matching algorithm is used, there will generally be some difference in the conditioning variables—or the propensity score—between participant and comparison cases. Although earlier work occasionally employed local linear regression adjustments to address this issue (Heckman et al., 1998), adjustments based on a single linear model have now become common following recent theoretical work by Abadie and Imbens (2006a) arguing that such an approach removes an asymptotic bias inherent in simple matching estimators. Although the discussion in Abadie and Imbens suggests that bias adjustment is likely to be less important in large samples like ours, we report below how such regression adjustment alters results.

A variety of alternative matching estimators have been proposed in recent years. Imbens (2000), for example, proposed an estimator in which comparison cases are weighted by the propensity-score ratio. In a Monte Carlo analysis, Frölich (2004) compared one-to-one matching

[11] As noted above, we wrote our own programs to undertake the matching, and the methods we use differ slightly from those used by psmatch2, a matching program written in STATA (Leuven & Sianesi, 2003). The psmatch2 program sorts both participants and the comparison group by propensity score (either ascending or descending) prior to undertaking the matching. We found that this approach produced somewhat different estimates than ours, although most were within the range of values produced by random sorting.

[12] Subclassification on the propensity score is also examined in Rosenbaum and Rubin (1984), and general issues of subclassification are considered in Cochran (1968).

[13] Although we have chosen to present equation (3) in such a way as to highlight the symmetrical contribution of treatment and comparison cases in the estimation, the average treatment effect specified in equation (4) is numerically identical to that where the mean for all comparison cases within the specified stratum is taken as the comparison outcome for each treatment case in that stratum. It therefore corresponds to the approach used by Dehejia and Wahba (2002).

with estimators based on kernel weighting, local linear regression, local linear ridge regression (designed to assure stability of linear regression adjustments in small samples), and a weight based on the propensity-score ratio. The weight based on the propensity-score ratio performed very poorly relative to the others. The ridge estimator performed best, although Frölich noted that where the sample size was large—as in our case—the kernel estimator's performance was similar.

In terms of theory, Angrist and Hahn (1999) argue that there is little basis for preferring propensity-score matching to Mahalanobis matching or to other metrics. Hirano, Imbens, and Ridder (2003) show that weighting by the inverse of a nonparametric estimate of the propensity score will produce an efficient impact estimate. As far back as Rosenbaum and Rubin (1985), it has been suggested that matching be undertaken on the propensity score combined with other variables. Zhao (2004) has proposed alternative metrics that weight variables by coefficients reflecting their impact on outcome measures as well as propensity score. Such an approach improves matches for those variables that could cause outcomes to differ for treatment and comparison cases. His Monte Carlo experiments suggest that in small samples his alternative metrics may have benefits, but there is no approach that dominates all the others across environments. His work examines only one-to-one matching, and it considers treatment samples that are appreciably smaller than those we use here.

Rather than using the propensity score to match cases, Smith and Todd (2005a) match on the odds ratio of the propensity score ($p/(1-p)$). This approach has the advantage that results will be invariant to choice-based sampling. Since the odds ratio is a continuous transformation of the propensity score in the relevant range, where participant and comparison cases are matched closely, results using this approach should be very similar to methods that use propensity score.

Given that the performance of all of these alternative estimators appears to be similar to the performance of the more standard estimators with samples as large as we use, and given that all of these estimators are much less common, we have chosen not to focus on these alternative matching estimators.

E. *Comparing Comparable Cases: Common Support Conditions*

In order to obtain an estimate of the treatment effect on the treated, there must be some case in the comparison sample that is "close" to each treated case. Propensity-score methods have the advantage that they allow this common support problem to be reduced to a single dimension. Subject to the assumptions inherent in the estimation of the propensity score, if propensity-score values for the comparison cases are sufficiently dense in the neighborhood of each treated case, it will be possible to find an acceptable match. Heckman, Ichimura, and Todd (1997) illustrate that failure of overlap may play a particularly important role in biasing nonexperimental estimators. Black and Smith (2004) show in an analysis of college quality that estimation precision is dramatically reduced when the support condition is only weakly met (see also Dearden, Ferri, and Meghir, 2002). We report below the distribution of propensity scores for the treatment and comparison samples.

F. *Additional Issues in Implementing Matching*

There are a number of additional choices about how one actually forms a matched sample, such as the choice of whether to match with or without replacement, the choice of the number of nearest neighbors, the use of a caliper when matching, and the size of the strata or bandwidth, that warrant further discussion. The choice among these various options often involves a tradeoff between bias and efficiency. For example, matching with replacement will, in general, produce closer matches than matching without replacement and therefore will result in estimates with less bias. However, matching with replacement can also increase sampling error because an individual in the comparison group can be matched to more than one individual in the treatment group. Similar tradeoffs exist in deciding how many comparison cases to match with a given treatment case. Using a single comparison case minimizes bias, while using multiple comparison cases can improve precision. A caliper has the effect of omitting comparison cases that are poorly matched, which may reduce bias, but when treatment cases are omitted because a sufficiently close match is not found, the estimated treatment effect applies only for the included cases. Where impacts vary across individuals, estimates no longer capture the true average treatment effect. Which methods are appropriate depends on the overlap in the matching variables for the treatment and the comparison samples, the relative sample sizes, and the quality of the data.

To assess the effects that these choices have on our estimates, we present results based on a variety of matching methods. In particular, we present results based on matching with and without replacement; matching to one, five, and ten nearest neighbors; and matching using several different calipers. In addition, as indicated above, we present estimates based on standard matching without replacement, where the match is dependent on the order of the data, as well as estimates based on modified matching, where the matches are independent of the order of the data.

For each of the alternative methods for estimating the effect of the program, we construct two estimators, a cross-sectional estimator and a difference-in-difference estimator. The cross-sectional estimator is based on the difference in post-program earnings between the treatment and comparison samples. The difference-in-difference estimator is based on the difference for the comparison and treatment samples in the difference between pre- and post-program earnings. The advantage of the difference-in-difference estimator is that it allows one to control for any unobserved fixed individual factors that may affect program participation and earnings. Therefore, the difference-in-difference estimator is more likely to meet the assumption underlying matching that the determinants of program participation are independent of the outcome measure once observable characteristics are accounted for. The disadvantage of the difference-in-difference estimator, particularly in this setting, is that if there are any transitory shocks to pre-program earnings that affect program participation, this could bias the difference-in-difference estimator. Problems of estimating program effects in the presence of the now famous "Ashenfelter dip," where mean earnings for participants fall shortly before participation, illustrates the potential bias. Since the Ashenfelter dip is a transitory decline in earnings, later earnings are expected to increase even in the absence of intervention. If pre-program earnings are measured during the Ashenfelter dip, the difference-in-difference estimator will produce an upward biased estimate of the program's impact. Therefore, when measuring pre-program earnings we will try to do so prior to the onset of the Ashenfelter dip.

G. *Matching Variables*

The assumption that outcomes are independent of the treatment once we control for measured characteristics depends critically on the particular measured characteristics available. Any characteristic that is associated with both program participation and the outcome measure for nonparticipants, after conditioning on measured characteristics, can induce bias. It has long been recognized that controls for the standard demographic characteristics such as age, education, and race are critical. Labor market experience of the individual is also clearly relevant. Where program eligibility is limited, factors influencing eligibility have usually been included as well. LaLonde (1986) includes controls for age, education, race, employment status, prior earnings, and residency in a large metropolitan area, as well as prior year receipt of Aid to Families with Dependent Children (AFDC) and marital status, measures associated with eligibility in the program.

Several recent analyses (Friedlander & Robins, 1995; Heckman & Smith, 1999) have stressed the importance of choosing a comparison group in the same labor market. Since it is almost impossible to choose comparison groups in the same labor market as participants when drawing comparison groups from national samples, approaches that use such data are unlikely to produce good estimates, even if they are well matched on other individual characteristics. There is also a growing recognition that the details of the labor market experiences of individuals in the period immediately prior to program participation are critical. In particular, movements into and out of the labor force and between employment and unemployment in the eighteen months prior to program participation are strongly associated with both program participation and expected labor market outcomes (Heckman, Ichimura, & Todd, 1997; Heckman et al., 1998; Heckman, LaLonde, & Smith, 1999; Heckman & Smith, 1999).

Finally, Heckman, Ichimura, and Todd (1997) have argued that where outcome measures for participants and comparison group members come from different sources—as for example, where different types of surveys are used—this often induces serious bias, as systematic differences in the measures are incorporated into impact estimates.

III. The Data

This project uses administrative data deriving from three sources. We draw our sample of program participants from records of Missouri's JTPA program. We draw our comparison group sample from job exchange service records maintained by Missouri's Division of Employment Security (ES). The earnings data source is wage record data maintained as part of the Unemployment Insurance programs in Missouri and Kansas. Using these data we obtain both pre- and postenrollment earnings and information on employment status prior to enrollment for both participants and nonparticipants.

The JTPA data comprise all individuals who apply to and then enroll in the JTPA program. The data include basic demographic and income information collected at the time of application that is used to assess eligibility, as well as information about any subsequent services received. Our initial sample consists of all applicants in program years 1994 (July 1994 through June 1995) and 1995 (July 1995 through June 1996) who are at least 22 years old and less than 65 and who subsequently enroll in the Title IIa program. We focus on participants 22 years old and older because younger individuals are eligible for the youth program, which is governed by a different set of rules. Participants in Title IIa are eligible to participate in the JTPA program because they are judged to be economically disadvantaged.[14] We focus on these participants because they are a fairly homogeneous group and because they have been the focus of previous evaluations of JTPA using experimental data (for example, Orr et al., 1996). Finally, we eliminate records with invalid values for our demographic variables (race, sex, veteran status, education, and employment status).[15] Our final sample consists of 2,802 males and 6,393 females.

Our Employment Security (ES) data include all individuals who registered with the ES employment exchange

[14] JTPA also serves Title III participants, who are eligible for the program because they were displaced from their previous jobs. See Devine and Heckman (1996) for a discussion of the JTPA eligibility criteria.

[15] We eliminate around 10% of the original sample because of invalid or missing demographic variables.

service in program years 1994 and 1995. With some exceptions, individuals who receive Unemployment Insurance payments in Missouri were required to register for ES services during this period, although it is not clear how strictly this requirement was enforced. In general, ES services were not very intensive. Assistance could take a variety of forms, such as providing access to a list of job openings in an area, helping individuals prepare résumés, referring individuals to jobs, or referring individuals to other agencies for more extensive services. All residents of Missouri were eligible to receive the basic ES services such as access to the list of job openings or assistance in preparing a résumé. During the time of our sample almost every individual who wanted to obtain ES services registered at one of the Employment Security offices located around the state.[16] The ES data contain basic demographic and income information obtained on the initial application, as well as information about subsequent services received.

