SENSITIVITY TESTING OF NET IMPACT ESTIMATES OF WORKFORCE DEVELOPMENT PROGRAMS USING ADMINISTRATIVE DATA

Upjohn Institute Staff Working Paper No. 08-139

Kevin Hollenbeck
W.E. Upjohn Institute for Employment Research
Hollenbeck@upjohninstitute.org

February 2008

ABSTRACT

This paper addresses the question of whether administrative data sources, such as performance monitoring data, can be used for program evaluation purposes. It argues that under certain circumstances, such data can be used. In particular, program performance data that are routinely gathered and monitored by administrators of many workforce development programs meet these circumstances. The paper goes on to demonstrate the point by using administrative data from the state of Washington to examine the net impact on earnings and employment of services provided to adults under the Workforce Investment Act (WIA).

Because of a lack of consensus about appropriate net impact estimators, the strategy of this paper is to examine the sensitivity of the results to various estimation techniques. The paper describes the various estimation techniques, and it summarizes the net impact estimates that are generated for the State of Washington. For the most part, the results are fairly stable across the techniques, which the paper argues adds a degree of confidence in them. The final section of the paper offers guidance to policymakers and program administrators who may not be familiar with the technical details of various analytical approaches about how empirical results that may appear to be complex or unstable can be used for program improvement.

JEL Classification: J24, J48, C81

This paper is a revised version of a paper presented at the 3rd Annual IZA Conference on the Evaluation of Labor Market Programs, at IZA, Bonn, on October 19, 2007. Discussant comments at that conference by Ulf Rinne, IZA, are gratefully acknowledged. The paper draws on results presented at the 2004 National Workforce Investment Research Colloquium, sponsored by the U.S. Department of Labor, Employment and Training Administration, Arlington, VA: May 2004. Support from the U.S. Department of Labor is gratefully acknowledged. Thanks to Timothy Bartik, Randall Eberts, Carolyn Heinrich, Chris O'Leary, Jeff Smith, and David Stevens for comments. Finally, Wei-Jang Huang and Claire Black provided excellent assistance with empirical work and preparation of the paper. The viewpoints expressed are solely the author's and do not necessarily represent the views of the U.S. Department of Labor or the W.E. Upjohn Institute.

SENSITIVITY TESTING OF NET IMPACT ESTIMATES OF WORKFORCE DEVELOPMENT PROGRAMS USING ADMINISTRATIVE DATA

The purpose of this paper is to address the question of whether administrative data sources, such as performance monitoring data, can be used for program evaluation purposes. It argues that under certain circumstances such data can be used. In particular, program performance data that are routinely gathered and monitored by administrators of many workforce development programs meet these circumstances. The paper goes on to demonstrate the point by using administrative data from the state of Washington to examine services provided to adults under the Workforce Investment Act (WIA). Using the lingo of individuals who have formalized evaluation studies (e.g., Rossi and Freeman 1993), the work presented here uses a quasi-experimental method relying on *ex post* data.

A considerable literature has arisen concerning the various empirical techniques used in quasi-experimental evaluations (see the February 2004 *Review of Economics and Statistics* collection of papers and the many studies referenced there).[1] The general theme of this literature seems to be that there are many different econometric techniques for estimating program effectiveness that have appropriate asymptotic properties. Some papers in this literature go on to speculate about which estimators seem to work best under which conditions.

Because of the lack of a consensus about appropriate estimators, the strategy of this paper is to examine the sensitivity of the results to various estimation techniques.[2] The paper describes the various estimation techniques, and it summarizes the net impact estimators that have been generated for the state of Washington. For the most part, the results are fairly stable across the techniques; I argue that this adds a degree of confidence to them. The final section offers

[1] One of the articles in that collection, Michalopoulos, Bloom, and Hill (2004), addresses the question that is central to this paper, namely the advisability of using administrative data for program evaluation purposes.

[2] The approach in this paper is very similar to work described in Mueser, Troske, and Gorislavsky (2003).

guidance to policymakers and program administrators who may not be familiar with the technical details of various analytical approaches about how empirical results that may appear to be complex or unstable can be used for program improvement.

1. INTRODUCTION

Analyses of quantitative data about workforce development programs are valuable to at least two audiences: 1) individuals charged with administering the programs and 2) entities that invest resources in the programs. The first group, administrators, are accountable for the results of their programs and want to make sure that they are achieving maximum results given the resources they have. The second, investors (or funders), want to make sure that they are maximizing their returns on investment. Like ship captains, program administrators set directions and objectives to be reached, and they must get feedback to determine if and when directional adjustments need to be made. I use the term *performance monitoring* to refer to this kind of feedback. The owners of the shipping company, on the other hand, want to know their returns on investment in order to allocate or reallocate their resources. I use the term *net impact evaluation* to refer to this kind of information. A question that this paper addresses is whether performance monitoring data can be used for net impact evaluation.

The empirical results presented pertain to WIA as administered in Washington State during the program year July 2000 to June 2001. However, the evaluation purposes and methods discussed in the paper are relevant to a gamut of workforce development programs: federal job training programs such as WIA, formal postsecondary educational programs such as community colleges or four-year colleges and universities, apprenticeships, adult basic education, formal or informal on-the-job training, and secondary career and technical education.

2. NET IMPACT EVALUATION VERSUS PERFORMANCE MONITORING

Many references provide excellent discussions of social program evaluation (see, for example, Blalock 1990; Rossi and Freeman 1993; Mohr 1992; Wholey, Hatry, and Newcomer 1994). The emphasis of much of this literature is on the design of an evaluation for which the evaluator has control over the data collection. However, less attention has been paid to the role of performance monitoring in program evaluation. In recent years, performance monitoring has become an integral part of program administration as public resources have become tighter and tighter, forcing administrators to be held more and more accountable to measurable performance standards. A fortunate by-product of performance monitoring is the considerable individual-level data that have become available, which may be used for evaluation purposes as well.

2.1 Performance Monitoring

The purpose of performance monitoring is to measure the usage of resources and the flow of clients in order to manage as effectively as possible the resources that are available. In general, administrators are concerned about *efficiency*, which is providing the greatest amount and highest quality of service given the level of resources, and about *equity*, which is providing services fairly. Administrators need to ensure that the characteristics that are being measured (to which measurements the administrators are being held accountable) are important, not just things that are easily measured. Furthermore, administrators need to ensure that measures are consistently defined over a sufficient length of time to have some confidence in their levels and trends.

Performance monitoring is most useful when the information can be benchmarked. That is, administrators who are undertaking performance monitoring in order to improve their programs' effectiveness will need to make judgments about trends or levels in the data.

Benchmarks, which are summaries of comparable indicators from other agencies and programs or from other time periods, can be used to formulate those judgments. Performance standards are intended to be a method of benchmarking performance data.[3]

In short, the purpose of performance monitoring is to inform program improvement. The audience for such monitoring is administrators.

2.2 Net Impact Evaluation

The purpose of a net impact evaluation is to evaluate the outcomes of the program for participants relative to what would have occurred if the program did not exist. In other words, it answers the question of how the program has changed the lives of individuals who participated in it relative to their next best alternative. The data that are used to address this question are quantitative, and the evaluation should attempt to examine results by subgroup because there may be systematic relationships between program outcomes and participant characteristics. The audiences for a net impact evaluation are the funding agency (or agencies) and the program administrators. For publicly funded workforce development programs, the owners are the taxpayers, and their agents are state or federal legislators or evaluation branches of the executive agencies.

The attribution of the net impacts to the program intervention is confounded by at least four factors. The first factor is definition of the treatment. Social programs usually tailor services to the individuals being served. Thus, each participant may receive slightly different services. Furthermore, participants control their effort. So, even if participants are given the same treatment, they may exert more or less effort in learning and applying the skills or knowledge being delivered to them. Furthermore, some individuals may not complete the treatment.

[3] Heckman, Heinrich, and Smith (2002) provide a thorough analysis of the impact of performance standards, which tend to focus on short-run outcomes—i.e., on actual performance.

Second, in order to estimate the *net* impacts of a program, it is necessary to compare program participants to another group of individuals who represent the counterfactual—i.e., what would have happened to the participants absent the program. Designation of that comparison group and, concomitantly, having adequate data concerning members of the group are crucial for estimating net impacts. However, acquiring the data may be difficult because the comparison group members did not receive the treatment.

The third factor that may confound attribution is the definition and measurement of the outcomes. Performance measurement is aimed at inflows to and outflows from a program, whereas evaluation is likely to focus on outcomes after clients have received the treatment. The performance measurement system may not be designed to collect such information.

Finally, the fourth factor is that the dynamics of program interventions and outcomes may make attribution difficult. In particular, receiving the treatment may require a significant amount of time. So the question becomes whether outcomes should be measured after program entrance or after the treatment ends. (Furthermore, individuals who receive the treatment may not complete the program.) Observations that are well matched at the time of program entrance may differ considerably if the reference point is program exit simply because of the business cycle or other changes that may occur over time.

The four conditions, then, that must be met in order to use administrative, performance-monitoring data for evaluation purposes are as follows:

1) The treatment is defined in a general enough fashion to be meaningful for a sizable group of program participants. But, of course, the more general the definition of the treatment, the less useful it may be for program improvement purposes.
2) Administrative data must be available for a group of individuals that arguably make up a reasonable comparison group.
3) Outcome data must be available for both the treatment and the comparison group.

4) The time periods of observation and treatment for program participants and the comparison group must be reasonably close to each other, so that meaningful outcome comparisons can be made.

3. THE NET IMPACT EVALUATION PROBLEM AND KEY ASSUMPTIONS

3.1 Statement of Problem

The net impact evaluation problem may be stated as follows: Individual i, who has characteristics X_{it}, will be observed to have outcome(s) $Y_{it}(1)$ if he or she receives a treatment, such as participating in a training activity, at time t and will be observed to have outcome(s) $Y_{it}(0)$ if he or she doesn't participate. The net impact of the treatment for individual i is $Y_{it}(1) - Y_{it}(0)$. But, of course, this difference is never observed because an individual cannot simultaneously receive and not receive the treatment.

To simplify the notation without loss of generality, I will omit the time subscript in the following discussion. Let $W_i = 1$ if individual i receives the treatment, and $W_i = 0$ if i does not receive the treatment. Let T represent the data set with observations about individuals who receive the treatment for whom we have data, and let n_T represent the number of individuals with data in T. Let U represent the data set with observations about individuals who may be similar to individuals who received the treatment for whom we have data, and let n_U be its sample size. In some of the techniques described below, I identify a subset of U that contains observations that match those in T. I call this subset C, so let n_C be its sample size. The names that I use for these three data sets are Treatment sample (T), Comparison sample (U), and Matched Comparison sample (C).

Receiving the treatment is assumed to be a random event—individuals happened to be in the right place at the right time to learn about the program, or the individuals may have

experienced randomly the eligibility criteria for the program—so W_i is a stochastic outcome that can be represented as follows:

$$(1) \qquad W_i = g(X_i, e_i) \,,$$

where e_i is a random variable that includes unobserved or unobservable characteristics about individual i as well as a purely random component. An assumption that I make about $g(\bullet)$ is that $0 < \text{prob}(W_i = 1|X_i) < 1$. This is referred to as the "support" or "overlap" condition that is necessary so that the outcome functions described below are defined for all X.[4]

In general, outcomes are also assumed to be stochastically generated. As individuals in the treatment group encounter the treatment, they gain certain skills and knowledge and encounter certain networks of individuals. I assume their outcomes are generated by the following mapping:

$$(2) \qquad Y_i(1) = f_1(X_i) + e_{1i} \,.$$

Individuals not in the treatment group progress through time and also achieve certain outcomes according to another stochastic process, as follows:

$$(3) \qquad Y_i(0) = f_0(X_i) + e_{0i} \,.$$

Let $f_k(X_i) = E(Y_i(k)|X_i)$, so that e_{ki} is a deviation from expected values that reflects unobserved or unobservable characteristics for $k = 1,0$.

