Very preliminary and incomplete
Please do not quote or cite without permission

Using State Administrative Data to Measure Program Performance

by

Peter Mueser
University of Missouri-Columbia

Kenneth R. Troske
University of Missouri-Columbia

Alexey Gorislavsky
University of Missouri-Columbia

January 2000

Abstract

The goal of this paper, is to use administrative data from one state to construct a set of nonrandom comparison groups, and then use these comparison groups to estimate the effect on participants of participating in the Job Training Partnership Act (JTPA). As part of this project we evaluate various proposed methods for constructing nonrandom comparison groups based on both their accuracy and their simplicity in implementation. These methods include the Mahalonobis distance (or nearest neighbor) method, the propensity scoring method of Rosenbaum and Rubin (1983) and Dehejia and Wahba (1998), and a variation of the propensity score method, the Kernel density matching method, developed in Heckman, Ichimuria, Smith, and Todd (1998). Our results show that, conditional on comparing comparable individuals, our estimates are insensitive to the method used to construct the comparison group. This suggests two things. First, based on its relative simplicity, the propensity score method appears to be the preferred method. Second, conditional on having data with similar observable characteristics on individuals, most states should have available administrative data that would allow them to conduct similar evaluations.

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, Massachusettes, Oregon, and Texas, have all mandated that some type of evaluation of new state welfare programs must be undertaken. In addition, a number of states are trying to evaluate the effects of their workforce development systems in order to develop performance standards to aid in future evaluations of these program.

However, the best way for states to conduct these evaluations remains an unanswered question. Early efforts to evaluate the effect of government sponsored training program 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 and Card, 1985). This research culminated in the paper by LaLonde (1986) which concluded that nonexperimental evaluations had the potential for severe specification error, and that the only way to choose the correct specification for the model is through the use of experimental control groups. This led both researchers and policy makers to conclude that the only appropriate way to evaluate government programs is through the use of randomized social experiments.

However, recent critiques of social experiments (Heckman and Smith, 1995; Heckman, LaLonde and Smith, 1999) conclude that social experiments are seldom implemented appropriately, raising serious questions about whether control groups are truly random samples.

These studies also argue that, even if social experiments are implemented correctly, they do not produce estimates of the effect of government programs that are useful for policy makers in deciding whether to create new programs or to expand existing ones (see also Manski, 1996). In addition, if one wants to evaluate the long-term impact of a program, randomized social experiments can be quite costly to implement since they require evaluators to collect data from both program participants and nonparticipants over a long period of time.

Based in part on these concerns, recent research has focused on constructing nonexperimental comparison groups for use in evaluating various government programs (Rosenbaum and Rubin, 1983; Friedlander and Robins, 1995; Heckman, Ichimura, and Todd, 1997; Dehejia and Wahba, 1998; Heckman, Ichimura, Smith, and Todd, 1998). The results from these papers suggest that, conditional on having the appropriate set of observable characteristics for both participants and nonparticipants and the use of appropriate statistical methods, it is possible to obtain unbiased estimates of the effect of a program using nonexperimental comparison groups. Since most states already collect a variety of data on participants in various state programs, these results suggest that it may be possible to use already-existing data sources to conduct evaluation studies much more inexpensively that previously thought, while at the same time producing estimates of the overall effects of programs that are much more useful for policy makers.

The goal of this paper, is to use administrative data from one state, Missouri, to construct a set of nonrandom control groups and then use these control groups to estimate the effect on participants of participating in the Job-Training Partnership Act (JTPA). As part of this project we will evaluate various proposed methods for constructing nonrandom control groups based on

both their accuracy and their simplicity in implementation. These methods include the Mahalonobis distance (or nearest neighbor) method, the propensity scoring method of Rosenbaum and Rubin (1983) and Dehejia and Wahba (1998), and the Kernel density matching method developed in Heckman, Ichimura, and Todd (1997) and Heckman, Ichimura, Smith, and Todd (1998).

Our data on JTPA participants come from information collected by the state of Missouri to administer this program. Our control group consists of individuals registered in the state's Employment Security (ES) program. 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, we are able 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 composition bias (having individuals in the comparison group who are either current or recent participants in the program being evaluated). Finally, the data on wages and employment history are being generated by the same process for both participants and nonparticipants. Results in Heckman, Ichimura, Smith, and Todd (1998) indicate that these are three important factors in determining whether one can construct an appropriate nonrandom control group. In addition, the data we have from Missouri is similar to administrative data collected by other states in implementing various workforce development and UI programs. Therefore, it should be possible to use the results from our study when conducting evaluations of other state's programs.

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). We also find that, once we ensure that we only compare comparable individuals, our estimates are insensitive to the method used for constructing comparison groups. This is exactly what one should expect, conditional on having the appropriate set of observable characteristics for both participants and nonparticipants. These results lead us to conclude that it is possible to evaluate government program such as JTPA using administrative data that is currently being collected by most state governments.

The rest 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. Section V concludes.

II. Methods for Creating Nonexperimental Comparison Groups

Our goal is to estimate the effect of participating in the JTPA program on program participants. In order to make the discussion concrete, we will focus on a single outcome measure. In the case of our study, we examine earnings following the treatment. Let us specify that Y_1 is earnings for an individual following participation in the program while Y_0 is earnings for that individual in the absence of participation. Let these be functions of measured individual characteristics, listed in the vector X,

$$Y_0 = \mu_0(X) + U_0 \ , \qquad Y_1 = \mu_1(X) + U_1. \tag{1}$$

We take

$$\mu_0(X)=\mathrm{E}(Y_0|X)\ , \quad \mu_1(X)=\mathrm{E}(Y_1|X)\ ,$$

so that U_0 and U_1 are deviations from expected values, reflecting unmeasured factors.