When selecting our ES sample, we chose individuals who were at least 22 and less than 65 years old and were deemed economically disadvantaged. Since the ES program used the same criteria to determine whether someone was economically disadvantaged as the JTPA program, all of our ES participants should be eligible to participate in the JTPA program. In addition to these criteria, we also chose ES participants who were not enrolled in JTPA in the program year. We further eliminated records with missing or invalid demographic variables.[17] Our final sample consists of 45,339 males and 52,895 females.

The pre-enrollment and post-enrollment earnings for both our JTPA and ES samples come from the Unemployment Insurance (UI) "wage record" data. These data consist of quarterly files containing earnings for all individuals in Missouri and Kansas employed in jobs covered by the UI system.[18] Both the JTPA and ES data are matched to the UI data using Social Security number. If we are unable to match an SSN to earnings data in a quarter, we considered the individual not employed in that quarter and set earnings equal to zero.

Using these earnings data, we determined total quarterly earnings from all employers for individuals in the eight quarters prior to participation, in the quarter they begin participation, and in the subsequent eight quarters. For our cross-sectional estimator, we use post-program earnings measured as the sum of earnings in the fifth through the eighth quarters after the initial quarter of participation. For our difference-in-difference estimator, we measure the difference between the sum of earnings in the fifth through the eighth quarters after the initial participation quarter and the fifth through the eighth quarters prior to the initial quarter of participation. As Ashenfelter and Card (1985) note, taking differences for periods symmetric around the enrollment quarter assures that the difference-in-difference estimator is valid in the case where there is autocorrelation in the transitory component of earnings. In order to capture the dynamics of earnings immediately prior to participation, we also control for earnings in the first through the fourth quarters prior to the initial quarter of participation.

As noted above, the dynamic patterns of an individual's prior labor market status have been found to be important determinants of both program participation and subsequent earnings (Heckman & Smith, 1999). We capture these dynamics using a series of four employment transition dummy variables. From both the JTPA and ES data we know whether an individual is employed at the time of enrollment. From the UI data we know whether an individual is employed in each of the eight quarters prior to enrollment. For an individual employed at the time of enrollment, we coded the transition as not employed/employed if earnings were 0 in any of the eight quarters prior to enrollment and coded it as employed/employed if earnings in every quarter were positive. An individual not employed at the time of enrollment was coded employed/not employed if earnings were positive in any of the prior eight quarters and not employed/not employed otherwise.

Previous research has also found local labor market conditions to be an important determinant of program participation (Heckman et al., 1998). We capture this effect by including a dummy variable for the service delivery area (SDA) where an individual lives.[19]

Our measure of labor market experience is defined as

$$\text{Experience} = \text{Age} - \text{Years of Education} - 6.$$

We also include dummy variables indicating whether someone was employed in each of the four quarters prior to participation, to capture labor market experience immediately prior to participation. Someone is considered employed in a quarter if earnings are greater than 0.

Table 1 presents summary statistics for our JTPA and ES samples separately for males and females. For most of the demographic variables the two samples are similar.[20] However, looking at the labor market transition variables we see that JTPA participants are much more likely to be not

[16] Subsequently, many of these services became available online so individuals no longer needed to go into an ES office and register before obtaining services.

[17] Approximately 10% of the original ES sample was eliminated due to missing or invalid demographic variables.

[18] Inclusion of Kansas wage record data is valuable since a substantial number of Missouri residents in Kansas City and surrounding areas work in Kansas. The number of Missouri residents commuting across state lines is not significant elsewhere in the state.

[19] Under JTPA, there were fifteen SDAs in Missouri, each overseen by a Private Industry Council, with representatives from both the local private and public sectors. In general, SDAs are structured to identify labor market areas, corresponding to metropolitan areas and to relatively homogeneous collections of contiguous counties elsewhere. Under the Workforce Investment Act, which replaced JTPA, thirteen of the fifteen regions remain as administrative units, whereas two of the SDAs were combined.

[20] We have modified both the occupation and the education variables to ensure that they are comparable across the two files. The details of the modifications we made are provided in the data appendix.

Table 1.—Summary Statistics

	Males		Females	
	JTPA	ES	JTPA	ES
Average years of education	11.84	11.77	12.02	11.91
Average years of experience	17.98	16.43	15.47	15.38
Percent white non-Hispanic	63.0	68.0	66.9	63.7
Percent veteran	29.1	15.5	2.3	1.4
Labor market transitions (percent)				
Not empl./empl.	8.4	6.7	8.0	7.0
Empl./empl.	7.1	9.6	9.4	10.3
Empl./not empl.	23.7	36.7	15.6	34.5
Not empl./not empl.	60.8	47.0	67.0	48.2
Occupation (percent)				
Missing	54.6	35.6	65.7	40.0
Managers/supervisors	1.7	3.7	1.5	3.9
Professionals	1.7	3.1	2.6	4.9
Sales	2.8	2.5	4.7	6.7
Clerical	3.1	3.4	6.4	14.4
Service	8.9	7.9	11.9	12.9
Precision production, craft, construction	4.1	12.1	0.4	0.6
Machine operators, inspectors/transportation	9.6	19.1	4.5	11.8
Agricultural workers/laborers	13.3	12.1	2.2	4.8
Percent in Kansas City SDA	17.6	13.0	13.2	14.2
Percent in St. Louis SDA	15.3	14.7	9.0	14.7
Mean post-enrollment earnings (quarters 5 to 8)	7,595	7,708	6,543	6,392
Mean earnings in quarter of assignment	817	1,541	573	1,213
Mean earnings one quarter prior to enrollment	875	2,111	679	1,591
Mean earnings two quarters prior to enrollment	1,067	2,095	787	1,570
Mean earnings three quarters prior to enrollment	1,331	2,005	860	1,507
Mean earnings four quarters prior to enrollment	1,398	1,920	867	1,442
Mean pre-enrollment earnings (quarters −8 to −5)	5,405	6,616	3,633	5,004
Mean growth in pre-enrollment earnings	90	297	14	205
Mean difference between pre- and post-enrollment earnings	2,190	1,092	2,911	1,388
Mean estimated probability of participation	0.17	0.05	0.23	0.09
Number	2,802	45,339	6,395	52,895

employed over the entire eight quarters prior to beginning participation. Looking at earnings we see that mean post-enrollment earnings are similar for the two samples but that mean pre-enrollment earnings are lower for JTPA participants, particularly for female participants. The numbers in table 1 demonstrate that there are differences in the JTPA and ES samples, particularly in earnings and employment dynamics prior to participation. This suggests that we need to account for these differences when estimating the impact of program participation on JTPA participants.

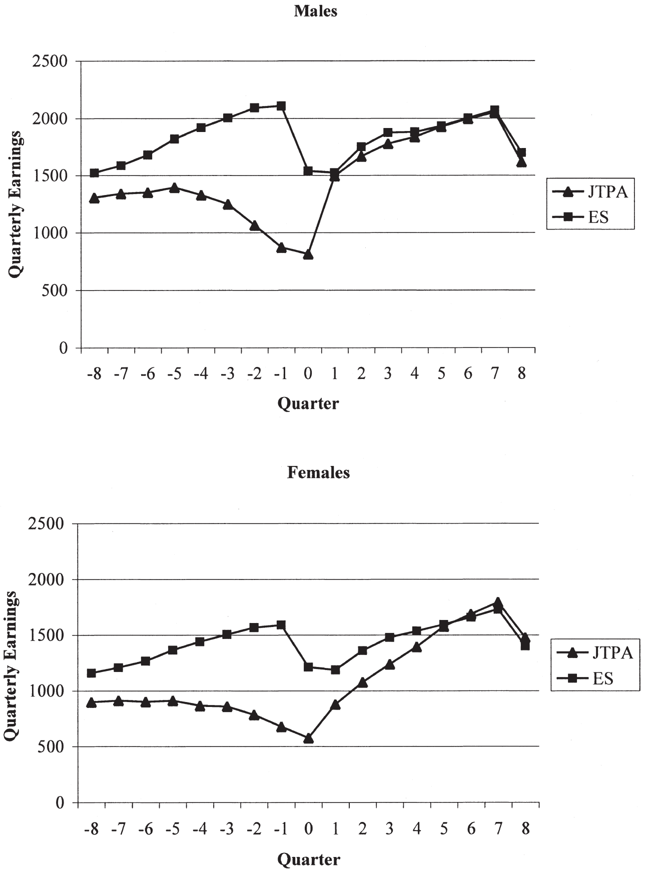
One of the conclusions reached by Heckman, LaLonde, and Smith in their chapter on program evaluation in the *Handbook of Labor Economics* (1999) is that "better data help a lot" (p. 1868) when evaluating government-sponsored training programs. The most important criteria they mention are that outcome variables should be measured in the same way for both participants and nonparticipants, that members of the treatment and comparison groups should be drawn from the same local labor markets, and that the data should allow one to control for the dynamics of an individual's labor force status prior to enrollment. With the exception that we are not able to distinguish the unemployed from those who are out of the labor market, our data satisfy each of these criteria. We therefore believe that they are nearly ideal for examining the impact of government-sponsored training programs. An additional advantage that we should mention is that Missouri is not unique. Almost every state in the union collects similar administrative data. Therefore, the type of analysis we perform could be conducted for other states as well. We next turn to examining the effects of alternative methods for constructing comparison groups on the estimated impact of treatment.