As mentioned, the problem is that $Y_i(1)$ and $Y_i(0)$ are never observed simultaneously. What is observed is the following:

$$(4) \qquad Y_i = (1 - W_i)Y_i(0) + W_iY_i(1) \,.$$

The expected value for the net impact of the treatment on the sample of individuals treated is as follows:

[4] Note that Imbens (2004) shows that this condition can be slightly weakened to $\Pr(W_i = 1|X_i) < 1$.

$$
\begin{aligned}
E[Y_i(1) - Y_i(0) \,|\, X, W_i = 1] &= E(\Delta Y \,|\, X, W = 1) \\
&= E[Y(1) \,|\, X, W = 1] - E[Y(0) \,|\, X, W = 0] \\
&\quad + E[Y(0) \,|\, X, W = 0] - E[Y(0) \,|\, X, W = 1] \\
&= \hat{f}_1(X) - \hat{f}_0(X) + \text{BIAS},
\end{aligned} \tag{5}
$$

where $\hat{f}_k(X)$, $k = 1,0$ are the outcome means for the comparison and treatment group samples, respectively, and BIAS represents the expected difference in the $Y(0)$ outcome conditional on X between the comparison group (actually observed) and the treatment group (the counterfactual).

A key assumption that allows estimation of Equation (5) is that $Y(0) \perp W|X$. This orthogonality assumption states that, given X, the outcome (absent the treatment), $Y(0)$, is not correlated with participation. This is equivalent to the assumption that participation in the treatment can be explained by X up to a random error term. In other words, there is no deterministic function of X that perfectly predicts participation (or nonparticipation). The assumption is called "unconfoundedness," "conditional independence," or "selection on observables." If the assumption holds, then the net impact is identified because BIAS goes to 0, or

$$
E[\Delta\, Y|X,\, W = 1] = \hat{f}_1(X) - \hat{f}_0(X)\,. \tag{6}
$$

In random assignment, the X and W are uncorrelated through experimental control, so the conditional independence assumption holds by design. In any other design, the conditional independence is an empirical question. Whether or not the data come from a random assignment experiment, however, because the orthogonality assumption holds asymptotically, in practice, it may make sense to regression-adjust Equation (6).

3.2 Regression and Quasi-experimental Estimation of Net Impacts

Burtless and Greenberg (2004) address the use of random assignment experiments to estimate the net impacts of programs. Clearly, a well-conducted experiment is the best solution to the attribution problem because it builds the assumption of unconfoundedness into the design. However, as many evaluators have pointed out, social experimentation is difficult to implement with total control and is therefore fraught with potential threats to validity. Furthermore, as Hollenbeck, King, and Schroeder (2003) point out, an experimental design may not be feasible for entitlement programs or may be prohibitively costly.

In short, for the purposes of this paper, I assume that experimental data are unavailable. Instead, I assume that I have one data set that contains information about individuals who have encountered a treatment (presumably collected as part of a performance monitoring system) and another data set that contains information about individuals who may comprise a comparison group for the treatment cases. The question that I address in this section is, "How should I proceed to derive defensible estimates of the net impact of the treatment?"

Figure 1 depicts the situation. The vertical axis suggests that there are eligibility conditions to meet in order to gain access to the treatment, which I assume is participation in a workforce development program. Individuals may be more or less eligible, depending on their employment situation or their location or other characteristics such as age or family income. The X-axis measures participation likelihood. Individuals who are highly eligible (observations that would be arrayed near the top of the graph) may or may not participate. On the other hand, individuals who are not eligible (near the bottom of the graph) may or may not have the desire to participate.

T represents the data set with treatment observations, and *U* represents the data set from which the comparison set of observations may be chosen. Note that *T* and *U* may come from the same source of data, or they may be entirely different data sets. In the former situation, *U* has been purged of all observations that are also in *T*.

Various estimation techniques have been suggested in the literature, but they may be boiled down to two possibilities: 1) use all of the *U* set, or 2) try to find observations in *U* that closely match observations in *T*. Note that identification of the treatment effect requires that none of the covariates *X* in the data sets are perfectly correlated with being in *T* or *U*. That is, given any observation X_i , the probability of that covariate being in *T* or in *U* is between 0 and 1. I will call techniques that use all of *U* full sample techniques.[5] Techniques that attempt to find matching observations will be called matching techniques. Each will be described in turn.

Full sample estimators. Assuming that *T* and *U* have some resemblance to each other, the evaluator should calculate the simple difference in means of the outcome variables as a baseline estimator.[6] This estimator essentially assumes away selection bias. It may be represented as follows:

$$(7) \qquad \tau = \frac{1}{n_T} \sum_{i \in T} Y_1(1) - \frac{1}{n_U} \sum_{i \in U} Y_j(0) \ .$$

This estimator can be regression-adjusted. If we assume that the same functional form holds for both *Y*(1) and *Y*(0), then the treatment effect can be estimated from a linear equation such as the following using the observations in the union of *T* and *U*:

$$(8) \qquad Y_i = a + B'X_i + \tau W_i + e_i \ .$$

[5] Some of these techniques trim or delete observations from *U* in order for support reasons, but I will still refer to them as full sample techniques.

[6] In comments on this paper, David Stevens points out that its emphasis is on the traditional focus of net-impact mean value estimates. David encourages readers to not neglect analysis of outliers, an evaluation focus that has been around for decades. He cites Klitgaard and Hall (1973, 1975).

More generally, τ can be estimated by using two separate regression functions for the two regimes ($Y(1)$ regressed on X in T and $Y(0)$ regressed on X in U), using both models to predict a "treated" and a "nontreated" outcome for all observations in both T and U.[7] The following average treatment effect can then be calculated:

$$(9) \qquad \tau = \frac{1}{N} \sum_{i \in T,U} \left[\hat{f}_1(X_i) - \hat{f}_0(X_i) \right],$$

where $N = n_T + n_U$ and $\hat{f}_k(X_i)$ is the predicted value for $k = 1, 0$.

Equation (8) and the more general regression in the first stage of Equation (9) require strong parameterization assumptions. Heckman et al. (1998) relax those assumptions in a nonparametric kernel method. This method amounts to weighting the observations in U so that the observations closest to the treatment observations receive the highest weights. This estimator may be written as follows (following Imbens 2004):

$$(10) \qquad \hat{f}_k(X_1) = \frac{\sum_j Y_j K\left(\frac{X_j - X_i}{h} \right)}{\sum_j K\left(\frac{X_j - X_i}{h} \right)} \text{ for } k = 1, 0,$$

where $j \in T$ if $k = 1$ and $j \in U$ if $k = 0$ and $K(\bullet)$ is a kernel function with bandwidth h. Thus,

$$(11) \qquad \tau = \frac{1}{N} \sum_i \left[\hat{f}_1(X_i) - \hat{f}_0(X_i) \right].$$

Several of the full sample estimators rely on the observations' propensity scores, which are the estimated probabilities of being in the treatment group. Rosenbaum and Rubin (1983)

[7] Imbens (2004) points out this generalization. The intuition is similar to that of the basic Roy (1951) model with two regimes and individuals pursuing the regime for which they have a comparative advantage. However, Imbens (2004) notes, "These simple regression estimators may be very sensitive to differences in the covariate distributions for treated and control units" (p.12). I produced these estimates in the empirical work, but the estimators and standard errors did not seem to make sense and were quite different from all other estimates. The regression parameters were quite unstable when estimated with full comparison and treatment samples. Consequently, I have not presented these results.

showed that the conditional independence assumption, $Y(0) \perp W|X$, implies that $Y(0) \perp W|p(X)$, where $p(X)$ is the conditional probability of receiving the treatment (= Prob($W = 1|X$)).

This result implies that the regression approaches in Equations (8) through (10) can be reestimated, at reduced dimensionality, with the X_i replaced by $p(X_i)$. That is, estimates can be generated as follows:

(8′) $Y_i = a + B'p(X_i) + \tau W_i + e_i$;

(9′) $$\tau = \frac{1}{N} \sum_{i \in T,U} \left[\left(\hat{f}_1 \left(p\left(X_i\right) \right) - \hat{f}_0 \left(p\left(X_i\right) \right) \right) \right] ;$$

(10′) $$\hat{f}_k \left(X_i\right) = \frac{\sum_j Y_j K \left(\frac{p\left(X_j\right) - p\left(X_i\right)}{h} \right)}{\sum_j K \left(\frac{p\left(X_j\right) - p\left(X_i\right)}{h} \right)} \text{ for } k = 1, 0 .$$

The final type of full sample estimator is computed by a technique known as blocking on the propensity score (see Dehejia and Wahba 1998). The intuition here is to partition the union of the treatment and the full sample into "blocks" or strata by propensity score, so that there is no statistical difference between the covariates, X, in each block. This essentially achieves the conditional independence assumption locally in each block. Then the average treatment effect is a weighted average of the treatment effects in each block.

Assume there are K blocks. Let the kth block be defined as all treatment or full comparison sample cases with values of X, so that $p(X) \in [p_{1k}, p_{2k}]$. Let NT_k be the number of treatment cases in the kth block and NU_k be the number of comparison cases from the full sample. The treatment effect with each block k is:

$$(12) \qquad \tau_k = \sum_{\substack{i=1 \\ i \in T}}^{NT_k} \frac{1}{NT_k} Y_i(1) - \sum_{\substack{j=1 \\ j \in U}}^{NU_k} \frac{1}{NU_k} Y_j(0) \ ,$$

and the overall estimated average treatment effect is:

$$(13) \qquad \tau = \sum_{k=1}^{K} \frac{NT_k}{N} \tau_k \ .$$

Matching estimators. As above, *U* denotes the set of observations from which I will choose the subset *C* (for matched comparison group) that will be used in the net impact analyses. The idea is to have *C* be composed of the observations where individuals are most "like" the individuals who make up *T*. Matching adds a whole new layer of complexity to the net impact estimation problem. The estimator becomes a function of how the match is done in addition to the characteristics of the sample. Since the matching process is a structured algorithm specified by the analyst, the statistical error associated with the net impact estimator now includes a component that may be identified as matching error in addition to the sampling error and model specification error.[8]

There is a substantial and growing literature on how to sample individuals to construct the comparison sample.[9] The first candidate approach is *cell-matching algorithms*. Variables that are common to both data sets would be used to partition (cross-tabulate) the data into cells. Then for each treatment observation the cell would be randomly sampled (with or without replacement) to select a comparison group observation. A substantial drawback to cell-matching is that the cross-tabulation of data, if there are many common variables, may result in small or empty cells.[10]

[8] This forces the analyst to use bootstrapping techniques to calculate standard errors.
[9] See Heckman, Lalonde, and Smith (1999) and references cited there.
[10] Center for the Study of Human Resources (1994) used a variation of this approach.

More sophisticated comparison group construction can be accomplished with *nearest-neighbor algorithms*. These algorithms minimize a distance metric between observations in T and U. If we let X represent the vector of variables that are common to both T and U, and let X_j, X_k be the values of X taken on by the jth observation in T and the kth observation in U, then C will be composed of the k observations in U that minimize the distance metric $|(X_j - X_k)|$ for all j. This approach is very mechanistic, but it does allow the use of all of the X variables.

The literature usually suggests that the distance metric be a weighted least squares distance, $(X_j - X_k)' \Sigma^{-1} (X_j - X_k)$, where Σ^{-1} is the inverse of the covariance matrix of X in the comparison sample. This is called the Mahalanobis metric. If we assume that the X_js are uncorrelated, then this metric simply becomes least-squared error. Imbens (2004) has a discussion of the effect of using different metrics, although in practice the Mahalanobis metric is used most often.[11]

In his work on training-program evaluation, Ashenfelter (1978) demonstrates that participants' preprogram earnings usually decrease just prior to enrollment in a program. This implies that a potential problem with the nearest-neighbor approach is that individuals whose earnings have dipped might be matched with individuals whose earnings have not. Thus, even though their earnings *levels* would be close, these individuals would not be good comparison-group matches.