If it were possible to observe Y_0 and Y_1 for each individual, a measure of the distribution of gains due to participation in the program as a function of X could then be calculated. However, for almost all programs we are interested in, it is impossible to observe both measures for a single individual. If we define D=1 for those who participate, and D=0 for the comparison group, 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 D is independent of Y_0 and Y_1 and the factors influencing them. In this case

$$\mathrm{E}(Y_1-Y_0) = \mathrm{E}(\Delta Y) = \mathrm{E}(Y|D=1) - \mathrm{E}(Y|D=0). \qquad (2)$$

Of course, in practice, such measures of program impact pertain not to all individuals but to the population who face randomized assignment.

Where D is not independent of factors influencing Y, participants may differ from nonparticipants in many ways, including the effect of the program. In most cases, we are interested in the effect of the program on the participants, $\mathrm{E}(\Delta Y|D=1)$. However, with nonrandom assignment, the simple difference in the outcome variable between participants and nonparticipants does not identify program impact,

$$\mathrm{E}(Y|D=1) - \mathrm{E}(Y|D=0) = \mathrm{E}(\Delta Y|D=1) + [\mathrm{E}(Y_0|D=1) - \mathrm{E}(Y_0|D=0)]\ . \qquad (3)$$

The term in brackets identifies bias due to the fact that even if they had not participated in the program, those who do participate would have faced different outcomes than the nonparticipants.

Matching and adjustment methods estimate $E(\Delta Y|D=1)$ under the assumption that, conditional on measured characteristics, participation is independent of outcomes,

$$(Y_0,Y_1) \perp\!\!\!\perp D|X. \tag{4}$$

If this condition holds, we know that

$$E(\Delta Y|X) = \mu_1(X) - \mu_0(X).$$

Under this assumption, since $\mu_1(X)=E(Y_1|D=1,X)$ and $\mu_0(X)=E(Y_0|D=0,X)$ and both μ_1 and μ_0 are observable, it is straightforward to estimate the impact of the program on the participants.

Based on (3), it can be seen that a weaker assumption than (4) suffices to allow estimation of the impact of the program on participants. If nonparticpant outcomes are independent of participation, conditional on *X*,

$$Y_0 \perp\!\!\!\perp D|X, \tag{5}$$

the program effect can be written as

$$E(\Delta Y|D=1,X) = \mu_1(X) - \mu_0(X) + E(U_1|D=1) = g(X) - \mu_0(X).$$

Since $g(X) = E(Y_1|D=1,X)$, and since *g(x)* is observable, the impact of the program on participants is again straightforward to estimate. We should emphasize that all matching techniques assume some version of (5). They differ in how they condition participation on *X*.

Simple Regression Adjustment

Given (4) the most common approach to estimating program impact is to assume that the earnings function is the same for participants and the comparison groups, except for a shifter δ,

$$\mu_1(X) = \mu_0(X) + \delta.$$

Further assuming a linear functional form, δ is estimated, along with the vector of parameters of the earnings function, β, by fitting the equation

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

where e is an error term independent of X and D. Although this approach can be pursued using more flexible functional forms, even with modifications, estimates of program impact rely on a parametric structure in order to compare participants and nonparticipants. Where the support of X differs for participants and the comparison group, these methods extrapolate outside the sample range, and, in effect, compare individuals who are not comparable.

Matching Methods

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 discrete values, it may be possible to calculate sample means that are direct estimates of $\mu_0(X)$ and $\mu_1(X)$. The estimated impact of the program is then,

$$\mathrm{E}(\Delta Y | D=1) = \frac{1}{N}\sum [\overline{\mu_1}(X) - \overline{\mu_0}(X)]N(X),$$

where $\bar{}$ identify sample means for particular values of X, $N(X)$ is the number of participants with values X, N is the total number of participants, and the summation is across all values of X. In most cases, there are too many observed values of X to make such an approach feasible.

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 one-to-one 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.

Mahalanobus Distance Matching

We first undertake one-to-one 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. Mahalanobis distance has the advantage 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.[1] This ensures that the difference between the two groups in

[1] In practice one must estimate V using either the sample of 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 from participants and nonparticipants with the weights being the proportion of each group in the sample population. Calculating V in this manner minimizes sampling error.

any linear function will be reduced (Rosenbaum and Rubin, 1985). Friedlander and Robins (1995) illustrate the use of Mahalanobis distance in program evaluation.

One-to-one matching is accomplished by first ordering the participants and the comparison group randomly. For the first participant, we match them to the comparison group member which 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 which minimizes $M(X', X'')$. The process continues through all participants until the participant or comparison group is exhausted.

We also considered a modified matching procedure in which 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 that were previously matched to participants. Here, a prior match is broken and a new matched 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.[2]

Of course, if the comparison group contains sufficient numbers of cases with very similar values on all X, the matching procedure will produce directly comparable groups. In most cases, however, there remain substantial differences between matched pairs. Therefore, we examine the impact of additional regression adjustment on estimates of program impact.

[2] One problem with the simple matching procedure is that the resulting matches are not invariant to the order in which the data are sorted prior to matching. The advantage of the second procedure is that the results should be invariant to the ordering of the data.

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 particpant. Rosenbaum and Rubin (1983) show that

$$(Y_0,Y_1) \perp\!\!\!\perp D|X \Rightarrow (Y_0,Y_1) \perp\!\!\!\perp D|P(X).$$

This means that if we consider participant and comparison group members with the same $P(X)$, the distribution of X will be the same across them. 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

$$E(\Delta Y|D=1) = E_X(\Delta Y|P(X)),$$

where E_X is the expectation across all values of X for participants. We estimate $P(X)$ using a logit specification with a high flexible functional form allowing for nonlinear effects and interactions. We first undertake one-to-one matching based on the propensity score using the methods described in the previous subsection. We also use a refinement of simple matching where we remove matches for which the difference in propensity scores between matched pairs exceeds some threshold or caliper. This is refereed to as "caliper matching." In the analysis we use calipers ranging from 0.05 to 0.2.