IV. Estimates of Program Effects Using Alternative Methods to Form Comparison Groups

A. *Specification Analysis*

Before presenting our estimates of the effect of the JTPA program on participants, we want to compare our various treatment and comparison samples and present the results from specification tests in order to assess whether our matching methods produce valid comparison samples. In this analysis we will focus on two variables, the sum of individual earnings in the eighth through the fifth quarters prior to beginning participation—what we call pre-program earnings—and the difference in earnings between the fifth and eighth quarters prior to beginning participation—what we call the growth in pre-program earnings. Our matching and adjustment procedures are based on individual demographic characteristics, and employment and earnings in the

FIGURE 1.—QUARTERLY EARNINGS OF JTPA AND ES PARTICIPANTS



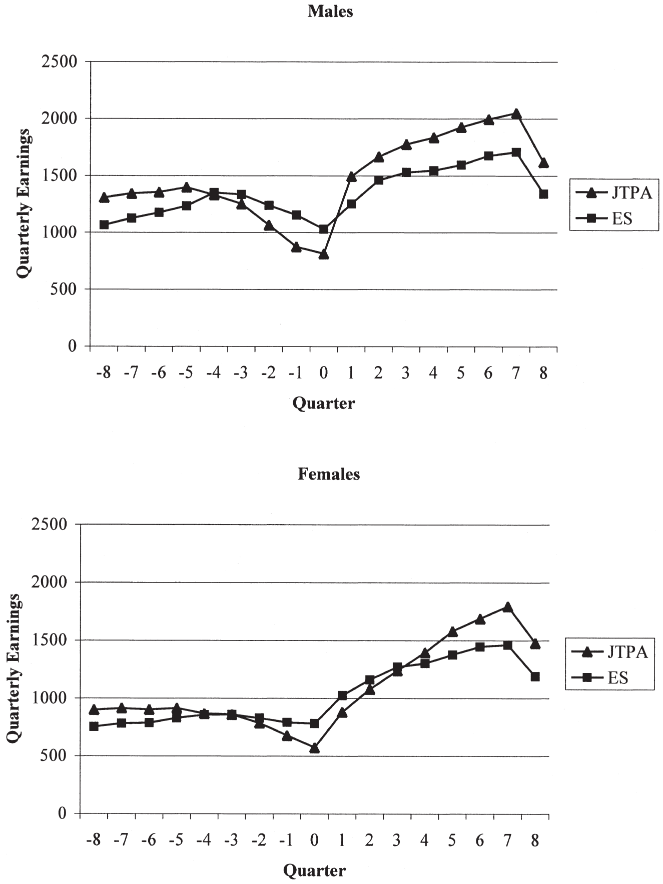
Note: Quarters are measured relative to the quarter of entry into the program. Quarter of entry is designated as quarter 0.

year prior to program enrollment, but our measure of pre-program earnings is not explicitly controlled in any of these approaches. Analysis of these earnings can therefore provide a specification test for our models. In particular, testing for differences between pre-program earnings for our treatment and comparison samples represents a test of our cross-sectional estimator (Heckman & Hotz, 1989). Similarly, testing for differences in the growth in pre-program earnings is a test for our difference-in-difference estimator, as it indicates whether pre-program growth in earnings differs for treatment and comparison samples.

Figure 1 plots quarterly earnings for our entire sample of JTPA and ES participants for the eight quarters prior to enrollment, the quarter of enrollment, and the eight quarters after enrollment. Earnings are plotted separately for men and women. Similar to table 1, figure 1 shows that JTPA and ES participants have very different earnings dynamics both prior to and after beginning participation. Prior to participation, the ES sample has much higher earnings levels and earnings growth than the JTPA sample. In addition, the Ashenfelter dip is present in both samples, although at somewhat different times. For the JTPA sample, quarterly earnings begin to decline four quarters prior to participation, whereas for ES participants earnings begin to decline one quarter prior to participation. The fact that earnings begin to decline four quarters prior to participation for JTPA participants is primarily why we measure pre-program earnings in the eighth through the fifth quarters prior to participation when constructing the difference-in-difference estimator.

Figure 2 presents the same information as figure 1 for our treatment and comparison samples created by matching on the Mahalanobis distance. As described above, for each participant in the JTPA sample, we choose a case from the comparison file for which the Mahalanobis distance is at its minimum, yielding a paired file. This pair matching method ensures that if there is at least one individual in the comparison sample who is similar on all values to each participant, the resulting matched comparison group will display the same variable distribution. In calculating the Mahalanobis distance, the characteristics in X' and X'' include education, race, prior experience, occupation (nine categories), our measures of employment status dynamics prior to enrollment (three dummy variables), dummy variables for whether an individual lived in either the St. Louis or Kansas City SDA, earnings for the four quarters prior to enrollment, and dummy variables indicating whether an individual was employed in each of the four quarters prior to enrollment.

FIGURE 2.—QUARTERLY EARNINGS OF MATCHED JTPA AND ES PARTICIPANTS—MATCHED USING MAHALANOBIS DISTANCE



Note: Quarters are measured relative to the quarter of entry into the program. Quarter of entry is designated as quarter 0.

The figure shows that while matching has produced a comparison sample with mean earnings prior to participation that are closer to the mean earnings of the treatment sample, they are still not identical. However, it does appear that the treatment and comparison samples experience similar growth in earnings prior to participation. Also, the timing of the Ashenfelter dip corresponds more closely for these two groups, although earnings begin falling one or two quarters earlier for JTPA participants and the decline in earnings is smaller for ES participants. The fact that earnings in the four quarters prior to participation are higher for ES participants is surprising because these earnings are included in the *X* vector used for matching. This suggests that Mahalanobis distance matching may fail to select a comparison group that corresponds closely even on the variables that are used in the matching process. We will see below that propensity-score matching is generally more effective.

One of the advantages of propensity-score matching is that the propensity score provides a simple measure to compare the overlap between the treatment and comparison samples. Properly estimating the effect of a program requires one to compare comparable individuals, which will occur only when the two samples have common support. In addition, the amount of overlap between the treatment and comparison samples determines the appropriate matching method and will affect the quality of the resulting matches. Figure 3 presents the distribution of propensity scores for both the JTPA and ES participants, separately for males and females. To estimate the propensity score, we use a logit function to predict participation in the sample combining the JTPA and ES samples. In addition to the variables used for matching for Mahalanobis distance, we tested nearly 300 interactions between these variables, using a stepwise procedure to enter all interactions that were statistically significant at the 5% level.[21]

Although figure 3 shows that a larger percentage of ES participants have propensity scores between 0.0 and 0.1, there are substantial numbers of ES participants with larger propensity scores. In fact, both samples span the entire range from 0.0 to 1.0. This suggests that, conditional on the assumptions of propensity-score matching, it will be possible to form samples of comparable individuals. The large overlap between the ES and JTPA samples also suggests that it will be possible to produce close matches using matching without replacement.

Figure 4 provides evidence on the comparability of our samples matched using the propensity score. This figure plots quarterly earnings both prior to and after the initial quarter of participation for our treatment and comparison samples.[22] We see that matching using the propensity score produces samples of JTPA and ES participants with similar pre-program earnings dynamics. Comparing figures 2 and 4 shows that, relative to matching using the Mahalanobis distance, matching using the propensity score produces samples that match more closely on earnings in the four quarters immediately prior to participation. Since these variables are used in both matching procedures, this suggests that propensity-score matching is more effective in practice. However, it is still the case that there are differences in pre-program earnings (earnings in the fifth through eighth quarters prior to participation), particularly for males.

FIGURE 3.—PROPENSITY-SCORE DISTRIBUTION FOR JTPA AND ES SAMPLES

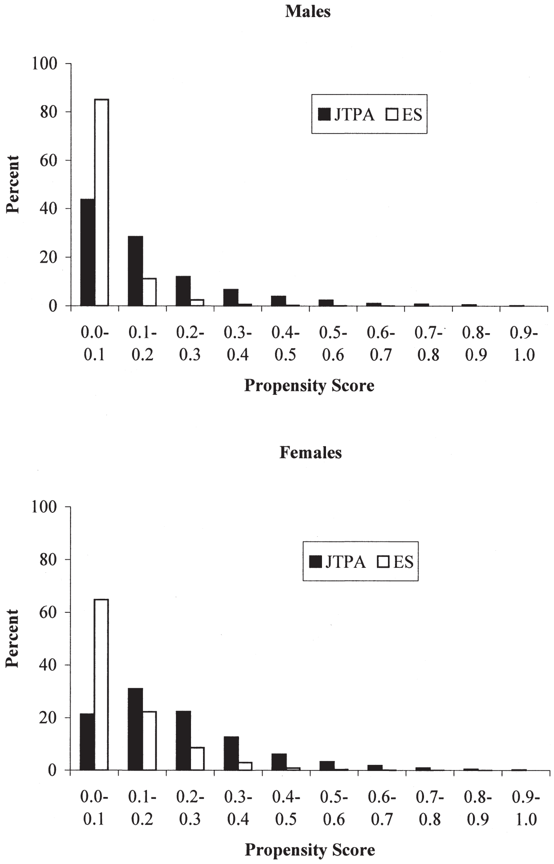
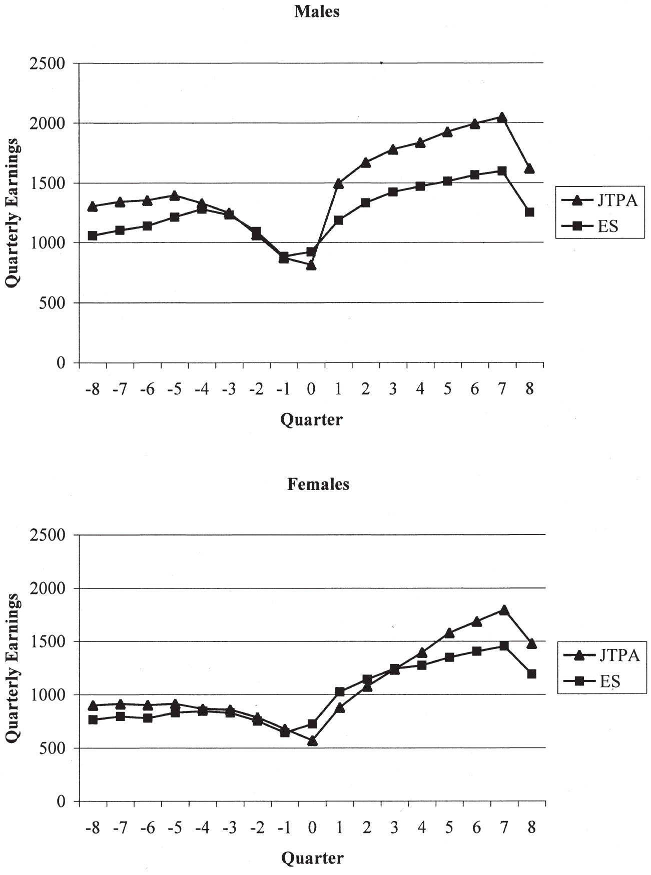


Figure 5 plots earnings for a sample of JTPA and ES participants matched using the propensity score and applying a caliper of 0.1. In caliper matching, we break any matches where distance exceeds the caliper value. Figure 5 shows that caliper matching produces samples that are very closely matched on earnings in the four quarters prior to the treatment, although there is still a difference in the level of pre-program earnings (the fifth through

[21] An alternative approach is to add interaction terms to the specification if the propensity score fails to assure balance, that is, if values of *X* differ systematically for participants and the comparison group at particular values of the predicted propensity score (for example, Smith & Todd, 2005a). We tested to be sure our propensity score was successful in balancing variable means, so results should be very similar to those obtained with such an approach.