An alternative nearest-neighbor type of algorithm involves use of propensity scores (see Dehejia and Wahba 1995). Essentially, observations in T and U are pooled, and the probability of being in T is estimated using logistic regression. The predicted probability is called a propensity

[11] Note that Zhao (2004) uses a metric that weights distances by the coefficients in the propensity score logit. This is similar to the technique that Schroeder implements in Hollenbeck, King, and Schroeder (2003).

score. Treatment observations are matched to the observations in the comparison sample with the closest propensity scores.

An important consideration in implementing the matching approach is whether to sample from *U* with or without replacement. Sampling with replacement reduces the distance between the treatment and comparison group cases, but it may result in the use of multiple repetitions of observations, which may artificially dampen the standard error of the net impact estimator. Another consideration is the number of cases to use from *U* in constructing *C*. Commonly, matching is done on a one-to-one basis, where the nearest neighbor is chosen. However, it is also possible to take multiple nearest neighbors. In the empirical work below, I experiment with one-to-five and one-to-ten matching.

The whole reason for matching is to find similar observations in the comparison group to those in the treatment group when the "overlap" or statistical support is weak. Consequently, the nearest-neighbor approach may be adjusted to require that the distance between the observations that are paired be less than some criterion distance. This is called *caliper* or *radii matching*.

Once the matched sample *C* has been constructed, the net impact estimation can be done using the estimators analogous to those in Equations (8) through (11). The outcome variable can be in terms of levels or difference-in-differences if the underlying data are longitudinal.

4. EMPIRICAL ESTIMATION OF THE NET IMPACT OF WIA SERVICES

4.1 Data

The treatment in this section of the paper is receipt of WIA intensive or training services by adults[12] who exited from WIA in Program Year 2000 (July 2000–June 2001) in the state of Washington. The counterfactual that I am using to construct a comparison group is that if there

[12] Note that I am only looking at individuals served in the adult program, not dislocated workers or youth.

were no WIA services, then individuals would receive services through the State Employment Service (i.e., Wagner-Peyser services).[13] Thus the pool of observations from which we construct the comparison groups is composed of individuals whose last reported service date in the Employment Service (ES) data was in the same program year. The administrative data from the WIA program and from the Employment Service have been linked to Unemployment Insurance wage records dating from 1990:Q1 through 2002:Q2.[14]

The empirical analyses are intended to be illustrative of the stability of the net impact estimates to various full sample or matched sample estimation techniques. So I have reduced the underlying data sets in two ways. First, I have reserved a randomly chosen 25 percent of the treatment data set for specification testing. Second, I have chosen half of the ES sample for use in the estimation in order to conserve on computational time. Table 1 presents descriptive data for the three samples by sex.

The table shows that the observations in the data from the Employment Service are substantially different from the treatment observations in both preprogram characteristics and outcomes. Between 2 and 3 percent of the comparison sample is disabled, compared to over 20 percent of the males in the treatment sample and about 15 percent of the females. Furthermore, a much higher percentage of comparison sample observations have educational attainment beyond a high school diploma. The employment and earnings histories of the individuals from the comparison pool are also quite different, although at the time of registration virtually none of the ES observations were employed, whereas one-sixth of the males and one-fourth of the females

[13] Carolyn Heinrich has pointed out that an implicit assumption in this empirical work is that the Employment Service (ES) is the "next best alternative" for WIA clients. If, in fact, WIA participants could have fared better in the labor market with no government assistance or with the assistance of some other institution than with the ES, then the net impact estimates are biased upward.

[14] Note that in much of the analysis described in this paper, I refer to preregistration employment and earnings data. To construct these variables, I used wage record data that started in 1997:Q3. Furthermore, note that Washington has an interstate agreement with contiguous states and Alaska to share wage record information for individuals who reside in Washington but work in one of these states.

that received training or intensive services from WIA were employed at the time of registration. Prior to program entry, the comparison sample's employment rate was almost 90 percent, with an average quarterly earnings of almost $6,400 for males and more than $5,000 for females. The WIA exiters' preprogram employment rate was about 75 percent, and average quarterly earnings were about $2,900 for males and $2,000 for females.

Table 1 displays descriptive statistics concerning outcomes as well as preprogram characteristics. Earnings, as measured by average quarterly earnings in the fourth quarter after leaving the program and as measured by average quarterly earnings for all quarters after leaving the program are higher for the comparison group than for the treatment group. However, the differences are not nearly as large as the differences in preprogram earnings. Furthermore, the differences in the employment rates after the program are virtually nil. Thus one expects that the difference-in-differences for earnings and employment would show that the treatment group did much better than the comparison group, which it does.

Figures 2 through 5 display the data for key outcome variables. The first two figures show quarterly earnings for males and females, respectively. The time series presented in these figures portray average quarterly earnings for the comparison and treatment samples prior to program entry and after program exit. As expected, the earnings for the comparison sample are much higher than those of the treatment sample prior to program entry. This reflects a more favorable labor market history and human capital characteristics. Note that the figures show the earnings dip that occurs prior to registration. Figures 4 and 5 show employment rates for the groups, where employment is defined as quarterly earnings exceeding $100. Again, the left side of the graphs shows the trend prior to program registration, and the right side pertains to the

quarters after program exit. As with the other figures, the comparison sample exhibits far higher employment rates than does the treatment sample.

4.2 Full Sample Estimators of Net Impact

Tables 2 and 3 provide estimates of the net impact of the treatment (which is having received WIA Intensive or Training Services) using several of the full sample estimation techniques for males and females, respectively. The first row of the table shows the simple differences in means between the treatment sample and the comparison sample. Columns (1) and (3) show the differences in the levels of the outcome variables, and we know from Table 1 that these will be negative and quite large because the comparison group had higher education levels and preprogram earnings and employment histories than the treatment sample. The entries in columns (2) and (4) show the mean of the difference-in-differences, and, as in Table 1, the employment and earnings advantages for the comparison group outcomes were not nearly as large as the preprogram differences, so the difference-in-differences are quite large and positive.[15]

The estimates in the first row are simply for baseline descriptive purposes because of the significant differences in the samples. The second row of the table regression-adjusts the results from the first row. For the most part, this reduces the magnitudes of the estimates significantly. The covariates used in the regression were measured at the time of registration with WIA or the ES. They are as follows: age, race/ethnicity, educational attainment, veteran status, disability status, limited English proficiency, employment status at registration, industry of current or most recent employment, labor market area, and employment and earnings history. Hollenbeck and Huang (2003) summarize the employment and earnings histories of individuals using the

[15] All of the earnings impacts in this paper are denominated in constant 2000 dollars.

following five variables: 1) percentage of quarters employed since entering employment, 2) conditional average earnings (preprogram), 3) trend in earnings levels (constant dollars), 4) variance in earnings levels, and 5) turnover. In this paper, I use these variables plus a measure of the preprogram dip in earnings that may have occurred in the preprogram earnings history.[16]

The third row of the table is another regression-adjustment technique in which I have substituted the propensity scores for the covariates in the model used in row (2). So in this row the estimators are regression-adjusted using a model with only two independent variables—1) propensity score and 2) treatment. As would be expected, the standard errors of the estimates increase significantly relative to the full regression model, although the estimates are not all that different qualitatively.

The next three rows show estimates derived using a kernel density nonparametric regression approach. Each row uses a different bandwidth for the basic Epanechnikov kernel. Mueser, Troske, and Gorislavsky (2003) and Imbens (2004) suggest that the bandwidth does not make much difference in the estimation, but the results here seem to indicate that bandwidth variation does make a considerable difference. With the exception of the postprogram employment rate, increasing the bandwidth significantly increases the magnitude of the estimates.

The last row of the table shows estimates that were calculated using the propensity score blocking approach. The algorithm that we employed in this approach uses the full comparison sample in principle, although we do trim some observations to guarantee full overlap. In particular, observations are eliminated from U if their p-score < min (p-score) for T, and observations are eliminated from T if their p-score > max (p-score) for U. We then "blocked" the

[16] The earnings dip variable is defined as max($0, [average quarterly earnings in preregistration quarters −3 to −8 minus average quarterly earnings in preregistration quarters −1 to −2]).

file into *p*-score deciles and performed an *F*-test to determine whether the distribution of key covariates (age, education, employment status at registration, race, and preemployment variables) were independent. If the *F*-test failed for any group, we split the cells in half and tested the new cells. The average treatment effects in the seventh row of the table are weighted averages of the cell-by-cell treatment effects, in which the weights are the proportion of treatment observations in the cell. The estimates, which are in the range of 15 to 20 percent for earnings and 10 to 15 percent for employment, are similar to the regression-adjusted estimates.

4.3 Matched Sample Estimators

Several different matched sample estimators were calculated. All of the approaches estimated the treatment effect by computing the average difference in outcomes for the treatment sample and the matched sample; they also estimated the treatment effect by adjusting those estimates by regression. Standard errors were estimated for the mean differences by bootstrapping with 100 replications. The standard errors for the regression-adjusted estimators come directly from the regression.

Match quality indicators and specification testing. Most of the matched sample estimators presented in this paper use a propensity score approach. This approach relies on predicted probabilities of being in the treatment. To compute these probabilities for each observation, I estimated a logit model with a binary dependent variable indicating whether the observation came from the treatment sample or not. I used the parameters estimated from this model to calculate a propensity score (*p*-score) for all observations in the treatment sample (*T*) and in the comparison sample (*U)*. These *p*-scores remained fixed on an observation-by-observation basis throughout the analyses to eliminate a source of variation in the estimators that are being compared.

When using a quasiexperimental, matched sample estimation technique, it is important to try to demonstrate the "quality" of the match. Several indicators are used in this paper. First of all, for *p*-score matching, I present the mean difference in the *p*-scores. Since the whole purpose of the matched sample estimation is to find observations that are as comparable as possible to the treatment cases, the smaller the mean difference, the higher the quality of the match, other things being equal. Next, I present the percentage of comparison sample observations that are unique (used only once in the match). For the matching without replacement estimators, this is 100.0 percent by construction. For the estimators derived by matching with replacement, higher percentages indicate that there were fewer cases used and that they were used more than once. The matching with replacement estimators yield lower mean differences in *p*-scores (higher quality), but using the same observation more than once will artificially reduce the variance and bias the standard error estimates. So, in comparing two matches done with replacement, the one with the higher percentage of unique cases is likely to be a higher quality match.

By reserving a quarter of the treatment sample, I am able to conduct specification testing on the matched comparison samples. Specifically, I conduct two *F*-tests to test the joint dependence between the matched comparison sample and the "reserved" subsample of the treatment cases. One of the *F*-tests uses all of the covariates available, and the other tests for joint dependence of only the six preregistration employment and earnings variables.

A final test of the "overlap" between the treatment sample and the comparison sample (recall that we assume that $0 < \text{prob [participation} \mid X] < 1$) is a test that I refer to as the 20th percentile indicator. This is the percentile of the *p*-score distribution for the comparison sample (*U*) at the first quintile point in the *p*-score distribution in the treatment sample. If the participation in the treatment model is "good," then most of the *p*-scores for treatment cases will

be near 1.0; and most of the *p*-scores for the comparison cases will be near 0. The mean for the former is expected to be much larger than the mean for the latter. Battelle Memorial Institute (n.d.) undertook an evaluation study using matched sample estimation and asserted that a reasonable assurance of overlap is that the *p*-score that identifies the lowest quintile of *p*-scores for the treatment sample should approximate the 80th percentile of the *p*-scores for the matched comparison set. The Battelle study does not really justify this assertion, but it turns out that the propensity estimates are very close to 80 percent—80.9 percent for males and 83.5 percent for females.

Characteristics matching. The first set of estimators that I present construct the matched comparison set by minimizing distances between characteristics using a Mahalanobis distance metric. The matching was done with replacement on a one-to-one basis. Tables 4 and 5 provide these estimates and the match quality indicators for males and females, respectively. For reference purposes, the first row of each table repeats the regression-adjusted difference in means for the full comparison sample. The second and third rows of the tables give the difference in means and the regression-adjusted difference in means for the matched comparison group and the treatment sample. Most of the estimates for females are statistically significant, and the regression-adjusted estimates are quite large in magnitude. For males, the earnings outcomes are not statistically significant, but the employment rate estimates are significant.