We then turn to two closely related approaches, also based on propensity score matching. Let the k^{th} strata or band be defined to include all cases with values of X such that $P(X) \in (P_1^k, P_2^k)$. Let N_k' be the number of participants within the k^{th} strata, N_k'' the number of individuals in the comparison group within the k^{th} strata, and N the total number of participants in our sample. Finally, let W_k' be a weight placed on each participant observation in strata k and W_k'' be a weight placed on each observation in the comparison sample in strata k

where $\sum_{i=1}^{N_k'} W_{ik}' = 1$ and $\sum_{j=1}^{N_k''} W_{jk}'' = 1$. In our first approach, our estimate of the treatment effect

within strata k is given by:

$$E_k(\Delta Y) = E_k(\Delta Y | P(X) \in P_1^K, P_2^K) = \sum_{i=1}^{N_k'} W_k' Y_{i1} - \sum_{j=1}^{N_k''} W_k'' Y_{j0} \quad (6)$$

where $W_k' = 1/N_k'$ and $W_k'' = 1/N_k''$. Our estimated average treatment effect across all strata is then given by:

$$E(\Delta Y) = \sum_k \frac{N_k'}{N} * E_k(\Delta Y). \quad (7)$$

In choosing P_1^k and P_2^k we follow the algorithm outlined in Dehajia and Wahba (1998). In particular, we choose P_1^k and P_2^k such that remaining differences in X between participants and nonparticipants within the strata are likely due to chance.

Our second approach is a slightly modified version of the kernel matching procedure developed by Heckman, Ichimura and Todd (1997), and Heckman, Ichimura, Smith and Todd

(1998). Our estimate of the treatment effect within the k^{th} strata is still given by (6). However, now our weights are given by:

$$W_i' = \frac{K(\tilde{P}(X) - P(X_i))}{\sum_{i=1}^{N_k'} K(\tilde{P}(X) - P(X_i))}, \quad W_j'' = \frac{K(\tilde{P}(X) - P(X_j))}{\sum_{j=1}^{N_k''} K(\tilde{P}(X) - P(X_j))} \tag{8}$$

where $\tilde{P}(X) = \frac{1}{2}(P_1^K + P_2^K)$ and K is a kernel. In general, a kernel is simply some distribution function such as the normal. In practice, the choice of P_1^k and P_2^k along with K are somewhat arbitrary. In our analysis we experiment with alternative choices and, as we indicate below, our results appear insensitive to our choice. In this second approach, our estimate of the average treatment effect is again given by (7), but with the weights defined by (8).

Control Variables

The assumption that outcomes are independent of treatment once we control for measured characteristics depends critically on the particular measured characteristics available. Any measured characteristic that is associated both with 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, residency in a large metropolitan area, as well as measures associated with eligibility in the program, which were prior year AFDC receipt and marital status.

Several recent analyses (Friedlander and Robins, 1995; Heckman Ichimura and Todd, 1997) 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 these 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 18 months prior to program participation are strongly associated with both program participation and expected labor market outcomes (Heckman, Ichimura and Todd, 1997; Heckman, Ichimura, Smith and Todd, 1998; Heckman, LaLonde and Smith, forthcoming).

Finally, Heckman, Ichimura and Todd (1997) have argued that differences in data sources, resulting from different data collection methods, are an important source of bias in attempts to estimate program impact using comparison groups.

III. The Data

The data for this project are all administrative data from Missouri deriving from three sources. The first is Missouri's JTPA program, from which we draw our sample of program participants. The second is Missouri's Division of Employment Security (ES), from which we

draw our comparison group sample. The final source is wage record data from the Unemployment Insurance program in Missouri. Using these data we obtain both pre- and post-enrollment earnings and information on labor force status prior to enrollment for both participants and nonparticipants.

The JTPA data include all individuals who apply to the JTPA program. The data consist of basic demographic and income information collected at the time of application which is used to assess eligibility, as well as information about any subsequent services received. Our initial sample consists of all applicants in program year 1995 (July, 1995 through June, 1996). We then eliminate records with missing or invalid values for social security number (SSN), race, sex, labor force status, Service Delivery Area (SDA), and highest grade completed. We also eliminate records for participants who live outside Missouri or who were never enrolled in the JTPA program (presumable because they were judged ineligible). Our final sample consists of data for 4050 adult males, 6669 adult females, and 706 youths.[3]

One of the primary tasks of ES in Missouri is to assist individuals in obtaining employment. Assistance can take a variety of forms such as maintaining a list of job openings in an area, helping individuals prepare a resume, or referring individuals to other agencies for more extensive training programs. During the time of our sample almost every individual who wanted to obtain services from ES applied at one of the Employment Security offices located around the state. The ES data contain basic demographic and income information obtained on the initial application, as well as information about subsequent services received. Our initial ES sample

[3] We define youths as being less than 22 years old when applying to the program. Youths include both males and females.

consists of all applicants who were registered with ES in program year 1995. From this original sample, we again eliminate records with missing or invalid values for SSN, race, sex, labor force status, SDA, and highest grade completed. In addition, we keep only individuals who are between 16 and 83 years old, who are not currently enrolled in JTPA, and who were not enrolled in JTPA either in the two previous program years (1993 and 1994) or in the subsequent program year (1996).[4] Finally, we keep only records for individuals who have been judged to be economically disadvantaged and therefore should be eligible for the JTPA program.[5] Our final ES sample consists of 23,706 adult males, 27,030 adult females and 12,228 youths.

The pre-enrollment and post-enrollment earnings for both our JTPA and ES sample come from the Unemployment Insurance (UI) data. These data consist of quarterly files containing total earnings for all individuals in the state employed in a job covered by the UI system.[6] Both the JTPA and ES data are matched to the UI data using SSN. If we are unable to match an SSN to earnings data in a quarter we consider the individual unemployed in that quarter and set earnings equal to zero.