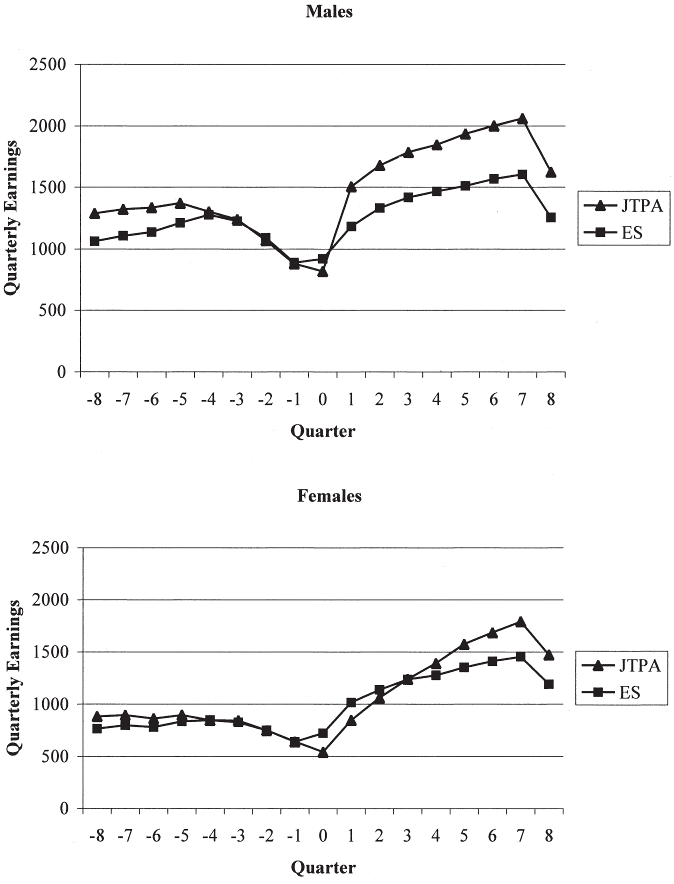
[22] These samples are formed using standard matching without a caliper. Further details on alternative matching procedures are provided below.

FIGURE 4.—QUARTERLY EARNINGS OF MATCHED JTPA AND ES PARTICIPANTS—MATCHED USING PROPENSITY SCORE



Note: Quarters are measured relative to the quarter of entry into the program. Quarter of entry is designated as quarter 0.

FIGURE 5.—QUARTERLY EARNINGS OF MATCHED JTPA AND ES PARTICIPANTS—MATCHED USING PROPENSITY SCORE WITH A 0.1 CALIPER



Note: Quarters are measured relative to the quarter of entry into the program. Quarter of entry is designated as quarter 0.

eighth quarters prior to participation) between the treatment and comparison samples. Of course, pre-program earnings are not used in the matching procedure, so this difference is not due to a technical shortcoming in the matching method.

Table 2 presents results from a more formal analysis of the difference in pre-program earnings levels and the difference in the growth in pre-program earnings between our treatment and comparison samples. The lines labeled "No regression adjustment" present the mean and standard error of the difference in the pre-program earnings level and the growth in pre-program earnings between our treatment and comparison samples. The lines labeled "Regression adjustment" present the coefficients and standard errors from a linear regression model where we include a dummy variable that equals 1 if an individual participated in JTPA. Controls in the model include the standard demographic variables (race, experience, experience squared, and veteran status), and years of education, along with a dummy variable identifying high school graduates and an additional term capturing years of schooling beyond high school, earnings in each of the four quarters prior to participation, dummy variables indicating whether the person worked in each of the four quarters prior to participation, our employment transition variables, dummy variables for nine occupations and for fifteen SDAs (identifying labor markets), and dummy variables indicating which calender quarter an individual entered either the JTPA or ES program.[23]

The results in table 2 summarize what we have seen in figures 1–2 and 4–5. Without controls, JTPA participants have appreciably lower earnings in the prior year (fifth through eighth quarters prior to enrollment), but using regression adjustment or any of the matching procedures—with or without additional regression adjustment—causes the difference to reverse. The differences in pre-program earnings for treatment and comparison samples are statistically significant (in the range of $800 for males and $400 for females) for all methods used. This suggests that there are differences between treatment and comparison groups that are not captured through matching or the controls in our regressions. As a result, the cross-sectional model based on post-program earnings may be misspecified, and so, in essence, it may not be comparing comparable individuals.

[23] These are the control variables we use throughout the paper when we undertake regression adjustment.

Table 2.—Estimates of Program Participation on Pre-Program Earnings and Earnings Growth

	Males		Females	
	Pre-Program Earnings Level	Growth in Pre-Program Earnings	Pre-Program Earnings Level	Growth in Pre-Program Earnings
Simple differences				
(1) No regression adjustment	−1,211	−207	−1,371	−192
	(162)	(34)	(84)	(18)
(2) Regression adjustment	854	−105	393	−62
	(96)	(34)	(51)	(18)
Mahalanobis distance matching				
(3) No regression adjustment	803	−76	478	−62
	(185)	(44)	(92)	(23)
(4) Regression adjustment	937	−74	448	−43
	(122)	(47)	(59)	(23)
P-score matching				
No caliper				
(5) No regression adjustment	823	−53	422	−62
	(177)	(45)	(88)	(24)
(6) Regression adjustment	811	−55	360	−63
	(130)	(44)	(64)	(24)
0.10 caliper				
(7) No regression adjustment	733	−61	327	−64
	(167)	(45)	(83)	(24)
	[*N* = 2,748]	[*N* = 2,748]	[*N* = 6,257]	[*N* = 6,257]
(8) Regression adjustment	777	−57	330	−61
	(128)	(45)	(65)	(24)
	[*N* = 2,748]	[*N* = 2,748]	[*N* = 6,257]	[*N* = 6,257]

Note: Standard errors are in parentheses. Bootstrap standard errors based on 100 replications are reported for propensity-score estimates. Analytical standard errors are reported for other estimates. There are 2,802 male participants and 6,395 female participants, except where numbers of participants are specified in brackets.

However, the results in table 2 also show that there are much smaller differences—differences that are not statistically significant for men—between our treatment and comparison samples in the growth of pre-program earnings. These results suggest that the unobserved difference between individuals in the treatment and comparison samples may be largely fixed over time and will be captured in our difference-in-difference specification. This, in addition to the fact that we measure pre-program earnings before the onset of the Ashenfelter dip, makes us optimistic that our difference-in-difference estimator may produce unbiased estimates of the effect of the program on participants.

B. *Estimates of Program Effects without Matching*

We start by considering the mean differences in earnings between our samples of JTPA and ES participants. The mean difference in postenrollment earnings between these two samples, as well as the mean difference in the difference between pre- and postenrollment earnings of the two samples, are presented in line 1 of table 3. Earnings differences between JTPA and ES are small and not statistically significant for either men or women. In contrast, males in the JTPA sample have almost a $1,100 greater *increase* in earnings relative to ES participants, while females in JTPA

Table 3.—Estimates of Program Effect Based on Simple Differences, Regression Analysis, and Mahalanobis Distance Matching

	Males		Females	
	Post-Program Earnings	Difference-in-Difference	Post-Program Earnings	Difference-in-Difference
(1) Simple differences	−113	1,098	151	1,522
	(173)	(190)	(93)	(104)
(2) Regression adjustment	1,481	628	1,087	693
	(157)	(177)	(86)	(98)
Mahalanobis distance matching				
Standard pair matching				
(3) No regression adjustment	1,267	465	1,067	589
	(194)	(216)	(108)	(122)
(4) Regression adjustment	656	719	1,054	606
	(197)	(220)	(110)	(121)
(5) Modified pair matching	1,285	482	1,13	620
	(194)	(216)	(108)	(122)
(6) Standard pair matching—triming tail	1,227	513	1,066	630
	(191)	(212)	(108)	(119)
	[*N* = 2,770]	[*N* = 2,770]	[*N* = 6,331]	[*N* = 6,331]

Note: Standard errors are in parentheses. There are 2,802 male participants and 6,395 female participants, except where numbers of participants are specified in brackets.

experience a $1,500 greater increase in earnings. Given the results presented in table 2, as well as the differences across groups in the mean values for other characteristics seen in table 1, these earnings differences at least in part reflect differences in pre-program characteristics.

Line 2 of table 3 presents estimates of program effects based on the simple linear regression model. The structure of these regressions and the control variables included are described in the previous section. The coefficient estimates for the control variables are reported in table A1 in the appendix. These coefficients generally correspond to expectations.

There are substantial differences between our cross-sectional and difference-in-difference estimates of program impact. For men the cross-sectional estimate is nearly $1,500, while the difference-in-difference estimate is only about $630. For females, the cross-sectional estimate is just under $1,100, while the difference-in-difference estimate is about $700. The results in table 2 suggest that, even after regression adjustment, the cross-sectional estimate is based on a misspecified model. The differences in the two estimates could well be the result of unobserved differences between the two groups.

As noted above, the critical question is whether regression adjustment is properly estimating what earnings would be in the absence of participation. Our large comparison sample has important advantages, but it also entails risks of misspecification. The estimated functional relationships will be largely determined by the comparison sample, and if values of control variables differ dramatically for participants, their potential earnings may be incorrectly estimated.

C. *Mahalanobis Distance Matching*

One natural approach is to choose a selection of cases from the comparison group that have similar values to those of participants. One measure of similarity is the Mahalanobis distance metric. Line 3 of table 3 shows our estimates of the program effects using the comparison sample formed by matching using the Mahalanobis distance. Comparing the cross-sectional estimates with the difference-in-difference estimates, we again see that the difference-in-difference estimates are much smaller.

Line 4 of table 3 presents our estimates of the effect of the program on participants using the matched samples and our basic linear regression model. To the extent that matching eliminates differences in the *X*s between the two samples, the estimates in lines 3 and 4 should be the same. While the estimates are similar for females, the regression produces a different estimate for males, suggesting that matching based on the Mahalanobis distance is not producing a sample with the same distribution of *X*s as the treatment sample. This is consistent with figure 2, where we saw differences in the level and growth of earnings immediately prior to participation.

Line 5 in table 3 presents our estimates based on our comparison sample matched using the Mahalanobis distance and our modified matching method. As we discussed above, when using the standard matching algorithm, the resulting matched sample depends on the order of the original data, whereas this is not the case with our modified matching algorithm. Line 6 presents results based on a comparison sample created by using the standard matching algorithm but then dropping 1% of the sample with the largest Mahalanobis distance. Comparing the estimates in lines 3, 5, and 6 shows that all of these techniques produce similar estimates.