As far as match quality goes, the preponderance of matched comparison set records are unique (used only once), although the percentage of observations used more than once for females is quite a bit higher than for males. The specification tests show that these matched samples do not replicate well the distribution of covariates in the treatment subsample that we reserved for such testing.

In short, this form of matched file estimation is probably not the preferred specification. The net impact estimates seemed to bounce around quite a bit, and the specification test failed. Note that other types of characteristics matching may provide much more stable estimates.

***P*-score matching.** In these techniques, observations in the treatment sample are matched to their nearest neighbors using differences in *p*-score values. Tables 6 and 7 show the impact of using this technique with and without replacement when the minimization is done for males and females, respectively. Note that the mean of the (absolute value of) the *p*-score differences is almost three times larger for the without-replacement estimator than for the one done with replacement. The estimated treatment effects for both procedures are reasonably similar, although the magnitudes of the estimates "with replacement" are usually larger. Seven of the eight estimates for females are statistically significant for the *p*-score matching with replacement.

In terms of match quality, as noted, the *p*-scores are much "closer" for matching with replacement. For both males and females, the percent of comparison observations that were used multiple times is not large, and the specification test shows that the distributions of the preregistration employment and earnings variables are independent for females. The specification tests are not consistent with statistical independence for males.

In Tables 8 and 9, I display the sensitivity of the impact estimators to the number of comparison sample observations chosen to match each treatment case. In particular, I show one-to-one, one-to-five, and one-to-ten nearest-neighbor estimates. Choosing more "nearest neighbors" seems to decrease the treatment effects on earnings for males, as well as their standard errors. The employment rate impacts are larger; however, again they have smaller standard errors. The picture is almost the exact opposite for females: the earnings estimates increase slightly with more nearest neighbors chosen, and the employment impacts decrease

slightly. Of course, the standard errors decrease for females when more nearest neighbors are chosen, as they do for males.

The match quality statistics conform to expectations. Choosing more observations to match causes the mean of the *p*-score differences to increase. The mean differences for the estimators using one-to-ten are three times as great as the mean differences for the one-to-one estimators. Furthermore, considerably fewer comparison file observations are used uniquely in the techniques that are one-to-many, and the maximum repetitions are quite large (especially for females). The specification tests for females indicate that the matched comparison sets do a good job of replicating the treatment subsample distribution of the preregistration employment and earnings variables for females, but the specification tests suggest systematic differences in the distribution for males.

Caliper matching. The purpose of the matching techniques is to find the observations in the comparison sample that most closely match the treatment cases. Empirically, it may turn out that for some observations in the treatment sample, there may not be close matches. Caliper (or radius) matching deletes from consideration matches where the distance between the treatment observation and its nearest neighbor exceeds a particular distance. This distance is the caliper or radius, and it is arbitrarily set. I demonstrate the effect of the caliper on the matching estimates in Tables 10 through 13. In the first two tables, I use calipers of 0.005 and 0.01 on the nearest-neighbor matching that was done with replacement. For males, these particular calipers do not change the estimates much. The treatment effects and standard errors in the second two panels of Table 10 are very similar to the estimates in the top panel, which were computed without a caliper. The match quality statistics are also quite comparable, although the mean *p*-score difference falls by almost 80 percent with the most binding caliper of 0.005, even though only 10

matches were deleted with this caliper. The outlying *p*-score differences in the top panel, the maximum of which was 0.0793, skew the mean difference considerably.

These particular calipers are more binding for females, and indeed, the estimates in the bottom two panels of Table 11 exhibit larger differences from the top panel than the corresponding panels in Table 10 for males. All of the estimates are attenuated toward 0, and the earnings estimates become statistically insignificant. As was the case for males in the previous table, the average *p*-score difference dropped dramatically; the mean in the bottom panel with the most binding caliper is 0.0003, compared to 0.0025 in the top panel. In this case, 37 matches (almost 10 percent) were deleted.

In Tables 12 and 13, I display the effects of calipers on results that were estimated by matching without replacement. In general, matches without replacement are not as close as matches with replacement, so the effects of using calipers are more dramatic. The results for males, displayed in Table 12, actually show fairly stable results across the three panels. The estimates decline slightly with the caliper of 0.01, but then they increase generally with the more binding caliper of 0.005. The average difference in *p*-scores tumbles by 90 percent, from 0.0031 to 0.0003, although the number of matches that are deleted is not great—9 and 15 for the less binding and more binding calipers, respectively. The effects of the calipers on estimates for females are similarly not all that large in magnitude, but in this case the calipers delete almost 15 percent of the matches—59 and 66 for the 0.01 and 0.005 calipers, respectively.

In short, the effects of using calipers on the *p*-score nearest-neighbor matches in this sample are not very large in magnitude, whether the match is with or without replacement. The use of calipers eliminates some matches that are not very close, but the treatment effects for these matches apparently do not vary greatly from the overall average treatment effects.

4.4 Summary of Net Impact Estimates

Tables 2 through 13 provide several dozen estimates of net impact estimates that exhibit significant variation. The question remains of whether there is enough stability or overlap in the estimates to draw a reasonable inference about the net impacts of WIA intensive or training services on adult clients in Washington who exited WIA in its first full year of implementation, i.e., PY 2000. Table 14 displays results from the previous tables that address this question. The columns in this table look at outcomes that have been calculated by using difference-in-differences. Both sexes are displayed in the table.

As a point of reference, the simple differences in means from the full sample are provided in the first row. In this particular sample, these differences are quite large and positive. They do not make reasonable estimates of the treatment effect, however, because the treatment and control samples were quite different prior to the program, as demonstrated in Table 1. So, the question becomes how best to estimate the treatment effect. The estimates in rows (2) through (5) are some of the full sample estimates, and those in rows (6) through (11) are some of the matched sample estimates. Note that all of these estimates come from a single set of data, so they are not independent pieces of information. The bottom row of the table provides means of the outcome variables for the preprogram period, which I display so that the treatment effects can be considered in percentage terms.

All of the earnings impacts presented in the table are positive for males, although only one of them is statistically significant at the 0.05 level. (Many of them are significant at the 0.10 level, however). The magnitudes of the estimates range from $166 to $553 in quarterly earnings. The mean of average quarterly earnings prior to the program is approximately $2,900, so this range corresponds to percentage increases of approximately 6 to 18 percent. The entries in the

second column of the table display estimates of the net impact on employment. In this case, many of the estimates are significant. They range from 5.5 to 12.3 percentage points. These impacts, on a percentage basis, range from about 7 to 16 percent. Consequently, these estimates suggest that WIA intensive and training services in Washington State in PY2000 had an impact on the earnings of adult males of approximately 10 to 12 percent; this impact appears mainly to result from these services' impact on employment.

All of the earnings impacts for females are also positive, and they are of larger magnitude than the estimates for males. Many of them are statistically significant at the 0.05 level. The magnitudes range from $391 to $894; the amounts correspond to effects that are between 20 to 45 percent. All of the employment impacts for females are significant, ranging from 5.0 to 17.2 percentage points. On a percentage basis, these employment impacts range from about 6 to 24 percent. Because the employment rate impacts are smaller than the earnings impacts, it must be the case that the program had positive net impacts on wage rates or hours worked. In short, these estimates suggest that WIA intensive and training services in Washington State in PY2000 had an impact on the earnings of adult females of approximately 20 to 25 percent—a difference that results from these services' impact on employment and on either wages or hours or both.

5. POLICY AND PROGRAMMATIC IMPLICATIONS

The empirical section of this study presented literally hundreds of estimates using different techniques to try to tease out the net impact of WIA. In this last section of the paper, I am going to try to take the perspective of a policymaker or program administrator who is confronted with all of these estimates, many of which are denoted as being significant. The question is, "What is such a policymaker or administrator to do with all of these results?" I am going to assume that this individual is interested in improving her program, and that she wants to

use results from empirical analyses of data as warranted. However, this individual has limited expertise in statistical analyses of data and wants to rely on studies done by experts. I am also going to assume that the studies being considered have gone through a peer review process and have achieved a level of professional acceptance. If this last assumption does not hold, the policymaker should be extremely cautious about relying on any findings.

I believe that there are six key principles that such an individual needs to keep in mind when considering the findings from studies.

Principle 1: Since all study results have some degree of uncertainty no matter what methodology is used, always consider the costs associated with Type I and Type II errors before instigating a programmatic change based on study findings.

The null hypothesis in a program evaluation would be that the treatment has no effect. A Type I error would mean rejecting a true null hypothesis. (If a Type I error has been made, then a false positive has been identified; i.e., the study found a significant treatment effect that was, in fact, not true.) A Type II error would mean accepting a null as true when in fact it is false. (This would be a false negative; i.e., the treatment-effect findings are not significant statistically, when in fact the null was false.) It is usually the case that Type I errors are much more expensive than Type II errors because they involve changing the status quo. Thus the administrator should be especially conservative or cautious with a study, such as the present one, that finds significant impacts, in case there turn out to be Type I errors.

Principle 2: Insist on multiple answers. Do not make high-stakes decisions based on a single study.

Policymakers or program administrators would only be considering major changes if they had been given a credible study that had convincing evidence. However, even in this case, the decision maker should actively seek out other sources of information, including qualitative data from staff persons and clients, before taking any sort of major programmatic action.

Principle 3: For quasi-experiments, insist on documentation of match quality. The author of the study needs to present evidence of sufficient overlap and, if possible, specification testing that confirms conditional independence.

Other things being equal, the validity of the estimates likely will increase with sample size, amount of overlap in covariates between the treatment and comparison samples, and similarity of the treatment and comparison samples. A consensus has formed around the notion that when employment-related outcomes are examined it is critical to require matches within—or at least control for—local labor market areas.

Principle 4: Apply the "smell" test.

Do the estimates seem reasonable? In all likelihood, the net impact of a program or change in a program on a particular outcome will be directly proportional to the size of the treatment. If only small, marginal changes are being made, or if the resources invested per recipient are modest, then the net impacts are likely to be modest also. This study presented estimated net impacts on earnings that were around 10 percent for males, and perhaps double that for females. Net impacts this large border on unreasonableness and should be considered with a healthy skepticism.

Principle 5: Insist on getting estimates of statistical uncertainty.

Policymakers and program administrators want to know *the* answer. But there will always be sources of error in the analyses of social programs because of the stochastic nature of client-program interaction, changes in the overall labor market, and pure chance. Furthermore, data generally come from samples of populations, so there is sampling error as well. When considering the size of an impact, it is always important to assess magnitudes within the context of the estimated statistical uncertainty.

Principle 6: Stability of estimates is probably good, but hard to assess.

First of all, the notion of stability has to be judged relative to the perturbation that has been introduced in order to compute different estimates. For instance, some of the estimates in this paper use entirely different estimation techniques and samples. (An example would be regression-adjusted full sample differences in means versus regression-adjusted matched sample differences in means when matching is done with replacement and selection of the 10 nearest neighbors for each treatment observation.) In other cases we made minor changes, such as trying a caliper of 0.005 instead of 0.01. Other things being equal, it is probably the case that stable estimates are more likely to approximate truth when the stability occurs in the presence of multiple data sets or substantially different estimation techniques. One should have less confidence in the results if the only results that are presented are stable but only minor estimation changes have been attempted, or if the results are not very stable when there are significant differences in the estimation techniques. One should be least comfortable with results that are highly variant to what appear to be minor changes in the estimation technique or samples.

6. SUMMARY

With government resources scarce, more and more emphasis has been placed on accountability and demonstrated return on investment. This trend, along with the dramatic decreases in the cost of information processing, has led to striking advances in the availability of program administrative data and to a demand for net impact evaluation. This paper demonstrates that administrative data can be used to support the hard, quantitative data demands of net impact estimation.