Post-program earnings are measured as the sum of an individual's earnings in the four quarters starting two quarters after the quarter in which an individual enrolls in the program.

---

[4] We should point out that while we eliminate individuals in the ES sample enrolled in the JTPA program we do not eliminate individuals receiving training through other private or public training programs.

[5] While economically disadvantage individuals are eligible for JTPA, other workers are also eligible. Therefore, our JTPA file contains individuals who are not considered economically disadvantaged.

[6] For workers with earnings from multiple employers we sum up earning across employers so that we have only one record per worker.

Thus, if someone enrolls in either JTPA or ES in the third quarter of 1995, we start measuring earnings in the first quarter 1996.[7] We construct seven different measures of preprogram earnings. The first three are the earnings in each of the three quarters immediately prior to enrolling in the program.[8] The fourth measure is the mean quarterly earnings in the fourth through the eighth quarters preceding enrollment. We also construct the linear trend in wages for the eight quarters prior to enrollment and the standard deviation of earnings over the eight quarters. Finally we have a count of the number of quarters among the eight prior quarters for which earnings are zero.

Previous research (Heckman and Smith, 1999) found that prior labor market status dynamics is an important determinant of both program participation and subsequent earnings. We capture these dynamics using a series of four dummy variables. From both the JTPA and ES data we know whether or not an individual is employed at the time of enrollment. From the UI data we know whether or not 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 zero 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.

---

[7] When we began this project we did not have earnings data for the third quarter 1997. Therefore, for individuals who enrolled either program in second quarter 1996, we only have three quarters of earnings data. For these individuals, we multiply the three quarters of earnings by 1.25. In future versions of the paper we will include four quarters of earnings data for all workers.

[8] All of the earnings data have been converted into 1993 dollars.

Another variable previous research has found to be an important determinant of program participation is local labor market conditions (Heckman, Ichimura, Smith and Todd, 1998). We capture this effect by including a dummy variable for the Service Delivery Area where an individual lives.[9]

Our measure of labor market experience is defined as:

Experience = Age - Years of Education - 8 + ((8 - Unemploy)/4)

where **Unemploy** is defined as the total number of quarters unemployed in the eight quarters prior to enrollment. Since we know actual labor market experience in the two years prior to enrollment, we use this information to improve the more traditional measure of experience.

Table 1 presents summary statistics for our JTPA and ES samples separately for our 3 groups, adult males, adult females, and youths. This table shows that for adult males and females, the JTPA sample has both higher post-program and pre-program earnings than the ES sample. In addition, the JTPA sample is more likely to be white and to have graduated from high school, has more labor market experience, and is more likely to be employed both prior to and at the time of enrollment. We believe these differences are an artifact of the sample we have drawn. As indicated earlier, we keep only individuals that were considered economically disadvantaged in the ES sample, while we made no such exclusion in the JTPA sample. While being economically disadvantaged is one criteria for qualifying for JTPA, it is not the only criteria. Therefore, the JTPA sample contains individuals who are not classified as economically disadvantaged, and this fact is reflected in the numbers in Table 1. In future versions of the paper we plan on excluding the non-economically disadvantaged participants from our JTPA sample.

[9] There are 15 SDAs in Missouri.

However, we should note that, as long as the only criteria for being classified as economically disadvantages is prior income, since we control for prior earnings in our analysis, this selection should not affect our results. In addition, since we are primarily interested in examining the impact of various methods for constructing the comparison group, the fact that we may not have the appropriate treatment sample should not affect this comparison.

One of the conclusions reached by Heckman, LaLonde, and Smith in their chapter on program evaluation in the new *Handbook of Labor Economics* (Heckman, LaLonde, and Smith, 1999) is that "better data help a lot" (pg. xxx). The most important criteria they mention are that outcome variables should be measured in the same way for both participants and non-participants, that members of the treatment and comparison groups should be drawn from the same local labor market, and that the data should allow one to control for the dynamics of an individual's labor force status prior to enrollment. Since our data meet all of these criteria we feel they are 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.

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

Table 1 makes clear that the post-enrollment earnings differ dramatically for the JTPA and ES samples The mean differences in post-enrollment earnings between the two samples are

listed in line 1 of Table 2 for our three groups. Adult males in JTPA earn nearly $4000 more than those in the ES sample, while, for females, those in the JTPA earn $1400 more. In contrast, among youths, JTPA participants earn less than those in the ES sample, with the difference greater than $1100. Given the difference across groups in the mean values for other characteristics listed in Table 1, these earnings differences could easily be due to differences in the pre-program characteristics of the two samples.

Simple Regression Adjustment

Line 2 of Table 2 presents adjusted estimates of program effects based on the simple linear regression model. The structure of the model and coefficient estimates for the primary control variables, which are, for the most part, conventional human capital measures, are listed in Table 3. Coefficient estimates for the controls generally correspond to expectations.

Our coding of education allows for a separate high school graduate effect, as well as allowing for an additional increment for years of higher education. We see that, in this specification, it is high school graduation that captures the primary impact of schooling. It is somewhat surprising that the estimated impact of experience is negative for males. For females, however, the experience profile has the expected shape, implying a positive but decreasing impact of experience on earnings through most of the working life.

The measures of employment status transitions prior to enrollment have substantial impacts on post-program earnings. Those employed continuously during the two years prior to the enrollment date have the highest earnings, with lower earnings for those who moved into employment. Those never employed in the period (the omitted category) and those who lost jobs

had lower earnings, although relative earnings for these groups differs for the three groups. Most adults are in this last category, of job losers.

We have controlled for earnings using the seven earnings variables described in the previous section. The coefficient estimates for the prior earnings measures generally accord to expectations, with positive coefficients for earnings in the prior two quarters and for the average of the six preceding quarters. In contrast, the trend variable has an unexpected negative sign (although it is only statistically significant for adult males). The standard deviation has an unexpected positive estimated coefficient for males but the expected negative coefficient for females. The number of prior quarters with zero earnings has the expected negative impact for females and youths, but no effect for males.