D. *Propensity-Score Matching*

Matching cases on the basis of propensity score promises substantial simplification as compared with any general distance metric. The theory assures us that the distribution of independent variables will be the same across cases with a given propensity score, even when values differ for a particular matched pair.

As we indicate above, our estimate of $P(X)$ is based on a logit model. For each case the predicted value from our estimated logit function provides an estimate of $P(X)$. Table 4 presents our estimates of the program effects using a variety of methods for creating comparison samples based on $P(X)$. Since standard formulas for estimating standard errors do not reflect the fact that our samples are matched using $P(X)$, which is measured with error, all estimates of the standard errors are estimated using bootstrapping.[24] Lines 1 and 2 of table 4 present estimates based on comparison samples created using standard one-to-one pair matching without replacement and without a caliper. Line 1 presents estimates without regression adjustment, while line 2 presents our estimates that correct for any difference between treatment and matched samples based on our linear regression model. Comparing the estimates in line 1 based on post-program earnings with the difference-in-difference estimates again shows that these estimates are significantly different. Since our previous analysis suggests that the cross-sectional estimates are based on a misspecified model, we focus on the difference-in-difference estimates.

Comparing the regression adjusted estimates (lines 2, 6, and 9) to the estimates without adjustment shows that regression adjustment reduces the variance of the estimates slightly but has an inconsistent and generally small effect on the estimate size. This is what one would expect if the matching was successful. These results suggest that the propensity-score matching method is more successful than the Mahalanobis distance matching in creating an appropriate comparison sample.

[24] We estimate standard errors using a bootstrap procedure (100 replications) whenever our estimates are based on propensity-score matching. A recent paper by Abadie and Imbens (2006b) shows that the bootstrap generally provides an inconsistent estimate of the true error. Although their simulations indicate that bootstrap estimates are not wildly inaccurate, we exercise caution in using them. Alternative approaches to estimating standard errors, such as those developed in Abadie and Imbens (2006a), have not been widely used to date.

TABLE 4.—ESTIMATES OF PROGRAM EFFECT ON ANNUAL EARNINGS BASED ON PROPENSITY-SCORE MATCHING

	Males		Females	
	Post-Program Earnings	Difference-in-Difference	Post-Program Earnings	Difference-in-Difference
Matching without replacement				
Standard pair matching				
(1) No regression adjustment	1,532	709	1,179	757
	(199)	(206)	(97)	(128)
(2) Regression adjustment	1,562	751	1,173	814
	(196)	(194)	(94)	(110)
Standard pair matching with caliper				
(3) 0.01 caliper	1,460	741	1,169 (120)	851
	(200)	(250)		(139)
	[N = 2,726]	[N = 2,726]	[N = 6,212]	[N = 6,212]
(4) 0.05 caliper	1,480	723	1,173	845
	(198)	(201)	(99)	(124)
	[N = 2,740]	[N = 2,740]	[N = 6,228]	[N = 6,228]
0.10 caliper				
(5) No regression adjustment	1,496	764	1,177	850
	(201)	(204)	(100)	(125)
	[N = 2,748]	[N = 2,748]	[N = 6,257]	[N = 6,257]
(6) Regression adjustment	1,522	746	1,187	857
	(199)	(198)	(96)	(112)
	[N = 2,748]	[N = 2,748]	[N = 6,257]	[N = 6,257]
(7) 0.20 caliper	1,525	727	1,184	847
	(200)	(205)	(105)	(117)
	[N = 2,765]	[N = 2,765]	[N = 6,318]	[N = 6,318]
Modified pair matching				
(8) No regression adjustment	1,731	822	1,165	722
	(199)	(212)	(97)	(124)
(9) Regression adjustment	1,707	947	1,136	776
	(196)	(202)	(92)	(103)
Matching with replacement				
(10) One nearest neighbor	1,682	701	1,253	892
	(249)	(306)	(154)	(163)
(11) Five nearest neighbors	1,661	70	1,204	754
	(203)	(221)	(104)	(130)
(12) Ten nearest neighbors	1,681	758	1,234	777
	(153)	(214)	(98)	(130)
(13) Matching by propensity-score category	1,608	782	1,209	787
	(135)	(164)	(87)	(106)
(14) Kernel density matching	1,574	790	1,226	798
	(163)	(171)	(88)	(116)

Note: Standard errors are in parentheses. Standard errors are calculated by bootstrap methods based on 100 replications. There are 2,802 male participants and 6,395 female participants, except where numbers of participants are specified in brackets.

Caliper matching differs from standard matching in that only matches within a specified distance are permitted, so not all participants may be matched. Lines 3–7 show how our estimates differ when the caliper is set to 0.01, 0.05, 0.1, and 0.2, respectively. The numbers in brackets show the number of cases in the treatment sample that are matched. In the full JTPA sample (after deletions of cases with missing data), there are 2,802 males and 6,395 females. Even when we impose the 0.01 caliper, the sample size does not drop by very much, implying that matches used in the estimates reported in line 1 are generally good. Comparing our estimates in lines 3–7 with our estimates in line 1 shows that imposing a caliper has very little effect on our estimates of the program effect.[25]

E. *Comparing Pair Matching Algorithms*

The matching algorithm used in the above analysis is the standard pair matching procedure. As we discussed in the previous section, we also consider a modified matching procedure that produces matched samples that are insensitive to sample ordering and should increase the quality of the final matches. In searching the comparison sample to find a match, this alternative procedure compares not only unmatched cases but also previously matched cases, breaking previous matches if the new match distance is smaller.

Table 4, lines 8 and 9, present results using this alternative matching technique. The average difference in propensity

[25] When we impose a 0.001 caliper, although we observe little change in impact estimates for females, for males estimates increase by about $200, approximately 1 standard error. Those omitted by the 0.001 caliper but not the 0.01 caliper have an average propensity score of 0.48, in contrast to an average in the remaining treated sample of 0.14. The 0.001 caliper estimates clearly apply to a somewhat different population than estimates using the 0.01 caliper.

scores between matched pairs was often appreciably smaller when this alternative was used. Focusing on the difference-in-difference estimates, we see that estimates for males are somewhat larger than those using standard matching, although differences are never more than 1 standard error. For women, estimates are slightly smaller than those using standard matching. The similarity in estimates reflects the fact that although this method often selects a different comparison case to be matched with a particular treated case, the overall composition of the comparison sample changes relatively little.

F. *Matching with Replacement*

Matching without replacement will work well when, for each combination of characteristics, there are at least as many comparison cases as treated cases, allowing each treated case to be matched with a distinct comparison case. However, where there are a small number of comparison cases with characteristics shared by many treated cases, researchers often use matching with replacement. In that case, an individual in the comparison sample can be matched to more than one person in the treatment sample. In order to examine the sensitivity of our estimates to this alternative matching strategy, we have constructed matched samples using matching with replacement. We have also matched each person in the treatment sample to one, five, and ten nearest neighbors in the comparison sample. Our estimates based on these samples are presented in lines 10–12 in table 4. Comparing these estimates with the estimates reported in line 1 again shows that this alternative matching method produces estimates that are quite similar to our original estimates. However, matching with replacement appears to increase standard errors in the case of one-to-one matching, which is what we would expect if the repeated use of comparison cases magnified variation due to sampling.[26] The standard errors for five and ten nearest neighbors are not different from those obtained using conventional matching.

G. *Matching by Propensity-Score Category*

All of the pair matching approaches described above have the important disadvantage that they require that we discard comparison group members who are not matched. In one-to-one matching, only one case from the larger sample can be used for each case in the smaller sample, resulting in an immediate loss of information. Where the distribution of participants and the comparison groups differs dramatically, either the matches will be poor, or, if a caliper is applied, additional cases will be lost.

Group matching relaxes the requirement that the two groups be matched on a one-to-one (or one-to-*N*) basis. In those regions of the data where there are some participants and some comparison group members, group matching allows us to use all the data. The only cases that must be discarded are those for which there are no similar cases in the other group. The approach we use is closely modeled on that recommended by Dehejia and Wahba (2002) and is described in section II above.

In order to ensure that the propensity ranges were sufficiently small, we calculated the mean differences on our primary independent variables between participant and comparison groups within a propensity category. We first considered uniform propensity categories of size 0.1. However, given the large number of cases with propensity values less than 0.1, we found that differences in our basic variables within the lowest group were often statistically significant. We ultimately created much smaller category widths at the lower end of the propensity distribution, corresponding approximately to deciles in the distribution of the combined sample.

The estimated program effects based on this approach are listed in line 13 of table 4. The estimates are quite similar to our initial estimates, although the standard errors are somewhat smaller. This supports the view that matching by category increases power by using more of the comparison cases in the data.

H. *Kernel Density Matching*

Following an approach outlined in Heckman, Ichimura, and Todd (1997, 1998) and Heckman et al. (1998), we employ a kernel density estimator to calculate the density of the propensity score and the means for post-program earnings by propensity score for participants and the comparison group.[27] In order to choose the bandwidth and kernel, we used least squares leave-one-out cross-validation (Black & Smith, 2004; Hall, Racine, & Li, 2004), ultimately choosing a 0.01 bandwidth and an Epanechnikov[28] kernel. The results are reported in line 14 of table 4. These estimates are again similar to the other estimates reported in table 4. Standard errors are smaller than those obtained with simple matching methods, and in most cases only slightly larger than those obtained using matching by propensity-score category, as might be expected given that both methods use all appropriate comparison case data.

[26] In a Monte Carlo analysis of one-to-one matching, Zhao (2004) also reports an increase in standard errors of estimates based on matching with replacement relative to matching without replacement.

[27] Heckman, Ichimura, and Todd (1997) and Heckman et al. (1998) recommend using local linear regression matching, which modifies kernel estimation methods by including a linear adjustment. We tried local linear regression matching but, given the size of our samples, it was extremely time consuming to implement. In addition, the results we obtained with this approach were similar to the results obtained using kernel density matching, so we choose to focus on the results from the kernel density matching.

[28] We compared bandwidths between 0.001 and 0.2, and considering the Gaussian, Epanechnikov, biweight, and tricube kernels. Our use of the Epanechnikov kernel is consistent with the recommendation of Silverman (1986), who compares kernel properties.