The paper has described a number of full sample and matched sample techniques for estimating the net impacts of workforce development programs. Furthermore, it provides empirical estimates of the impact of WIA services for adults in the state of Washington using several of these approaches. Virtually all of the techniques yielded estimates of positive impacts for both men and women. Men had earnings gains on the order of 10 percent that appear to have resulted mainly from increased employment rates of approximately the same amount. Women had larger earnings gains—perhaps 20 to 25 percent—that emanated from increased employment *and* wages or hours. These impacts are large and should be accepted with some caution. The final substantive section of the paper provides six principles that policymakers should apply when considering evaluation results in order to exercise an appropriate amount of healthy skepticism and caution.

References

Ashenfelter, Orley C. 1978. "Estimating the Effect of Training Programs on Earnings." *Review of Economics and Statistics* 60(1): 47–57.

Battelle Memorial Institute. n.d. "Net Impact Evaluation: Appendix A, Technical Appendix." Unpublished report. Seattle, WA: Battelle Memorial Institute.

Blalock, Ann Bonar, ed. 1990. *Evaluating Social Programs at the State and Local Level: The JTPA Evaluation Design Project*. Kalamazoo, MI: W. E. Upjohn Institute for Employment Research.

Burtless, Gary, and David Greenberg. 2004. "Evaluating Workforce Programs Using Experimental Methods." Paper presented at the 2004 National Workforce Investment Research Colloquium, held in Arlington, VA, May 24.

Center for the Study of Human Resources. 1994. *Texas JOBS Program Evaluation: Final Report*. Austin, TX: Lyndon B. Johnson School of Public Affairs, University of Texas at Austin.

Dehejia, Rajeev H., and Sadek Wahba. 1995. "Causal Effects in Non-Experimental Studies: Re-Evaluating the Evaluation of Training Programs." NBER Working Paper No. W6586. Cambridge, MA: National Bureau of Economic Research.

———. 1998. "Propensity Score Matching Methods for Non-experimental Causal Studies." NBER Working Paper No. 6829. Cambridge, MA: National Bureau of Economic Research.

Heckman, James J., Carolyn Heinrich, and Jeffery A. Smith. 2002. "The Performance of Performance Standards." *Journal of Human Resources* 37(4): 778–811.

Heckman, James J., Hidehiko Ichimura, Jeffrey A. Smith, and Petra Todd. 1998. "Characterizing Selection Bias Using Experimental Data." *Econometrica* 66(5): 1017–1098.

Heckman, James J., Robert J. LaLonde, and Jeffrey A. Smith. 1999. "The Economics and Econometrics of Active Labor Market Programs." In *Handbook of Labor Economics, Vol. 3A,* Orley C. Ashenfelter and David Card, eds. Amsterdam: Elsevier, pp. 1866–2097.

Hollenbeck, Kevin M., and Wei-Jang Huang. 2003. *Net Impact and Benefit-Cost Estimates of the Workforce Development System in Washington State.* Upjohn Institute Technical Report No. TR03-018. Kalamazoo, MI: W. E. Upjohn Institute for Employment Research.

Hollenbeck, Kevin M., Christopher T. King, and Daniel Schroeder. 2003. "Preliminary WIA Net Impact Estimates: Administrative Records Opportunities and Limitations." Paper presented at the Bureau of Labor Statistics and the Workforce Information Council symposium "New Tools for a New Era!" held in Washington, DC, July 23–24.

Imbens, Guido W. 2004. "Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review." *Review of Economics and Statistics*, 86(1): 4–29.

Klitgaard, Robert E., and George R. Hall. 1973. *A Statistical Search for Unusually Effective Schools*. RAND Report No. R-1210-CC/RC. Santa Monica, CA: RAND. http://www.rand.org/cgi-bin/Abstracts/e-getabbydoc.pl?R-1210.

———. 1975. "Are There Unusually Effective Schools?" *Journal of Human Resources* 10(1): 90–106.

Michalopoulos, Charles, Howard S. Bloom, and Carolyn J. Hill. 2004. "Can Propensity-Score Methods Match the Findings from a Random Assignment Evaluation of Mandatory Welfare-to-Work Programs?" *Review of Economics and Statistics* 86(1): 156–179.

Mohr, Lawrence B. 1992. *Impact Analysis for Program Evaluation*. Thousand Oaks, CA: Sage.

Mueser, Peter R., Kenneth R. Troske, and Alexey Gorislavsky. 2003. *Using State Administrative Data to Measure Program Performance.* IZA Discussion Paper No. 786. Bonn, Germany: Institute for the Study of Labor.

Rosenbaum, Paul R., and Donald B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70(1): 41–55.

Rossi, Peter H., and Howard E. Freeman. 1993. *Evaluation: A Systematic Approach, 5th ed*. Thousand Oaks, CA: Sage Publications.

Roy, Andrew D. 1951. "Some Thoughts on the Distribution of Earnings." *Oxford Economic Papers* 3(2): 135–146.

Wholey, Joseph S., Harry P. Hatry, and Kathryn E. Newcomer, eds. 1994. *Handbook of Practical Program Evaluation*. San Francisco: Jossey-Bass.

Zhao, Zhong. 2004. "Using Matching to Estimate Treatment Effects: Data Requirements, Matching Metrics, and Monte Carol Evidence." *Review of Economics and Statistics* 86(1): 91–107.

Table 1 Summary Statistics

Characteristic	Male			Female		
	Treatment Sample		Comparison (ES) sample	Treatment Sample		Comparison (ES) sample
	Spec. testing subsample	Analysis subsample		Spec. testing subsample	Analysis subsample	
Age (years)	34.2	35.5	37.0**	35.9	36.1	38.3**
Disability	21.2	20.2	2.7**	11.9	17.4	2.2**
White	72.7	73.6	75.1	72.2	74.4	78.1
Veteran	24.2	20.2	12.8**	1.6	2.1	1.6
Limited English proficiency (LEP)	6.1	8.2	5.9	2.4	7.2	4.8
Education Completed						
< high school	17.1	17.8	15.8	9.5	13.1	12.3
High school	56.6	49.0	41.1**	50.8	50.0	37.2**
> high school	26.3	33.9	43.1**	39.7	37.1	50.5**
Employed at reg.	16.2	18.2	1.1**	27.8	24.8	1.1**
Preprogram Employment						
Employment rate (%)	73.2	73.1	87.7**	74.7	74.4	88.5**
Avg. earnings	2609.1	2908.7	6398.1**	1860.2	2008.9	5059.5**
Earnings trend	−243.5	−173.3	197.4**	−67.3	−100.6	177.9**
Variance earnings	4.73	5.70	12.90**	1.79	2.95	7.23
Percent of employed qtrs. w/ mult. employers	22.7	22.2	17.1**	23.1	21.1	16.7**
Earnings dip, mean	1670.2	1388.3	671.5**	608.6**	973.1	523.6
Outcomes						
Earnings in Quarter 4	2746.5	2844.0	4235.1**	2474.6	2460.1	3602.0**
Avg. earnings	4122.7	4176.5	6299.3**	3713.9	3593.0	5099.3**
Employment rate (%)	62.6	65.1	66.4**	64.3	66.8	67.3
Difference in earnings	−653.3	−1143.7	−2964.3**	175.8	−26.7	−2083.4**
Difference in avg. earnings	435.2	230.7	−920.7**	1247.3	946.2	−596.4**
Difference in empl. rate	−5.1	−0.6	−15.1**	−0.2	2.4	−15.5**
Ever employed (%)	58.6	61.6	63.2	57.1	61.6	65.4
Sample size	99	292	39,241	126	391	28,733

NOTE: Treatment samples are observations from PY2000 WIASRD file that reported receiving intensive or training services. These observations were randomly divided into an analysis subsample (75 percent) and a specification testing subsample (25 percent). Comparison samples are a random 50 percent sample from ES records.
** represents means that are statistically significantly different from the analysis subsample at the $p < 0.05$ level.

Table 2 Net Impact Estimates Using Full Sample Estimation Techniques, Males

Estimator	Outcome			
	Postprogram earnings (4th qtr.)	Difference-in-Difference	Postprogram employment rate	Difference-in-Difference
(1) Difference in means (baseline)	−1391.2***	1818.6***	−1.3	14.5***
	(194.0)	(253.1)	(2.4)	(3.1)
Regression adjustment				
(2) Regression adjustment	197.9	314.7	4.3	5.5**
	(258.8)	(258.0)	(2.4)	(2.6)
(3) Regression adjustment (*p*-score as sole regressor)	302.8	166.5	7.1***	8.4***
	(288.0)	(386.9)	(2.5)	(2.8)
Kernel density estimation				
(4) Bandwidth = 0.01	−31.3	552.6**	6.1**	8.7***
	(205.3)	(269.9)	(2.5)	(3.1)
(5) Bandwidth = 0.05	−701.4***	1131.0***	2.4	9.8***
	(199.3)	(264.6)	(2.4)	(3.0)
(6) Bandwidth = 0.10	−883.6***	1342.7***	1.9	11.2***
	(204.3)	(261.4)	(2.4)	(3.0)
(7) Propensity score blocking	198.2	399.8	7.6***	8.0**
	(202.0)	(262.6)	(2.5)	(3.2)

NOTE: Table entries are estimated average treatment effects. Except as noted, regression adjustment includes the following independent variables: age, age^2, disability, race/ethnicity, veteran status, LEP status, educational attainment, employment status at registration, exit quarter, preprogram employment and earnings, summary variables, industry of most recent employment, and labor market area. Standard errors for kernel density estimates calculated by bootstrapping (100 replications).

*** denotes significant at the 0.01 level; ** denotes significant at the 0.05 level.

Table 3 Net Impact Estimates Using Full Sample Estimation Techniques, Females

Estimator	Outcome			
	Postprogram earnings (4th qtr.)	Difference-in-Difference	Postprogram employment rate	Difference-in-Difference
(1) Difference in means (baseline)	−1141.9***	2056.7***	−0.5	17.9***
	(163.7)	(206.4)	(2.1)	(2.5)
Regression adjustment				
(2) Regression adjustment	204.5	419.8	2.1	5.0**
	(192.7)	(222.2)	(2.1)	(2.4)
(3) Regression adjustment (*p*-score as sole regressor)	399.2	486.4	6.2***	9.4***
	(223.2)	(282.5)	(2.3)	(2.6)
Kernel density estimation				
(4) Bandwidth = 0.01	253.8	736.2***	7.0***	11.8***
	(166.5)	(205.1)	(2.3)	(2.8)
(5) Bandwidth = 0.05	−144.3	1249.0***	6.5***	15.7***
	(158.8)	(188.9)	(2.1)	(2.8)
(6) Bandwidth = 0.10	−395.1	1413.4***	5.2**	16.5***
	(158.9)	(182.8)	(2.0)	(2.7)
(7) Propensity score blocking	389.2**	604.2***	8.0***	11.0***
	(186.3)	(276.3)	(2.7)	(3.0)

NOTE: Table entries are estimated average treatment effects. Except as noted, regression adjustment includes the following independent variables: age, age^2, disability, race/ethnicity, veteran status, LEP status, educational attainment, employment status at registration, exit quarter, preprogram employment and earnings, summary variables, industry of most recent employment, and labor market area. Standard errors for kernel density estimates calculated by bootstrapping (100 replications).
*** denotes significant at the 0.01 level; ** denotes significant at the 0.05 level.