It is clear that the regression-adjusted estimates of program impact for adults are appreciably smaller because the JTPA samples have much more advantaged backgrounds than the ES comparison sample. Clearly, higher prior earnings for this sample are most important in causing this change, although greater levels of education and a larger proportion of whites plays a role as well.

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. As noted above, even under the maintained assumption that there are no unmeasured factors that distinguish participants from the comparison group, if differences in measured variables are great enough, regression adjustment may be predicting outcomes for participants by extrapolation. The large size of our comparison group has substantial advantages but it also entails substantial risks of misspecification. In particular, if most of the comparison sample has characteristics that

are quite distinct from those of the participants, coefficients will be estimated based largely on relationships for individuals with very different characteristics from participants. If the functional relationships differ by values of *X*, the regression function may be poorly estimated. There are no assurance regarding the direction of the bias for such regression adjustment.

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. As our measure of similarity, we have chosen the Mahalanobis distance metric, since it has a number of attractive features, as noted above. For each participant, we choose a case from the comparison file for which the Mahalanobis distance is at its minimum, yielding a paired file. This one-to-one matching method ensures that if there is at least one individual in the comparison sample that is similar on all values to each participant, the resulting matched comparison group will display the same variable distribution.[10] In calculating the Mahalanobis distance, the characteristics in X' and X'' include education, race, prior experience, occupation (any versus none), our measures of employment status prior to enrollment (three dummy variables), dummy variables for whether an individual lived in either the St. Louis or Kansas City SDA, and our seven measures of pre-enrollment earnings.

Line 3 of Table 2 shows how post-program earnings differ between the participant file and the matched comparison file. If the regression adjustment was accurate, and if the Mahalanobis matching method produced a comparison sample very similar to the participant group, the estimates in lines 2 and 3 should be very close. In fact, the difference is substantial.

[10]The matching method used here is that described by Rosenbaum and Rubin (1985). We describe it in detail above, where we also consider an alternative algorithm.

In each case, the matched pair estimate is closer to the simple difference listed in line 1 than it is to the regression-adjusted estimate.

There are two possibly overlapping explanations for the discrepancy between the estimates in lines 2 and 3. First, if the pair matching was successful, the discrepancy would indicate that the regression specification inadequately captured nonlinear effects or interactions between independent variables. In this case, line 3 would be an appropriate estimate of program impact. The second possibility is that pairs are not very closely matched because the comparison sample does not contain a sufficient number of cases that are similar to those in the participant sample. In this case, the matched pair estimate could be seriously biased. If this were the case, we might be tempted to place greater confidence in the regression estimate, but failure of the pair matching is a warning that values of the control variables do not overlap in the participant and comparison sample. Insofar as there is little overlap, the danger of misspecification in the regression model may be substantial, reducing our confidence in line 2 estimates.

Propensity Score Matching

Matching cases on the basis of propensity score promises substantial simplification as compared with any general distance metric. The Mahalanobis distance between any pair of cases will only approach zero if all values of X are the same for both cases. In contrast, if cases are matched by propensity score, two cases with same propensity score will be matched perfectly even if values of X differ. Hence, the matching process is reduced to a single dimension. 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 (assuming that participation is independent of outcomes, conditional on observable characteristics).

The success of propensity score matching depends on the estimation of the propensity function. We use a logit function to predict participation in the sample combining the JTPA and ES samples. In addition to the variables listed in table 3, we tested nearly 300 interactions between these variables, using a stepwise procedure to enter any that were statistically significant at the 5 percent level

For each case, the predicted value based on the estimated logit function provides an estimate of $P(X)$. Table 4 lists the results of one-to-one matching based on the propensity score, providing a comparison with Mahalanobis distance matching. Comparison of lines 1 and 2 shows that the two matching methods produce very similar estimates for adult males and females. The estimates differ somewhat more for youths, but they are still in the same range.

Caliper matching differs from simple matching in that only matches within a specified distance are permitted, so not all participants may be matched. Lines 3-5 show how estimates differ when the caliper is set to 0.2, 0.1, and 0.05, respectively. The caliper of 0.2 removes three-fifths of the adult male participant cases, while it removes about two-fifths of female and youth cases. This clearly shows that the simple matching procedure matches many cases that are not very similar. Even with our very large comparison sample, more than half of the matched cases are not similar. The estimated impact of participation changes dramatically when we omit poor matches, declining by nearly 50 percent for males, by about 40 percent for females. The estimated impact for youths changes sign and is no longer statistically significant.

As the caliper is set more stringently, the effective sample size declines further, but the loss of cases is moderate. When the caliper is reduced from 0.2 to 0.1, estimates for males and females decline by about 10 percent. The estimate for youths increases by 50 percent, but the

increase is less than one standard error. Finally, reducing the caliper to 0.05 causes still smaller changes in the sample size and estimates. It would appear that reduction of the caliper to 0.1 is sufficient to assure that matched cases are comparable.

The sixth line of Table 4 shows how the estimate changes when a linear regression control is applied to the matched sample produced by the 0.1 caliper. The regression controls correspond to those listed in Table 3. The linear controls alter estimates for males and females by less that 10 percent, while the estimate for youths shifts substantially but by less than a standard error. It is clear that caliper matching, using a 0.1 caliper, is sufficient to produce a comparable sample. It does not appear that substantial differences remain between participants and the nonparticipants in terms of the variables that we have matched on.

Comparing One-to-One Matching Algorithms

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

Table 4 lines 7-9 presents results using this alternative matching technique. We found that the average difference in propensity scores between matched pairs was often appreciably smaller when this alternative was used. Nonetheless, it is clear that, for adult males and females, the effect of this alternative matching algorithm is small. The impact on estimates for youths is somewhat larger. In each case, the alternative algorithm matches fewer cases than does the

standard approach. This indicates that the search for the best match causes matches to be broken for cases which have no alternative comparison case within the caliper range.