TABLE 5.—SUMMARY OF ESTIMATES OF PROGRAM EFFECT

	Males		Females	
	Post-Program Earnings	Difference-in-Difference	Post-Program Earnings	Difference-in-Difference
(1) Simple difference	−113	1,098	151	1,522
	(173)	(190)	(93)	(104)
(2) Regression adjustment	1,481	628	1,087	693
	(157)	(177)	(86)	(98)
(3) Mahalanobis distance matching	1,267	465	1,067	589
	(194)	(216)	(108)	(122)
P-score matching without replacement				
(4) No caliper	1,532	709	1,179	757
	(199)	(206)	(97)	(128)
(5) 0.10 caliper	1,496	764	1,177	850
	(201)	(204)	(100)	(125)
	[N = 2,748]	[N = 2,748]	[N = 6,257]	[N = 6,257]
P-score matching with replacement				
(6) One nearest neighbor	1,682	701	1,253	892
	(249)	(306)	(154)	(163)
(7) Matching by P-score category	1,608	782	1,209	787
	(135)	(164)	(87)	(106)
(8) Kernel density matching	1,574	790	1,226	798
	(163)	(171)	(88)	(116)

Note: Standard errors are in parentheses. Bootstrap standard errors based on 100 replications are reported for propensity-score estimates. Analytical standard errors are reported for other estimates. There are 2,802 male participants and 6,395 female participants, except where numbers of participants are specified in brackets.

I. *Further Modifications*

Our implementation of the difference-in-difference estimator takes as its dependent variable the difference between earnings prior to and after treatment, where the prior year earnings are for the period comprising the fifth through the eighth quarters before treatment. Earnings during this period are not otherwise controlled in the analysis. It is natural to ask how the cross-sectional estimates would change if earnings in this period were included as controls. We wish to know whether including this information in the analysis predicting earnings would yield estimates that correspond more closely to the difference-in-difference estimates.

We implemented one-to-one matching as reported in line 1 of table 4 but with eight quarters of earnings included in the determinants of the propensity score. For both males and females, the estimates based the earnings outcome declined, up to 10% for men and about 5% for women. In each case, the decline was less than a standard error. We also calculated kernel matching estimators based on the propensity score using all eight quarters of earnings, with very similar results. Our conclusion is that the difference between the difference-in-difference estimates and estimates using the earnings outcome are not a result of the additional information used to construct the difference. Rather, the discrepancy is due to the modeling assumption implicit in the difference-in-difference specification.[29]

Heckman and Smith (1999) suggest that where treatment and comparison group members are drawn from distinct labor markets, estimates of impact are very likely to suffer bias. In all the above analyses, propensity-score estimates include dummies for labor market areas in Missouri, but treatment and comparison cases may be from different areas so long as they have similar propensity scores.[30] We have therefore replicated our matching structure imposing the constraint that matches on the fifteen SDAs must be exact. We find that estimates change by less than 1% for males, and estimates increase by about 7% for females, but still less than three-quarters of 1 standard error. We conclude that, in our sample, there is no evidence that differences in labor market not captured by propensity score have impacted our results.

J. *Summary of Estimated Program Effects*

Table 5 presents selected estimates from tables 3 and 4. We see that, in each case, the estimate based on Mahalanobis distance is the smallest one reported in table 5, and usually the difference between this estimate and the others is appreciable. Recall that results presented in figure 2 suggested that Mahalanobis distance matching was not successful in producing samples that were comparable on the measures used for matching. Looking at the other methods that control for independent variables, we see that the range of estimates is moderate. Estimates differ by a maximum of about 30%, and in no case is the difference as great as 2 standard errors. Overall, the results in table 5 show that, with the exception of Mahalanobis distance matching, which we have found does not effectively control

[29] When all eight quarters of earnings are controlled, the difference-in-difference estimates are quite close to those obtained with earnings as the outcome variable, which is expected. This underscores the point that the difference-in-difference estimates are only distinct if the prior earnings measure is omitted from the control variables. In effect, the difference-in-difference specification restricts the impact of prior earnings.

[30] We know that if matching by propensity score is perfect, on average labor markets will correspond, but not necessarily for any one match. Where propensity score is not matched exactly, there may be overall deviations.

TABLE 6.—COMPARISON OF ESTIMATED PROGRAM EFFECTS USING DIFFERENCE-IN-DIFFERENCE ESTIMATOR WITH EFFECTS BASED ON RANDOMIZED CONTROL GROUPS—ANNUAL EARNINGS

	Orr et al. (1996)		Current Analysis			
	Months 7–18	Months 19–30	Propensity Score, No Caliper	Propensity Score, 0.10 Caliper	Propensity Score Categories	Kernel Density Matching
Men	666	1,001	709	764	780	790
	(478)	(511)	(206)	(204)	(190)	(171)
Women	1,015	990	757	850	787	798
	(288)	(319)	(128)	(125)	(111)	(116)

Note: Standard errors are in parentheses. The Orr et al. (1996) estimates are taken from exhibit 4.6, page 107. They have been adjusted for inflation so that they are comparable to the estimates from the current analysis.

independent variables, estimates of program effect on participants are relatively insensitive to the methods used to form comparison groups and weight the data.

As we have noted previously, our specification tests in table 2 show that cross-sectional estimates are likely to be biased, as they depend on comparison of individuals whose pre-program earnings differ. However, there is evidence in table 2 suggesting that once individual fixed effects are removed, earnings patterns are similar, so that difference-in-difference estimates may be valid. Among the difference-in-difference estimates (omitting lines 1 and 3), we see that, for men, our estimates range from a low of $628 to a high of $790. For women the estimates range from $693 to $892.

K. *Comparison with Previous Estimates of Treatment Effects Based on Randomized Control Groups*

Table 6 compares our estimated effects on annual earnings for program participants with those reported in Orr et al. (1996, p. 107, exhibit 4.6), which are based on an experimental evaluation of the JTPA program.[31] The Orr et al. estimates are for individuals who entered JTPA from November 1987 through September 1989 at sixteen sites nationwide. Although individuals are randomly assigned to the treatment, many of the controls were provided with a list of training providers and approximately a third of them received services of some kind outside the JTPA system (Heckman, Hohmann, and Smith, 2000). In contrast, in our analysis, the comparison group consists of those who registered for job exchange services. An unknown number may have received other services as well. There is no way to estimate the extent to which such service "substitution" effects may bias our results.

We have adjusted the Orr et al. estimates for inflation so that they are comparable to ours. Since our estimates are for months 13–24 after assignment, we present the Orr et al. estimates for months 7–18 and months 19–30 after assignment. Comparing our estimates for men with the Orr et al. estimates shows that our estimates lie between theirs. For women our estimates are below those reported by Orr et al., but the difference is not generally statistically significant. Overall, our estimates based on nonexperimental data appear similar to the estimates produced using experimental data for the earlier cohort of JTPA participants.

V. Robustness of Results to Limitations in Data Quality

The results reported in the previous section—especially those based on a difference-in-difference specification—suggest that program effect estimates are robust to alternative methods of matching and weighting the data. In this section we examine the sensitivity of our results to the quality of the data used to perform the analysis. We will focus on two key aspects of data quality, the observable characteristics for participants, and the size of the treatment and comparison samples.

A. *Sensitivity of Results to Observable Characteristics*

We begin by examining the robustness of our results to changes in the characteristics available for individuals. We will examine this by dropping variables from our analysis that previous researchers have found to be important when estimating program effects (see Heckman, LaLonde, & Smith, 1999). The variables we will drop are our variables measuring employment transitions prior to entering the program, the SDA dummy variables, which capture an individual's local labor market, and the variables measuring employment status and earnings in the four quarters prior to participation. We will focus on estimates produced from treatment and comparison samples that are matched using the propensity score with a 0.1 caliper.[32] For this analysis, when we drop a set of variables, we reestimate the propensity score without those variables in the logit regression. Next we match the treatment and comparison sample using the new $P(X)$. We then compute the estimates using the new matched sample. We also drop the variables from any subsequent regression adjustment.

The results from this analysis are presented in table 7. The first five lines of the table present estimates with no regression adjustment, while lines 6–10 present results based on our linear regression model. The estimates in lines 1 and 6 are identical to the estimates found in lines 5 and 6 in table 4 and are repeated here for ease of comparison.

[31] The estimates reported by Orr et al. (1996) include an adjustment for the fact that some of those assigned to treatment never enrolled. Since our data pertain to those who enroll, this is the appropriate estimate for comparison.

[32] We use the standard pair matching algorithm without replacement.

TABLE 7.—ESTIMATES OF PROGRAM EFFECT DROPPING CERTAIN VARIABLES—PROPENSITY-SCORE MATCHING WITH 0.1 CALIPER

	Males		Females	
	Post-Program Earnings	Difference-in-Difference	Post-Program Earnings	Difference-in-Difference
No regression adjustment				
(1) Including all variables	1,496	764	1,177	850
	(201)	(204)	(100)	(125)
(2) Dropping employment transitions	1,498	462	1,421	981
	(203)	(256)	(99)	(117)
(3) Dropping SDA	1,552	899	1,157	790
	(210)	(230)	(120)	(124)
(4) Dropping earnings 4 quarters prior	957	643	1,025	803
	(235)	(279)	(129)	(136)
(5) Dropping all 3	924	838	980	1,077
	(217)	(244)	(134)	(139)
Regression adjustment				
(6) Including all variables	1,522	746	1,187	857
	(199)	(198)	(96)	(112)
(7) Dropping employment transitions	1,409	627	1,410	998
	(202)	(246)	(102)	(115)
(8) Dropping SDA	1,614	843	1,198	779
	(206)	(228)	(118)	(121)
(9) Dropping earnings 4 quarters prior	752	910	944	899
	(222)	(262)	(122)	(132)
(10) Dropping all 3	994	829	1,021	1,089
	(211)	(240)	(128)	(122)

Note: Standard errors are in parentheses. Standard errors are calculated by bootstrap methods based on 100 replications.

The largest change is observed in the estimate based on the cross-sectional model for males. When the four quarters of prior earnings are no longer controlled, the estimate declines by nearly two-fifths (compare lines 1 and 4). Although dropping these variables also causes a decline in the estimate for females, the decline is much smaller. Of interest is that dropping the employment transition measures increases estimates for females appreciably but has no impact for males. Dropping all three sets of variables causes estimates to decline slightly relative to dropping the prior earnings measures (compare lines 4 and 5). In the case of regression adjustment, we observe that dropping all measures causes an increase in the estimate relative to the case where the earnings measures are dropped, but the estimate remains smaller than the baseline (compare line 10 with lines 6 and 9).