Table 4 Net Impact Estimates and Match Quality Indicators Using Characteristics Matching, Males

Estimator	Outcome			
	Postprogram earnings (4th qtr.)	Difference-in-Differences	Postprogram employment rate	Difference-in-Differences
(1) Full sample, difference in means, regression-adjusted	197.9	314.7	4.3	5.5**
	(258.8)	(258.0)	(2.4)	(2.6)
Mahalanobis distance matching (with replacement)				
(2) Difference in means	−5.1	286.1	3.8	12.3***
	(256.8)	(344.2)	(3.6)	(4.2)
(3) Regression-adjustment	473.7	529.4	7.4**	12.3***
	(272.5)	(315.9)	(3.6)	(4.0)
Match quality				
(a) % comparison sample observations that are unique		96.8		
(b) Maximum repetition		4		
(c) *F*-test, all covariates (d.f.)[a]		3.60	(30, 360)	$p < 0.001$
(d) *F*-test, preregistration employment and earnings (d.f.)[a]		9.33	(6, 384)	$p < 0.001$

NOTE: Table entries are estimated average treatment effects. Except as noted, regression adjustment includes the following independent variables: age, age^2, disability, race/ethnicity, veteran status, LEP status, educational attainment, employment status at registration, exit quarter, preprogram employment and earnings, summary variables, industry of most recent employment, and labor market area. Standard errors for difference in means that are not regression-adjusted calculated by bootstrapping (100 replications).

*** denotes significant at the 0.01 level; ** denotes significant at the 0.05 level.

[a] d.f. stands for degrees of freedom.

Table 5 Net Impact Estimates and Match Quality Indicators Using Characteristics Matching, Females

Estimator	Outcome			
	Postprogram earnings (4th qtr.)	Difference-in-Differences	Postprogram employment rate	Difference-in-Differences
(1) Full sample, difference in means, regression-adjusted	204.5	419.8	2.1	5.6**
	(192.7)	(222.2)	(2.1)	(2.4)
Mahalanobis distance matching (with replacement)				
(2) Difference in means	22.4	837.5***	4.4	13.3***
	(212.1)	(243.8)	(2.7)	(3.4)
(3) Regression-adjustment	784.6***	894.5***	10.8***	17.2***
	(213.4)	(244.8)	(3.0)	(3.5)
Match quality				
(a) % comparison sample observations that are unique		90.9		
(b) Maximum repetition		13		
(c) *F*-test, all covariates (d.f.)[a]		2.84	(31, 485)	$p < 0.001$
(d) *F*-test, preregistration employment and earnings (d.f.)[a]		7.10	(6, 510)	$p < 0.001$

NOTE: Table entries are estimated average treatment effects. Regression adjustment includes the following independent variables: age, age^2, disability, race-ethnicity, veteran status, LEP status, educational attainment, employment status at registration, exit quarter, pre-program employment and earnings, summary variables, industry of most recent employment, and labor market area. Standard errors for difference in means that are not regression-adjusted calculated by bootstrapping (100 replications).

*** denotes significant at 0.01 level; ** denotes significant at 0.05 level.

[a] d.f. stands for degrees of freedom.

Table 6 Net Impact Estimates and Match Quality Indicators for *P*-score Matching, With and Without Replacement, Males

Estimator	Outcome			
	Postprogram earnings (4th qtr.)	Difference-in-Differences	Postprogram employment rate	Difference-in-Differences
(1) Full sample, difference in means, regression-adjusted	197.9	314.7	4.3	5.5**
	(258.8)	(258.0)	(2.4)	(2.6)
P-score Matching (without replacement)				
(2) Difference in means	341.9	223.0	6.1	6.4
	(254.3)	(330.3)	(3.2)	(3.8)
(3) Regression-adjustment	466.5	369.1	6.4	7.8**
	(253.3)	(309.2)	(3.4)	(3.8)
Match quality				
(a) Mean *p*-score difference		0.0031		
(b) % comparison obs. unique		100.0		
(c) Maximum repetition		1		
(d) *F*-test, all covariates (d.f.)[a]		2.14	(30, 360)	$p < 0.001$
(e) *F*-test, preregistration employment and earnings (d.f.)[a]		2.63	(6, 384)	$p = 0.016$
P-score matching (with replacement)				
(4) Difference in means	438.1	263.0	4.8	4.9
	(263.6)	(362.8)	(3.7)	(4.2)
(5) Regression-adjustment	586.4**	515.3	5.5	6.9
	(247.5)	(301.8)	(3.4)	(3.8)
Match quality				
(a) Mean *p*-score difference		0.0011		
(b) % comparison obs. unique		92.2		
(c) Maximum repetition		3		
(d) *F*-test, all covariates (d.f.)[a]		2.19	(30, 360)	$p < 0.001$
(e) *F*-test, pre-registration employment and earnings (d.f.)[a]		2.87	(6, 384)	$p = 0.010$

NOTE: Table entries are estimated average treatment effects. Regression adjustment includes the following independent variables: age, age^2, disability, race/ethnicity, veteran status, LEP status, educational attainment, employment status at registration, exit quarter, preprogram employment and earnings, summary variables, industry of most recent employment, and labor market area. Standard errors for difference in means that are not regression-adjusted calculated by bootstrapping (100 replications).

*** denotes significant at the 0.01 level; ** denotes significant at the 0.05 level.

[a] d.f. stands for degrees of freedom.

Table 7 Net Impact Estimates and Match Quality Indicators for *P*-score Matching, With and Without Replacement, Females

Estimator	Outcome: Postprogram earnings (4th qtr.)	Outcome: Difference-in-Differences	Outcome: Postprogram employment rate	Outcome: Difference-in-Differences
(1) Full sample, difference in means, regression-adjusted	204.5	419.8	2.1	5.6**
	(192.7)	(222.2)	(2.1)	(2.4)
P-score Matching (without replacement)				
(2) Difference in means	310.4	546.9**	7.4***	10.1***
	(171.1)	(241.4)	(2.2)	(3.2)
(3) Regression-adjustment	398.4	400.7	7.1**	11.3***
	(204.5)	(258.6)	(2.9)	(3.3)
Match quality				
(a) Mean *p*-score difference		0.0439		
(b) % comparison obs. unique		100.0		
(c) Maximum repetition		1		
(d) *F*-test, all covariates (d.f.)[a]		1.59	(31, 485)	$p = 0.025$
(e) *F*-test, pre-registration employment and earnings (d.f.)[a]		1.27	(6, 510)	$p = 0.271$
P-score matching (with replacement)				
(4) Difference in means	421.0**	484.5	10.1***	14.5***
	(200.7)	(237.6)	(2.6)	(3.6)
(5) Regression-adjustment	512.0**	531.3**	10.6***	15.5***
	(202.3)	(235.0)	(2.9)	(3.3)
Match quality				
(a) Mean *p*-score difference		0.0025		
(b) % comparison obs. unique		88.9		
(c) Maximum repetition		13		
(d) *F*-test, all covariates (d.f.)[a]		1.80	(31, 485)	$p = 0.006$
(e) *F*-test, preregistration employment and earnings (d.f.)[a]		1.50	(6, 510)	$p = 0.178$

NOTE: Table entries are estimated average treatment effects. Regression adjustment includes the following independent variables: age, age^2, disability, race/ethnicity, veteran status, LEP status, educational attainment, employment status at registration, exit quarter, preprogram employment and earnings, summary variables, industry of most recent employment, and labor market area. Standard errors for difference in means that are not regression-adjusted calculated by bootstrapping (100 replications).

*** denotes significant at the 0.01 level; ** denotes significant at the 0.05 level.

[a] d.f. stands for degrees of freedom.

Table 8 Net Impact Estimates and Match Quality Indicators for *P*-score Matching, With Replacement, Selecting 1, 5, and 10 Nearest Neighbors, Males

Estimator	Outcome: Postprogram earnings (4th qtr.)	Outcome: Difference-in-Differences	Outcome: Postprogram employment rate	Outcome: Difference-in-Differences
(1) Full sample, difference in means, regression-adjusted	197.9	314.7	4.3	5.5**
	(258.8)	(258.0)	(2.4)	(2.6)
P-score Matching (with replacement, 1-to-1)				
(2) Difference in means	438.1	263.0	4.8	4.9
	(263.6)	(362.8)	(3.7)	(4.2)
(3) Regression-adjustment	586.4**	515.3	5.5	6.9
	(247.5)	(301.8)	(3.4)	(3.8)
Match quality				
(a) Mean *p*-score difference		0.0011		
(b) % comparison obs. unique		92.2		
(c) Maximum repetition		3		
(d) *F*-test, all covariates (d.f.)[a]		2.19	(30, 360)	$p < 0.001$
(e) *F*-test, preregistration employment and earnings (d.f.)[a]		2.87	(6, 384)	$p = 0.010$
P-score matching (with replacement, 1-to-5)				
(4) Difference in means	271.0	207.0	6.4**	6.5**
	(223.3)	(289.3)	(2.6)	(3.4)
(5) Regression-adjustment	369.9	226.5	6.7**	8.6***
	(193.5)	(233.7)	(2.6)	(2.9)
Match quality				
(a) Mean *p*-score difference		0.0011		
(b) % comparison obs. unique		85.5		
(c) Maximum repetition		7		
(d) *F*-test, all covariates (d.f.)[a]		2.08	(30, 1528)	$p < 0.001$
(e) *F*-test, preregistration employment and earnings (d.f.)[a]		2.58	(6, 1552)	$p < 0.001$
P-score matching (with replacement, 1-to-10)				
(6) Difference in means	262.8	181.5	6.1**	6.1**
	(218.9)	(282.0)	(2.5)	(3.2)
(7) Regression-adjustment	348.0	252.2	6.7***	8.1***
	(183.8)	(217.1)	(2.4)	(2.7)
Match quality				
(a) Mean *p*-score difference		0.0034		
(b) % comparison obs. unique		81.9		
(c) Maximum repetition		11		
(d) *F*-test, all covariates (d.f.)[a]		1.96	(30, 2988)	$p = 0.001$
(e) *F*-test, preregistration employment and earnings (d.f.)[a]		2.39	(6, 3012)	$p = 0.026$

NOTE: Table entries are estimated average treatment effects. Regression adjustment includes the following independent variables: age, age^2, disability, race/ethnicity, veteran status, LEP status, educational attainment, employment status at registration, exit quarter, preprogram employment and earnings, summary variables, industry of most recent employment, and labor market area. Standard errors for difference in means that are not regression-adjusted calculated by bootstrapping (100 replications).

*** denotes significant at the 0.01 level; ** denotes significant at the 0.05 level.

[a] d.f. stands for degrees of freedom.

Table 9 Net Impact Estimates and Match Quality Indicators for *P*-score Matching, With Replacement, Selecting 1, 5, and 10 Nearest Neighbors, Females

Estimator	Outcome: Postprogram earnings (4th qtr.)	Outcome: Difference-in-Differences	Outcome: Postprogram employment rate	Outcome: Difference-in-Differences
(1) Full sample, difference in means, regression-adjusted	204.5	419.8	2.1	5.6**
	(192.7)	(222.2)	(2.1)	(2.4)
P-score Matching (with replacement, 1-to-1)				
(2) Difference in means	421.0**	484.5**	10.1***	14.5***
	(200.7)	(237.6)	(2.6)	(3.6)
(3) Regression-adjustment	512.0**	531.3**	10.6***	15.5***
	(202.3)	(235.0)	(2.9)	(3.3)
Match quality				
(a) Mean *p*-score difference		0.0025		
(b) % comparison obs. unique		88.9		
(c) Maximum repetition		13		
(d) *F*-test, all covariates (d.f.)[a]		1.80	(31, 485)	$p = 0.006$
(e) *F*-test, preregistration employment and earnings (d.f.)[a]		1.50	(6, 510)	$p = 0.178$
P-score matching (with replacement, 1-to-5)				
(4) Difference in means	421.3***	666.0***	8.2***	10.3***
	(162.4)	(213.0)	(2.2)	(2.9)
(5) Regression-adjustment	494.4***	599.4***	8.8***	11.6***
	(140.1)	(162.3)	(2.3)	(2.5)
Match quality				
(a) Mean *p*-score difference		0.0047		
(b) % comparison obs. unique		82.1		
(c) Maximum repetition		29		
(d) *F*-test, all covariates (d.f.)[a]		2.11	(31, 2049)	$p < 0.001$
(e) *F*-test, preregistration employment and earnings (d.f.)[a]		1.17	(6, 2074)	$p = 0.322$
P-score matching (with replacement, 1-to-10)				
(6) Difference in means	419.5***	701.0***	8.5***	11.2***
	(158.6)	(209.9)	(2.0)	(2.6)
(7) Regression-adjustment	501.1***	604.1***	8.8***	12.6***
	(131.1)	(152.4)	(2.2)	(2.4)
Match quality				
(a) Mean *p*-score difference		0.0081		
(b) % comparison obs. unique		75.8		
(c) Maximum repetition		44		
(d) *F*-test, all covariates (d.f.)[a]		1.98	(31, 4004)	$p = 0.001$
(e) *F*-test, preregistration employment and earnings (d.f.)[a]		1.34	(6, 4029)	$p = 0.236$

NOTE: Table entries are estimated average treatment effects. Regression adjustment includes the following independent variables: age, age^2, disability, race/ethnicity, veteran status, LEP status, educational attainment, employment status at registration, exit quarter, preprogram employment and earnings, summary variables, industry of most recent employment, and labor market area. Standard errors for difference in means that are not regression-adjusted calculated by bootstrapping (100 replications).