In large part, differences in estimates for the conventional and alternative algorithm can be explained in terms the number of cases matched. Estimates that are produced using the alternative algorithm and a caliper of 0.2 are similar to estimates produced with the conventional approach using a caliper of 0.1, while estimates using the alternative algorithm and a caliper of 0.1 are similar to conventional estimates using a caliper of 0.05. It would appear that while in theory there might be some basis to prefer the alternative algorithm, in practice there is little impact on estimates.

Matching by Propensity Score Group

All of the one-to-one matching approaches described above have the important disadvantage that they require that we discard participants and comparison group members who are not matched. A first obvious limitation is that only one case from the larger sample can be used for each case in the smaller sample, resulting immediately in loss of information. Where the distribution of participants and the comparison groups differ dramatically, in the case of simple matching, the matches will be poor, whereas, for caliper matching, additional cases will be lost.

The results of the caliper matching reported above make clear that the participant and comparison groups differ dramatically on propensity score and thus on underlying variables. Figure 1 shows the distribution of the JTPA and ES samples by propensity score. The overwhelming majority of the ES cases for all three groups we are considering have probabilities of participation below 0.1, while the JTPA sample is distributed more uniformly over the full

range of propensity scores. This makes clear why caliper matching discards so many participants. Even though the ES sample is very large, in many propensity score categories, there are many participants for each ES sample member, so many of the participants must be discarded.

Group matching relaxes the requirement that the two groups be matched on a one-to-one 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 (1998).

Cases are first grouped on the basis of propensity score. Within a range of propensity scores, the mean difference in the outcome, post-program earnings, is taken as an estimate of $E(\Delta Y \mid P(X))$. These estimates are then averaged across all propensity groups, weighting each by the proportion of participants with scores in that range. The method requires that propensity ranges be small enough that there are no important difference between the participant and comparison groups but large enough that there are sufficient numbers of participants and comparison group members.

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 0.1 size. However, given the large number of cases with propensity values less than 0.1, we found that differences in our basic variables within this 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.[11]

The estimated program effects based on this approach are listed in Line 6 of Table 2. For males and females, the estimates are quite similar to those obtained using the caliper matching, and standard errors are somewhat larger. For youths, the estimate differs quite dramatically from the earlier one, and the standard error is somewhat smaller.

One explanation for the observed differences between estimates is that the one-to-one matching may have omitted a large share of participants for whom the program effects were different from others. Figure 2 plots the program effects for our three groups by propensity score. We see that estimates do vary by propensity score, and that these differences may be of some importance.

Kernel Density Matching

The estimates based on propensity score grouping use estimates of $E(\Delta Y|P)$ that are simple sample averages that combine cases with similar values of P. Following an approach recommended by Heckman, Ichimura and Todd (1997) and Heckman, Ichimura, Smith and Todd (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. In forming our estimates we experimented with a variety of different kernels and we

[11]We tested for statistically significant differences between participants and the comparison group on the propensity score, years of education, experience, the three employment dummies, race, occupation (any versus none), dummies for the Kansas City and St. Louis SDAs, three dummies for quarter of enrollment, earnings in the two quarters prior to enrollment, average earnings in the third through eighth quarters prior to enrollment, and the standard deviation of this average. Fewer than 10 percent of the differences were statistically significant (at the 5 percent level) for the final category widths we used.

considered bandwidths from 0.01 to 0.1. We found that the choice of kernel and bandwidth made little difference in our estimates. Therefore, we report estimates for each group based on a bi-weight kernel using the bandwidth calculated as optimal for each group under standard assumptions.[12] These bandwidths are 0.5 for males, 0.4 for females, and 0.7 for youths. The results are reported in Line 7 in Table 2.

Table 2 makes clear that estimates based on this approach are very similar for males and females to those based on one-to-one matching with a caliper of 0.1 and with the estimates produced when matching by propensity score group. In contrast, the estimate for youths is discrepant with that in line 5, but is similar to the estimate in line 6. Overall, the results from Table 2 suggests that, at least for men and women, conditional on comparing comparable individuals, estimates of the effect of the program on participants are relatively insensitive to the methods used to construct the comparison group.

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

Table 5 compares our estimated program effects with those reported in Orr, el. al. (1996) which are based on an experimental evaluation of the JTPA program. For both Men and Youths our estimates are larger than those reported in Orr, et. al. while our estimates for women are smaller. However, this is one place where our sample selection criteria is likely making a difference. The Orr, et. al. sample just includes economically disadvantaged workers, whereas our JTPA sample includes both disadvantaged and other eligible workers.

[12] The bandwidth used is the width that would minimize the mean integrated squared error if the data had a Gaussian distribution and the Gaussian kernel were used.

IV. Conclusion

Results in the previous section show that when estimating the effect of a program on participants, it is vital that one compare comparable individuals. This is consistent with the results from previous research (Heckman, Ichimura, and Todd, 1997; Heckman, Ichimura, Smith and Todd, 1998; Heckman LaLonde and Smith, 1999). However, conditional on comparing comparable individuals, our results also suggests that estimates of the effect of the program on participants are relatively insensitive to the method used to form the nonexperimental comparison group. This leads us to two conclusions. First, since matching by propensity group is the simplest method, we suggest using this method when forming comparison groups. Second, conditional on having data with a similar set of observable characteristics for individuals, our results show that it is possible for other states to use existing administrative data to estimate the impact of training programs on the participants in these programs.

References

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

Ashenfelter, Orley and David Card. "Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs." Review of Economics and Statistics, 67 (November 1985), pp. 648-60.

Bassi, L. "Estimating the Effect of Training Program with Non-random Selection," *Review of Economics and Statistics*, 66 (February 1984): 36-43.