The difference-in-difference estimates are not generally as sensitive to dropping any of these variables, but there are some shifts. One surprising result is that, for men, dropping the employment transition variables reduces the estimate substantially, but dropping all three sets of variables produces estimates that are slightly higher than estimates produced controlling all of the variables. The pattern is different for women, since dropping all three classes of variables has the largest impact, increasing estimated effects by about 25%.

Overall, estimated impacts—especially the difference-in-difference estimates—appear remarkably robust to dropping these variables. In addition, the regression adjustment has very little impact on our estimates. We suspect that the discrepancy between our results and those in the literature reflects the use of a somewhat different comparison sample, in combination with the limitation to a single state. Our local labor markets are all within a single state, and so we expect differences to be much smaller than labor market variation in a national sample. Our use of ES registrants as the comparison group may be significant as well, since it means the comparison sample is of those who are seeking support in obtaining employment. If this is the case, analyses may benefit by using comparison groups of such individuals.

B. *Sensitivity to Changes in Sample Size*

We next ask to what degree our conclusions may be generalized to analyses where treatment or comparison sample sizes are substantially smaller than ours. Our first concern is the degree to which expected values of estimates based on smaller data sets correspond with ours. The theory assures us that as sample size increases, estimates approach true values, but expected values of estimates from small samples need not correspond with the true values. Our second concern is with the extent to which sampling error increases as sample size declines.

In order to examine the sensitivity of our estimates to the size of the sample, we vary our sample in three ways. First, we set the size of the comparison sample equal to the size of the treatment sample. Second we reduce the size of the treatment sample by 90%, holding constant the size of the comparison sample. Finally we reduced the size of both the treatment and the control sample by 90%. To form the smaller samples, we draw a sample of a given size with replacement from the original treatment and comparison samples. We then performed the analysis with this new

Table 8.—Sensitivity of Estimates to Changes in the Size of the Samples

	Males						Females					
	Regression Adjustment		Propensity Score, 0.1 Caliper		Propensity-Score Categories		Regression Adjustment		Propensity Score, 0.1 Caliper		Propensity-Score Categories	
	Post-Program Earnings	Diff-in-Diff	Post-Program Earnings	Diff-in-Diff	Post-Program Earnings	Diff-in-Diff	Post-Program Earnings	Diff-in-Diff	Post-Program Earnings	Diff-in-Diff	Post-Program Earnings	Diff-in-Diff
(1) Full sample	1,481	628	1,496	764	1,608	782	1,087	693	1,177	850	1,209	787
	(157)	(177)	(201)	(204)	(135)	(164)	(86)	(98)	(100)	(125)	(87)	(106)
(2) Comparison sample =	1,508	848	1,635	966	1,635	849	1,096	684	1,290	918	1,284	877
treatment sample	(224)	(246)	(280)	(327)	(273)	(317)	(147)	(203)	(139)	(149)	(147)	(161)
(3) Reduce treatment	1,493	646	1,662	850	1,518	905	1,126	751	1,274	791	1,116	845
sample by 90%	(448)	(514)	(633)	(728)	(488)	(556)	(263)	(288)	(424)	(444)	(277)	(278)
(4) Reduce both samples by	1,428	618	1,516	756	1,524	758	1,119	692	1,224	862	1,194	778
90%	(448)	(533)	(625)	(679)	(517)	(625)	(268)	(324)	(333)	(358)	(325)	(374)

Note: Mean estimates in lines 2–4 are based on 100 replications, and the standard deviations of estimates are in parentheses.

sample. We repeated the process one hundred times. For each repetition, we calculated program effects for regression adjustment, propensity-score pairwise matching with a 0.1 caliper, and estimation based on propensity-score category.[33] In each case, we present cross-sectional estimates and difference-in-difference estimates. Table 8 reports the mean and standard deviation of these one hundred estimates of the program effect.

Comparing the mean estimates reported in lines 2–4 with the estimate based on the full sample reported in line 1 shows that changing the sample size does not usually have an effect on the expected estimate value, since most differences could easily be due to sampling error.[34] However, there are some exceptions. Four of the six mean values for the difference-in-difference estimator are substantially higher when the comparison sample is reduced to equal the treatment sample (line 2), and it is clear that these differences could not be due to chance. This suggests that having a sufficiently large comparison sample may be of importance. Interestingly, the difference appears smaller in lines 3 and 4, as the treatment sample is reduced. Of course, while there is clearly a difference in the expected value of the estimate as the size of the comparison sample declines, it is modest relative to the standard deviation of the estimate, which is the appropriate measure of the standard error of the estimate in the reduced sample.

Comparing the standard deviation of our estimates in lines 2–4 with the standard error of our estimates reported in row 1 shows that reducing the sample size results in a substantially less precise estimate of the program effect. Focusing on the difference-in-difference estimates, we see that the standard deviations of our estimates are between 1.5 and 3 times the standard error of the original estimate. The greatest increase occurs when we reduce the treatment sample size. If the actual effect was as estimated in our full sample, for men the estimated effect using the smaller samples in lines 3 or 4 would produce estimates that were statistically significant fewer than one out of five times. Even for women, the effect would be significant in only two out of three samples. Assuming a 5% discount rate and taking the increment in earnings to be stable throughout a thirty-year working life, there is a one in five chance the program would not pass a benefit-cost test for males. In contrast, under these permissive standards, the program would almost certainly be judged cost beneficial for females.[35]

In short, reducing the sample size has relatively little impact on the expected estimates but does result in a substantial fall in the precision of the estimates, making it more difficult to find significant effects. Nonetheless, we find some evidence that a large comparison sample may serve to stabilize expected estimates, quite independent of effects on precision.

VI. Conclusion

Our results suggest that a variety of matching methods produce estimates of program effect that are quite similar if they are based on the same control variables. The most important exception is that we find Mahalanobis distance matching is less successful than the other methods in producing a comparison sample that is comparable. Regression adjustment, based on a simple linear model, seems to perform surprisingly well.

Specification tests suggest that cross-sectional program impact estimates are likely to suffer bias. In contrast, difference-in-difference estimators appear less likely to exhibit bias. Remarkably, difference-in-difference estimators

[33] Because we are using a caliper for pairwise propensity-score matching, not all records in the treatment sample are matched. The number of matched records varies for the repetitions and depends on the actual sample drawn.

[34] With one hundred replications, the standard error of the mean of the effect estimate reported in the table can be estimated as one-tenth of the standard deviation.

[35] This is based on our own calculations for the JTPA program in Missouri, which suggest a cost of $2,500. Orr et al. (1996) estimate costs under $2,000. See Dyke et al. (2006) for estimates of program impact over an extended period following job training program participation.

are not only quite robust to the particular matching method that is used, but they also remain relatively stable in the face of changes in the available control variables.

We have focused here on program impact on a single year's earnings, but the same approach could be extended to consider earnings over a longer period. Judgments of program efficacy depend critically on how long earnings benefits remain. As an example, assume that the earnings increment is $750 beginning in the second year after program participation, as our estimates suggest, and that program cost is $2,500 per person. If the earnings increment remains constant throughout a thirty-year working life, the internal rate of return is 30%, but if the benefits depreciate by 30% each year, the internal rate of return is only 8%.[36] Of course, in calculating actual costs of any particular program it would be necessary to identify the impacts of distortionary taxes used to raise revenues for the program (Heckman, LaLonde, & Smith, 1999).

Our work shows that estimating program impacts is feasible based on administrative data that are collected and maintained in most states. Our findings suggest that researchers use a difference-in-difference estimator. Controlling for variables can be accomplished in a variety of ways, although we believe that those based on propensity-score matching are most likely to provide robust estimates. Among the propensity-score methods, we have a preference for matching by propensity group or kernel density matching, but other methods will produce similar estimates.

REFERENCES

Abadie, Alberto, and Guido W. Imbens, "Large Sample Properties of Matching Estimators for Average Treatment Effects," *Econometrica* 74:1 (January 2006a), 235–267.

——— "On the Failure of the Bootstrap for Matching Estimators," Unpublished paper, John F. Kennedy School of Government, Harvard University, mimeograph (2006b).

Angrist, Joshua D., and Jinyong Hahn, "When to Control for Covariates? Panel-Asymptotic Results for Estimates of Treatment Effects," NBER technical working paper no. 241 (1999).

Ashenfelter, Orley C., "Estimating the Effect of Training Programs on Earnings," this REVIEW 60:1 (1978), 47–57.

Ashenfelter, Orley C., and David Card, "Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs," this REVIEW 67:4 (November 1985), 648–660.

Barnow, Burt S., "The Impact of CETA Programs on Earnings: A Review of the Literature," *Journal of Human Resources* 22:2 (Spring 1987), 157–193.

Barnow, Burt S., Glenn G. Cain, and Arthur S. Goldberger, "Issues in the Analysis of Selectivity Bias," in E. Stromsdorfer and G. Farkas (Eds.), *Evaluation Studies,* vol. 5 (Beverly Hills, CA: Sage Publications, 1980).

Bassi, Laurie J., "Estimating the Effect of Training Program with Nonrandom Selection," this REVIEW 66:1 (February 1984), 36–43.

Black, Daniel A., and Jeffrey A. Smith, "How Robust Is the Evidence on the Effects of College Quality? Evidence from Matching," *Journal of Econometrics* 121 (July-August 2004), 99–124.

Card, David, and Daniel G. Sullivan, "Measuring the Effects of CETA Participation on Movements In and Out of Employment," *Econometrica* 56 (1988), 497–530.

Cochran, William G., "The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies," *Biometrics* 24:2 (June 1968), 295–313.

Dearden, Lorraine, Javier Ferri, and Costas Meghir, "The Effect of School Quality on Educational Attainment and Wages," this REVIEW 84:1 (2002), 1–20.

Dehejia, Rajeev H., and Sadek Wahba, "Causal Effects in Non-Experimental Studies: Re-Evaluating the Evaluation of Training Programs," *Journal of the American Statistical Association* 94 (December 1999), 1053–1062.

——— "Propensity Score-Matching Methods for Nonexperimental Causal Studies," this REVIEW 84:1 (2002), 151–161.

Devine, Theresa J., and James J. Heckman, "The Structure and Consequences of Eligibility Rules for a Social Program: A Study of the Job Training Partnership Act (JTPA)" (pp. 111–170), in Solomon W. Polachek (Ed.), *Research in Labor Economics,* volume 15 (Greenwich, CT: JAI Press, 1996).

Dyke, Andrew, Carolyn J. Heinrich, Peter R. Mueser, Kenneth R. Troske, and Kyung-Seong Jeon, "The Effects of Welfare-to-Work Program Activities on Labor Market Outcomes," *Journal of Labor Economics* 24:3 (2006), 567–608.