*** denotes significant at the 0.01 level; ** denotes significant at the 0.05 level.

[a] d.f. stands for degrees of freedom.

Table 10 Net Impact Estimates and Match Quality Indicators for *P*-score Matching, With Replacement, Calipers = 0.005 and 0.01, Males

Estimator	Outcome: Postprogram earnings (4th qtr.)	Outcome: Difference-in Differences	Outcome: Postprogram employment rate	Outcome: Difference-in-Differences
(1) Full sample, difference in means, regression-adjusted	197.9	314.7	4.3	5.5**
	(258.8)	(258.0)	(2.4)	(2.6)
P-score Matching (with replacement)				
(2) Difference in means	438.1	263.0	4.8	4.9
	(263.6)	(362.8)	(3.7)	(4.2)
(3) Regression-adjustment	586.4**	515.3	5.5	6.9
	(247.5)	(301.8)	(3.4)	(3.8)
Match quality				
(a) Mean *p*-score difference		0.0011		
(b) % comparison obs. unique		92.2		
(c) Maximum repetition		3		
(d) *F*-test, all covariates (d.f.)[a]		2.19	(30, 360)	$p < 0.001$
(e) *F*-test, preregistration employment and earnings (d.f.)[a]		2.87	(6, 384)	$p = 0.010$
P-score matching (with replacement, caliper = 0.01)				
(4) Difference in means	423.2	305.4	4.6	5.6
	(268.3)	(354.9)	(3.7)	(4.3)
(5) Regression-adjustment	601.9**	550.9	5.6	6.8
	(251.5)	(307.5)	(3.4)	(3.9)
Match quality (deleted 8 matches)				
(a) Mean *p*-score difference		0.0002		
(b) % comparison obs. unique		92.8		
(c) Maximum repetition		3		
(d) *F*-test, all covariates (d.f.)[a]		2.12	(30, 352)	$p < 0.001$
(e) *F*-test, preregistration employment and earnings (d.f.)[a]		2.88	(6, 376)	$p = 0.009$
P-score matching (with replacement, caliper = 0.005)				
(6) Difference in means	437.6	323.6	4.5	5.6
	(270.6)	(368.2)	(3.7)	(4.4)
(7) Regression-adjustment	609.8**	560.8	5.5	6.4
	(251.7)	(307.9)	(3.4)	(3.9)
Match quality (deleted 10 matches)				
(a) Mean *p*-score difference		0.0002		
(b) % comparison obs. unique		92.7		
(c) Maximum repetition		3		
(d) *F*-test, all covariates (d.f.)[a]		2.09	(30, 350)	$p < 0.001$
(e) *F*-test, preregistration employment and earnings (d.f.)[a]		2.91	(6, 374)	$p = 0.009$

NOTE: Table entries are estimated average treatment effects. Regression adjustment includes the following independent variables: age, age^2, disability, race/ethnicity, veteran status, LEP status, educational attainment, employment status at registration, exit quarter, preprogram employment and earnings, summary variables, industry of most recent employment, and labor market area. Standard errors for difference in means that are not regression-adjusted calculated by bootstrapping (100 replications).

*** denotes significant at the 0.01 level; ** denotes significant at the 0.05 level.

[a] d.f. stands for degrees of freedom.

Table 11 Net Impact Estimates and Match Quality Indicators for *P*-score Matching, With Replacement, Calipers = 0.005 and 0.01, Females

Estimator	Outcome: Postprogram earnings (4th qtr.)	Outcome: Difference-in-Differences	Outcome: Postprogram employment rate	Outcome: Difference-in-Differences
(1) Full sample, difference in means, regression-adjusted	204.5 (192.7)	419.8 (222.2)	2.1 (2.1)	5.6** (2.4)
P-score Matching (with replacement)				
(2) Difference in means	421.0** (200.7)	484.5 (237.6)	10.1*** (2.6)	14.5*** (3.6)
(3) Regression-adjustment	512.0** (202.3)	531.3** (235.0)	10.6*** (2.9)	15.5*** (3.3)
Match quality				
(a) Mean *p*-score difference		0.0025		
(b) % comparison obs. unique		88.9		
(c) Maximum repetition		13		
(d) *F*-test, all covariates (d.f.)[a]		1.80	(31, 485)	$p = 0.006$
(e) *F*-test, preregistration employment and earnings (d.f.)[a]		1.50	(6, 510)	$p = 0.178$
P-score matching (with replacement, caliper = 0.01)				
(4) Difference in means	316.8 (205.0)	436.9 (241.3)	8.0*** (2.6)	11.7*** (3.6)
(5) Regression-adjustment	348.5 (210.7)	391.1 (245.3)	7.5** (3.0)	12.7*** (3.3)
Match quality (deleted 27 matches)				
(a) Mean *p*-score difference		0.0005		
(b) % comparison obs. unique		89.8		
(c) Maximum repetition		5		
(d) *F*-test, all covariates (d.f.)[a]		1.75	(31, 458)	$p = 0.009$
(e) *F*-test, preregistration employment and earnings (d.f.)[a]		1.83	(6, 483)	$p = 0.092$
P-score matching (with replacement, caliper = 0.005)				
(6) Difference in means	281.1 (207.7)	420.6 (245.1)	7.4*** (2.7)	10.9*** (3.6)
(7) Regression-adjustment	325.9 (215.2)	364.0 (250.7)	6.5** (3.0)	11.8*** (3.4)
Match quality (deleted 37 matches)				
(a) Mean *p*-score difference		0.0003		
(b) % comparison obs. unique		91.0		
(c) Maximum repetition		4		
(d) *F*-test, all covariates (d.f.)[a]		1.76	(31, 448)	$p = 0.008$
(e) *F*-test, preregistration employment and earnings (d.f.)[a]		1.86	(6, 473)	$p = 0.085$

NOTE: Table entries are estimated average treatment effects. Regression adjustment includes the following independent variables: age, age^2, disability, race/ethnicity, veteran status, LEP status, educational attainment, employment status at registration, exit quarter, preprogram employment and earnings, summary variables, industry of most recent employment, and labor market area. Standard errors for difference in means that are not regression-adjusted calculated by bootstrapping (100 replications).

*** denotes significant at the 0.01 level; ** denotes significant at the 0.05 level.

[a] d.f. stands for degrees of freedom.

Table 12 Net Impact Estimates and Match Quality Indicators for *P*-score Caliper Matching, Without Replacement, Calipers = 0.005 and 0.01, Males

Estimator	Outcome: Postprogram earnings (4th qtr.)	Outcome: Difference-in-Differences	Outcome: Postprogram employment rate	Outcome: Difference-in-Differences
(1) Full sample, difference in means, regression-adjusted	197.9	314.7	4.3	5.5**
	(258.8)	(258.0)	(2.4)	(2.6)
P-score Matching (without replacement)				
(2) Difference in means	341.9	223.0	6.1	6.4
	(254.3)	(330.3)	(3.2)	(3.8)
(3) Regression-adjustment	466.5	369.1	6.4	7.8**
	(253.3)	(309.2)	(3.4)	(3.8)
Match quality				
(a) Mean *p*-score difference		0.0031		
(b) % comparison obs. unique		100.0		
(c) Maximum repetition		1		
(d) *F*-test, all covariates (d.f.)[a]		2.14	(30, 360)	$p < 0.001$
(e) *F*-test, preregistration employment and earnings (d.f.)[a]		2.63	(6, 384)	$p = 0.016$
P-score matching (without replacement, caliper = 0.01)				
(4) Difference in means	360.2	311.9	5.9	7.1
	(259.3)	(342.5)	(3.3)	(3.8)
(5) Regression-adjustment	502.5	425.2	6.6	7.3
	(257.7)	(316.9)	(3.4)	(3.9)
Match quality (deleted 9 matches)				
(a) Mean *p*-score difference		0.0003		
(b) % comparison obs. unique		100.0		
(c) Maximum repetition		1		
(d) *F*-test, all covariates (d.f.)[a]		2.07	(30, 351)	$p = 0.001$
(e) *F*-test, preregistration employment and earnings (d.f.)[a]		2.69	(6, 375)	$p = 0.145$
P-score matching (without replacement, caliper = 0.005)				
(6) Difference in means	305.5	237.3	5.7	6.9
	(259.4)	(342.4)	(3.3)	(3.9)
(7) Regression-adjustment	460.3	370.0	6.3	7.3
	(258.4)	(316.0)	(3.4)	(3.9)
Match quality (deleted 15 matches)				
(a) Mean *p*-score difference		0.0002		
(b) % comparison obs. unique		100.0		
(c) Maximum repetition		1		
(d) *F*-test, all covariates (d.f.)[a]		2.08	(30, 354)	$p = 0.001$
(e) *F*-test, preregistration employment and earnings (d.f.)[a]		2.85	(6, 369)	$p = 0.010$

NOTE: Table entries are estimated average treatment effects. Regression adjustment includes the following independent variables: age, age^2, disability, race/ethnicity, veteran status, LEP status, educational attainment, employment status at registration, exit quarter, preprogram employment and earnings, summary variables, industry of most recent employment, and labor market area. Standard errors for difference in means that are not regression-adjusted calculated by bootstrapping (100 replications).

*** denotes significant at the 0.01 level; ** denotes significant at the 0.05 level.

[a] d.f. stands for degrees of freedom.

Table 13 Net Impact Estimates and Match Quality Indicators for *P*-score Caliper Matching, Without Replacement, Calipers = 0.005 and 0.01, Females

Estimator	Outcome: Postprogram earnings (4th qtr.)	Outcome: Difference-in-Differences	Outcome: Postprogram employment rate	Outcome: Difference-in-Differences
(1) Full sample, difference in means, regression-adjusted	204.5	419.8	2.1	5.6**
	(192.7)	(222.2)	(2.1)	(2.4)
P-score Matching (without replacement)				
(2) Difference in means	310.4	546.9**	7.4***	10.1***
	(171.1)	(241.1)	(2.2)	(3.2)
(3) Regression-adjustment	398.4	400.7	7.1**	11.3***
	(204.5)	(238.6)	(2.9)	(3.3)
Match quality				
(a) Mean *p*-score difference		0.0439		
(b) % comparison obs. unique		100.0		
(c) Maximum repetition		1		
(d) *F*-test, all covariates (d.f.)[a]		1.59	(31, 485)	$p = 0.025$
(e) *F*-test, preregistration employment and earnings (d.f.)[a]		1.27	(6, 510)	$p = 0.271$
P-score matching (without replacement, caliper = 0.01)				
(4) Difference in means	278.1	673.4***	7.3***	11.3***
	(186.1)	(261.9)	(2.4)	(3.4)
(5) Regression-adjustment	318.5	377.2	5.8	10.6***
	(226.9)	(263.3)	(3.2)	(3.6)
Match quality (deleted 59 matches)				
(a) Mean *p*-score difference		0.0004		
(b) % comparison obs. unique		100.0		
(c) Maximum repetition		1		
(d) *F*-test, all covariates (d.f.)[a]		1.86	(31, 426)	$p = 0.004$
(e) *F*-test, preregistration employment and earnings (d.f.)[a]		1.30	(6, 451)	$p = 0.254$
P-score matching (without replacement, caliper = 0.005)				
(6) Difference in means	271.8	657.7**	7.0***	11.2***
	(182.3)	(253.7)	(2.4)	(3.3)
(7) Regression-adjustment	356.0	416.5	6.2**	10.9***
	(223.6)	(259.0)	(3.1)	(3.5)
Match quality (deleted 66 matches)				
(a) Mean *p*-score difference		0.0002		
(b) % comparison obs. unique		100.0		
(c) Maximum repetition		1		
(d) *F*-test, all covariates (d.f.)[a]		2.12	(31, 419)	$p < 0.001$
(e) *F*-test, preregistration employment and earnings (d.f.)[a]		1.31	(6, 444)	$p = 0.252$

NOTE: Table entries are estimated average treatment effects. Regression adjustment includes the following independent variables: age, age^2, disability, race/ethnicity, veteran status, LEP status, educational attainment, employment status at registration, exit quarter, preprogram employment and earnings, summary variables, industry of most recent employment, and labor market area. Standard errors for difference in means that are not regression-adjusted calculated by bootstrapping (100 replications).