Dehejia, Rajeev H. and Wahba, Sadek. "Causal Effects in Non-Experimental Studies: Re-Evaluating the Evaluation of Training Programs." Unpublished paper, Columbia University, April 1998.

Friedlander, Daniel and Robins, Phillip. "Evaluating Program Evaluations: New Evidence on commonly Used Nonexperimental Methods." *American Economic Review* 85 (September 1995): 923-937.

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

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

Heckman, James J. and Jeffery A. Smith, "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.

Heckman, J., H. Ichimura, J. Smith, and P. Todd. "Characterizing Selection Bias Using Experimental Data." *Econometrica*, 66 (September 1998): 1017-1098.

Heckman, J., H. Ichimura, and P. Todd, "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme." *Review of Economics Studies*, 64, (1997): 605-654.

Lalonde, Robert J. "Evaluating the Econometric Evaluations of Training Programs with Experimental Data." *American Economic Review*, 76, (September 1986), pp. 604-20.

Manski, Charles F. “Learning About Treatment Effects from Experiments with Random Assignment of Treatments.” Journal of Human Resources, 31, (Fall, 1996): 709-33.

Orr, Larry L., Howard Bloom, Stephen Bell, Fred Doolittle, Winston Lin, George Cave. *Does Training for the Disadvantaged Work? Evidence from the National JTPA Study*. Washington D.C.: The Urban Institute Press, 1995.

Rosenbaum, P. and D. Rubin. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70 (1983): 41-55.

Rosenbaum, P. and D. Rubin. “Constructing a Control Group Using Multivariate Matched Sampling Methods the Incorporate the Propensity Score.” *The American Statistician*, Vol. 39 (February 1985): 33-38.

Table 1: Summary Statistics

	Adult Males		Adult Females		Youths	
	JTPA	ES	JTPA	ES	JTPA	ES
Percent High School Grad	86.8	72.9	86.4	75.5	30.3	52.1
Experience	15.6	14.3	13.6	12.9	0.42	0.97
Experience^2	382.4	294.6	344.0	262.4	1.15	2.65
Percent White (non Hispanic)	72.7	68.3	72.5	64.3	69.1	68.5
Percent Male	1.0	1.0	0.0	0.0	53.8	42.8
Labor Market Transitions						
Percent Not Empl./Empl	5.1	6.4	6.5	6.9	4.0	9.2
Percent Empl./Empl.	5.6	5.2	7.2	5.6	2.1	3.4
Percent Empl./Not Empl.	77.7	70.1	71.1	67.4	46.3	66.3
Percent Not Empl./Not Empl.	17.2	18.3	15.2	20.1	47.6	21.1
Percent in Kansas City SDA	17.9	11.8	16.9	13.2	14.9	10.7
Percent in St. Louis SDA	14.0	14.4	10.7	14.6	11.0	11.8
Mean post enrollment earnings	10296.29	6367.57	6576.60	5150.19	2586.51	3716.11
Mean earnings one quarter prior to enrollment	2869.69	1816.31	1896.76	1422.45	490.06	815.92
Mean earnings two quarters prior to enrollment	3157.01	1832.60	2044.06	1424.14	506.90	745.85
Mean earnings three quarters enrollment	3356.08	1781.95	2094.68	1387.07	485.02	678.10
Mean earnings four to eight quarters prior to enrollment	3396.27	1575.19	2105.86	1201.10	339.78	478.04
Mean estimated probability of participation	42.7	9.8	61.7	9.5	50.8	2.8
Number	4050	23760	6669	27030	706	12288

Table 2: Summary of Estimates of Program Effects

	Adult Males		Adult Females		Youths	
	Estimate	(Std. Err.)	Estimate	(Std. Err.)	Estimate	(Std. Err.)
1. Simple Difference	3928.72	(181.07)	1426.41	(95.34)	-1129.60	(159.44)
2. Regression Adjustment	1209.11	(134.69)	139.69	(92.92)	-213.27	(162.97)
One-to-One Matching Methods						
3. Mahalanobis Distance Matching	3164.14	(196.25)	1202.59	(107.25)	-653.05	(189.18)
4. Propensity Score Matching	3143.28	(218.28)	1072.42	(116.57)	-933.65	(220.72)
5. Propensity Score, .10 Caliper	1448.10	(222.78)	582.23	(145.56)	431.34	(271.28)
Matching for Groups						
6. Propensity Score Categories	1692.86	(289.87)	569.20	(162.63)	-19.83	(198.01)
7. Propensity Score Kernel Density Matching	1720.18	(274.79)	605.63	(169.64)	-61.95	(238.27)

Table 3: Regressions Predicting Post-Program Earnings

	Adult Males	Adult Females	Youths
	Coefficient (Standard Error)	Coefficient (Standard Error)	Coefficient (Standard Error)
Participation in JTPA	1209.11 (134.64)	139.69 (92.92)	-213.27 (162.97)
Years of Education	68.49 (58.74)	21.96 (47.35)	164.70 (47.35)
High School Graduation	870.85 (164.42)	606.39 (126.29)	800.72 (113.46)
Years of Higher Education	131.79 (69.27)	456.04 (54.18)	109.88 (92.14)
Experience	-24.20 (15.81)	49.07 (10.71)	-52.38 (66.66)
Experience^2	-1.04 (.38)	-1.81 (.25)	13.24 (14.75)
Not Employed/ Employed	1540.57 (202.98)	1693.75 (137.50)	1335.37 (136.47)
Employed/ Employed	2948.67 (249.41)	2629.68 (167.38)	1904.84 (221.81)
Employed/ Not Employed	-673.36 (147.41)	75.45 (99.07)	217.05 (98.32)
White	871.54 (115.04)	-218.31 (80.64)	335.55 (89.48)
Male	-	-	470.01 (75.92)
Enrollment 95:3	-103.09 (128.27)	-248.33 (85.93)	200.66 (96.15)
Enrollment 95:4	-44.42 (130.57)	-260.33 (90.98)	-18.45 (101.31)
Enrollment 96:1	268.61 (131.13)	55.69 (91.46)	212.45 (102.34)
Earnings 1 Quarter Prior	.6311 (.0288)	.3450 (.0199)	.7583 (.0485)
Earnings 2 Quarters Prior	.2206 (.0321)	.3035 (.0272)	.1510 (.0582)
Average Earnings Prior Quarters 3-8	.5686 (.0377)	.4991 (.0359)	.8765 (.0861)
Trend in Quarters 3-8	-.3838 (.0897)	-.1378 (.0837)	-.1053 (.1681)
Standard Deviation in Quarters 3-8	.4458 (.1853)	-1.0487 (.1398)	-.8681 (.3531)
Prior Quarters Zero Earnings	.6771 (33.87)	-1.0488 (.1398)	-149.50 (29.58)
9 Occupation Dummies	X	X	X
14 Service Delivery Area Dummies	X	X	X
Adjusted R^2	.2532	.2049	.2496
N	27810	33699	12994