Fraker, Thomas M., and Rebecca A. Maynard, "The Adequacy of Comparison Group Designs for Evaluation of Employment-Related Programs," *Journal of Human Resources* 22:2 (Spring 1987), 194–227.

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

Frölich, Markus, "Finite-Sample Properties of Propensity-Score Matching and Weighting Estimators," this REVIEW 86:1 (2004), 77–90.

Hall, Peter, Jeffrey S. Racine, and Qi Li, "Cross-Validation and the Estimation of Conditional Probability Densities," *Journal of the American Statistical Association* 99 (December 2004), 1015–1026.

Hansen, Ben B., "Full Matching in an Observational Study of Coaching for the SAT," *Journal of the American Statistical Association* 99 (September 2004), 609–618.

Heckman, James J., Neil Hohmann, and Jeffrey A. Smith, "Substitution and Dropout Bias in Social Experiments: A Study of an Influential Social Experiment," *Quarterly Journal of Economics* 115:2 (2000), 651–694.

Heckman, James J., and V. Joseph Hotz, "Choosing Among Alterative Methods for Estimating the Impact of Social Programs: The Case of Manpower Training," *Journal of the American Statistical Association* 84 (1989), 862–874.

Heckman, James J., Hidehiko Ichimura, Jeffrey A. Smith, and Petra E. Todd, "Characterizing Selection Bias Using Experimental Data," *Econometrica* 66 (September 1998), 1017–1098.

Heckman, James J., Hidehiko Ichimura, and Petra E. Todd, "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme," *Review of Economic Studies* 64 (October 1997), 605–654.

——— "Matching as an Econometric Evaluation Estimator," *Review of Economic Studies* 65 (April 1998), 261–294.

Heckman, James J., Robert J. LaLonde, and Jeffery A. Smith, "The Economics and Econometrics of Active Labor Market Programs," in Orley Ashenfelter and David Card (Eds.), *Handbook of Labor Economics,* vol. 3 (Amsterdam: North-Holland, 1999).

Heckman, James J., and Jeffery A. Smith, "Assessing the Case for Social Experiments," *Journal of Economic Perspectives* 9 (Spring 1995), 85–110.

——— "The Pre-programme Earnings Dip and the Determinants of Participation in a Social Programme: Implication for Simple Programme Evaluation Strategies," *The Economic Journal* 109 (July 1999), 313–348.

Hirano, Keisukem, Guido W. Imbens, and Geert Ridder, "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica* 71:4 (2003), 1161–1189.

Hollenbeck, Kevin M., Wei-Jang Huang, Christopher T. King, Peter R. Mueser, and Daniel Schroeder, "Initial WIA Net Impacts in Seven States," prepared for the U.S. Department of Labor (December 2004).

Imbens, Guido W., "The Role of the Propensity Score in Estimating Dose-Response Functions," *Biometrica* 87 (2000), 706–710.

[36] We have assumed that the $750 increment occurs in the first and second year after training and then declines each year thereafter.

——— "Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review," this REVIEW 86:1 (2004), 4–29.

LaLonde, Robert J., "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," *American Economic Review* 76 (September 1986), 604–620.

Leuven, Edwin, and Barbara Sianesi, "Psmatch2: Stata Module to Perform Full Mahalanobis and Propensity Score Matching, Common Support Graphing, and Covariate Imbalance Testing" (2003), http://ideas.repec.org/c/boc/bocode/s432001.html.

Manski, Charles F., "Learning About Treatment Effects from Experiments with Random Assignment of Treatments," *Journal of Human Resources* 31 (Fall 1996), 709–733.

Orr, Larry L., Howard S. Bloom, Stephen H. Bell, Fred Doolittle, Winston Lin, and George Cave, *Does Training for the Disadvantaged Work? Evidence from the National JTPA Study* (Washington, DC: Urban Institute Press, 1996).

Rosenbaum, Paul R., *Observational Studies* (New York: Springer-Verlag, 2002).

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

——— "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score," *Journal of the American Statistical Association* 79 (September 1984), 516–524.

——— "Constructing a Control Group Using Multivariate Matched Sampling Methods that Incorporate the Propensity Score," *The American Statistician* 39 (February 1985), 33–38.

Silverman, Bernard W., *Density Estimation for Statistics and Data Analysis* (New York: Chapman and Hall, 1986).

Smith, Jeffrey A., and Petra E. Todd, "Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators?" *Journal of Econometrics* 125 (March-April 2005a), 305–353.

——— "Rejoinder," *Journal of Econometrics* 125 (March-April 2005b), 365–375.

Zhao, Zhong, "Data Issue of Using Matching Methods to Estimate Treatment Effects: An Illustration with NSW Data Set," Peking University working paper no. #203004 (2003).

——— "Using Matching to Estimate Treatment Effects: Data Requirements, Matching Metrics, and Monte Carlo Evidence," this REVIEW 86:1 (2004), 91–107.

Data Appendix

Occupational Codes

There are two major differences in the occupational variables in the JTPA and ES files. The first is that the JTPA file contains a larger share of records that have missing occupation codes than the ES file. Both programs ask applicants to report occupational information for their current or their most recent job. However, for applicants who have not been recently employed, this information was not considered relevant and is frequently left blank. As can be seen in table 1, JTPA applicants are much more likely to have been unemployed for all eight quarters prior to beginning participating in JTPA, and this partly accounts for why JTPA participants are more likely to have missing occupation data. In order to use occupational information for matching individuals, we felt it was important to ensure that the probability of having a missing occupation code was similar for comparable ES and JTPA participants. To accomplish this we first estimate the probability that a record in the JTPA file has a missing occupation code using a logit model. In this model we control for whether individuals were employed in the quarter of enrollment, whether they were employed in each of the four quarters prior to enrollment, their earnings in each of the four quarters prior to enrollment (with earnings set to 0 for individuals who were not employed in the quarter), along with a complete set of interactions between these variables. We estimate this model separately for men and women. We use the results from these regressions to compute the estimated probability that someone in either the JTPA or ES file has a missing occupation code. For men we set the occupation variable equal to missing when the estimated probability is greater than 0.5. We do the same for women when the estimated probability is greater than 0.55. For those whose occupation code was already missing, we left it as missing.

The second difference in the occupation variable is that in the JTPA data, occupation is coded using the Occupational Employment Statistics (OES) codes, while in the ES data occupation is coded using the Dictionary of Occupational Title (DOT) codes. To create similar codes in both files, we first used a crosswalk obtained from the Bureau of Labor Statistics to convert the DOT codes into OES codes. We then used the OES codes to create nine occupation groups: managers/supervisors (OES codes 13–19, 41, 51, 61, 71, 81); professionals (OES codes 21–39); sales (43–49); clerical (53–59); precision production, craft, and construction (85, 87, 89, 95); machine operators, inspectors, transportation (83, 91–93, 97); service workers (63–69); agricultural workers/laborers (73–79, 98); and missing.

Education

In both programs, applicants are asked about the highest grade they completed. Up through a high school diploma, this information is coded as the number of years of schooling, so someone whose education stopped with a high school diploma will have the value 12. For individuals who complete more than 12 years of schooling but do not obtain a degree, the highest grade completed is again coded as the number of years of schooling. However, for individuals who complete a post–high school degree, different codes are entered into this field indicating what degree they completed. This is also true for individuals who obtained a high school equivalency certificate (GED) prior to entering the program. We converted this information into years of schooling as follows: GED = 12 years; associate of arts degree = 14 years; BA/BS degree = 16 years; master's degree = 17 years; PhD = 20 years.

TABLE A1.—REGRESSION PREDICTING POST-PROGRAM EARNINGS

	Males		Females	
	Post-Program Earnings	Difference in Earnings	Post-Program Earnings	Difference in Earnings
Participation in JTPA	1,481.37	627.76	1,086.99	693.71
	(156.84)	(176.84)	(86.11)	(98.16)
Years of education	33.51	−40.76	−14.19	−120.34
	(48.83)	(55.06)	(40.16)	(45.79)
High school graduation	935.09	797.95	1,038.20	721.74
	(136.75)	(154.20)	(106.40)	(121.30)
Years of higher education	325.97	399.34	735.94	787.25
	(62.68)	(70.68)	(48.88)	(55.73)
Experience	−50.03	−122.75	49.24	−0.59
	(14.50)	(16.36)	(10.26)	(11.70)
Experience^2	−0.90	0.04	−2.13	−1.87
	(0.34)	(0.39)	(0.25)	(0.28)
Not employed/employed	2,294.88	2,263.94	1,641.44	1,617.58
	(161.55)	(182.28)	(111.96)	(127.63)
Employed/employed	3,483.03	4,312.43	2,350.91	2,975.9
	(177.50)	(200.14)	(126.56)	(144.28)
Employed/not employed	−67.30	1,255.52	−288.00	525.62
	(144.55)	(162.99)	(107.95)	(123.06)
White	946.47	504.32	−61.74	−295.22
	(97.98)	(110.48)	(71.23)	(81.20)
Veteran	580.98	1,105.02	316.86	1,297.11
	(101.22)	(114.13)	(211.40)	(240.99)
Earnings 1 quarter prior	0.73	0.48	0.56	0.38
	(0.03)	(0.03)	(0.03)	(0.03)
Earnings 2 quarters prior	0.32	0.05	−0.12	−0.60
	(0.04)	(0.04)	(0.03)	(0.03)
Earnings 3 quarters prior	0.27	−0.29	0.37	−0.04
	(0.04)	(0.04)	(0.03)	(0.04)
Earnings 4 quarters prior	0.36	−1.70	0.57	−1.22
	(0.03)	(0.03)	(0.03)	(0.03)
Employed 1 quarter prior	−482.82	−660.51	56.86	77.79
	(147.51)	(166.32)	(103.69)	(118.20)
Employed 2 quarters prior	−352.56	−144.29	421.98	576.58
	(137.63)	(155.19)	(943.22)	(107.40)
Employed 3 quarters prior	−310.55	245.08	−219.78	96.53
	(132.63)	(152.93)	(95.62)	(109.00)
Employed 4 quarters prior	160.79	−531.79	−319.49	−1,460.64
	(133.00)	(149.96)	(93.93)	(107.08)
Quarter of enrollment dummies	Yes	Yes	Yes	Yes
8 occupation dummies	Yes	Yes	Yes	Yes
14 service delivery area dummies	Yes	Yes	Yes	Yes
Adjusted R^2	0.23	0.18	0.19	0.17
N	48,141	48,141	59,290	59,290

Note: Standard errors are in parentheses.