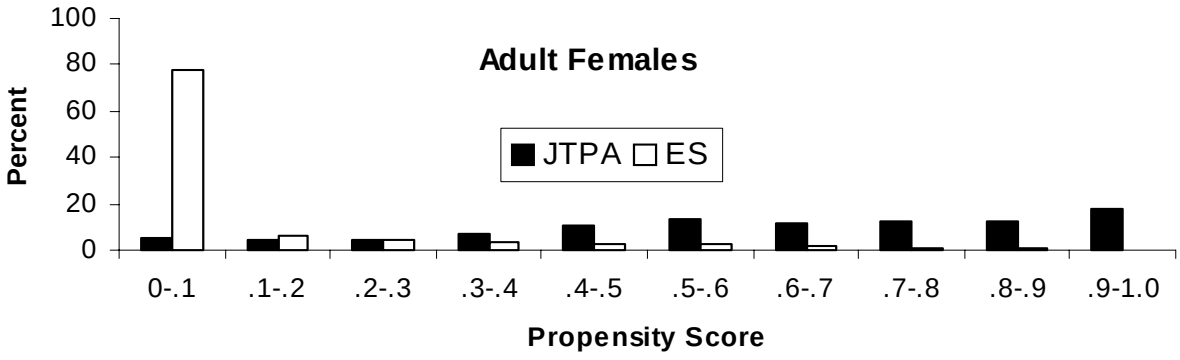
Table 4: Estimates of Program Effects Using One-to-One Matching Methods

	Adult Males		Adult Females		Youths	
	Estimate	(SE) N	Estimate	(SE) N	Estimate	(SE) N
1. Mahalanobis Distance Matching	3164.14	(196.25) 8100	1202.59	(107.25) 6669	-653.05	(189.18) 706
Propensity Score Matching						
2. Simple Matching	3143.28	(218.28) 8100	1072.42	(116.57) 6669	-933.65	(220.72) 706
3. Caliper = 0.2	1604.70	(228.88) 3270	637.68	(141.71) 4071	287.14	(269.94) 447
4. Caliper = 0.1	1448.10	(222.78) 3138	582.23	(145.56) 3774	431.34	(271.28) 416
5. Caliper = 0.05	1487.16	(225.33) 3097	552.52	(149.33) 3684	527.58	(279.14) 404
6. Caliper = 0.1 with Regression Adjustment	1474.16	(196.55) 3138	628.77	(136.69) 3774	293.87	(257.78) 416
7. Caliper = 0.2, modified matching method	1447.92	(219.38) 3178	525.03	(145.84) 3815	346.83	(275.26) 423
8. Caliper = 0.1, modified matching method	1472.50	(223.11) 3108	542.49	(146.62) 3687	452.47	(282.79) 403
9. Caliper = 0.05, modified matching method	1486.81	(225.27) 3091	563.00	(148.16) 3670	452.47	(282.79) 403

Table 5: Estimated Program Effects Based on Randomized Control Groups

		Current Analysis		
	Orr, et. al. (1996)	Propensity Score, 0.10 Caliper	Propensity Score Categories	Kernel Density Matching
Men	$970	$1448	$1693	$1720
Women	$960	$582	$569	$605
Youths	-$171	$431	-$20	-$62

Figure 1: Propensity Score Distribution for JTPA and ES Samples



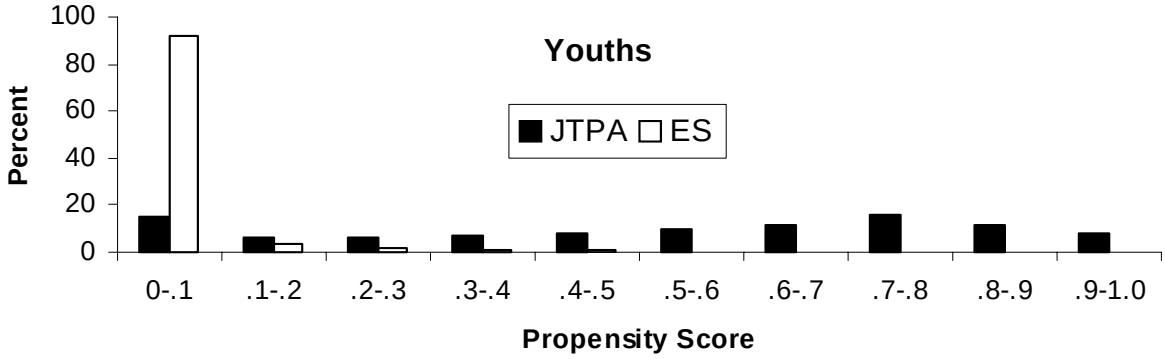


Figure 2: Estimates of Program Effect by Propensity Score