*** denotes significant at the 0.01 level; ** denotes significant at the 0.05 level.

[a] d.f. stands for degrees of freedom.

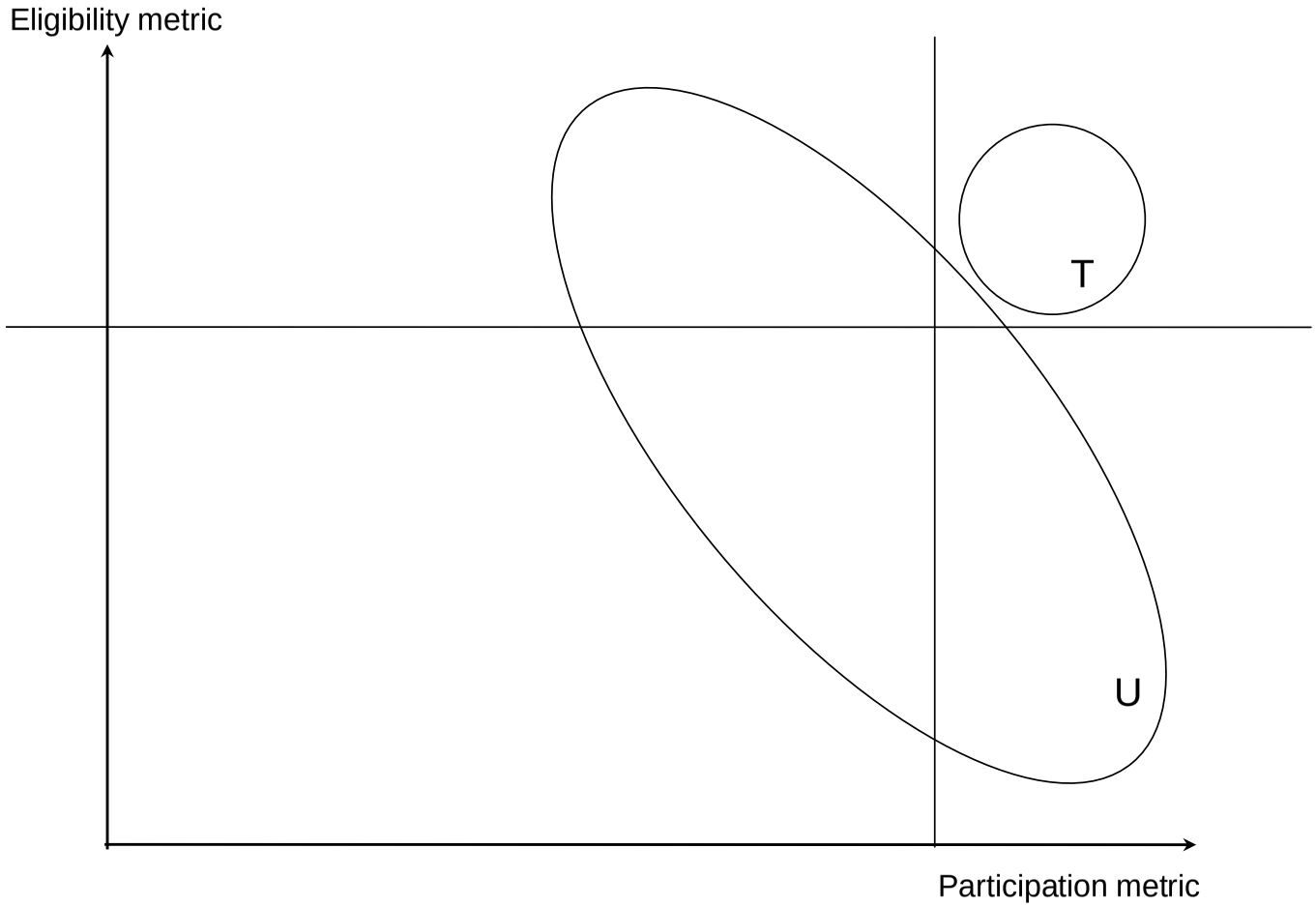
Table 14 Summary of Net Impact Estimates

Estimator	Male		Female	
	Earnings (D-in-D)	Employment (D-in-D)	Earnings (D-in-D)	Employment (D-in-D)
(1) Full sample, difference in means, unadjusted	1818.6***	14.5***	2056.7***	17.9***
	(253.1)	(3.1)	(206.4)	(2.5)
(2) Full sample, regression adjusted	314.7	5.5**	419.8	5.0**
	(258.0)	(2.6)	(222.2)	(2.4)
(3) Full sample, regression adjusted (*p*-score only)	166.5	8.4***	486.4	9.4***
	(386.9)	(2.8)	(282.5)	(2.6)
(4) Full sample, kernel density, bandwidth = 0.01	552.6**	8.7***	736.2***	11.8***
	(269.9)	(3.1)	(205.1)	(2.8)
(5) *p*-score blocking	399.8	8.0**	604.2***	11.0***
	(262.6)	(3.2)	(226.3)	(3.0)
(6) Characteristics matching (Mahalanobis metric), regression adjusted	529.4	12.3***	894.5***	17.2***
	(315.9)	(4.0)	(244.8)	(3.5)
(7) *p*-score matching, w/o replacement, regression adjusted	369.1	7.8**	400.7	11.3***
	(309.2)	(3.8)	(258.6)	(3.3)
(8) *p*-score matching, w/o replacement, 0.01 caliper, regression adjusted	370.0	7.3	416.5	10.9***
	(316.0)	(3.9)	(259.0)	(3.5)
(9) *p*-score matching, w/replacement, regression adjusted	515.3	6.9	531.3**	15.5***
	(301.8)	(3.8)	(235.0)	(3.3)
(10) *p*-score matching, w/replacement, 1-to-5, regression adjusted	226.5	8.6***	592.4***	11.6***
	(233.7)	(2.9)	(162.3)	(2.5)
(11) *p*-score matching, w/replacement, 0.01 caliper, regression adjusted	550.9	6.8	391.1	12.7***
	(307.5)	(3.9)	(245.3)	(3.3)
(12) Treatment sample mean levels at time of program registration	2908.7	73.1	2008.9	74.4

NOTE: Table entries are estimated average treatment effects. Regression adjustment includes the following independent variables: age, age^2, disability, race/ethnicity, veteran status, LEP status, educational attainment, employment status at registration, exit quarter, preprogram employment and earnings, summary variables, industry of most recent employment, and labor market area. Standard errors for row (4) calculated by bootstrapping (100 replications). D-in-D means difference-in-differences.

*** denotes significant at the 0.01 level; ** denotes significant at the 0.05 level.

Figure 1 Treatment Sample and Full Sample from which Matched Comparison Sample may be Drawn



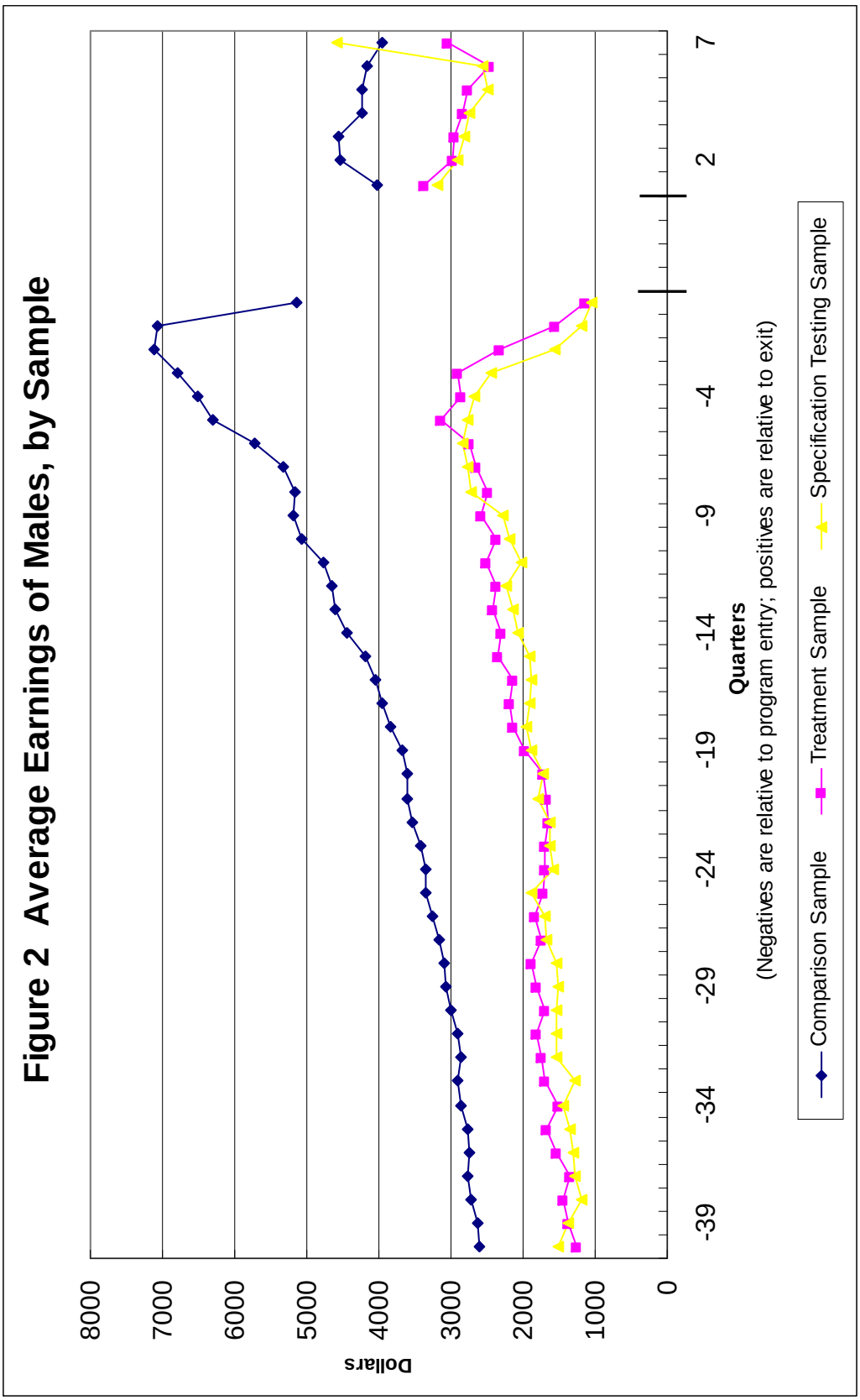


Figure 2 Average Earnings of Males, by Sample
Dollars
8000
7000
6000
5000
4000
3000
2000
1000
0
-39
-34
-29
-24
-19
-14
-9
-4
2
7
Quarters
(Negatives are relative to program entry; positives are relative to exit)
Comparison Sample
Treatment Sample
Specification Testing Sample

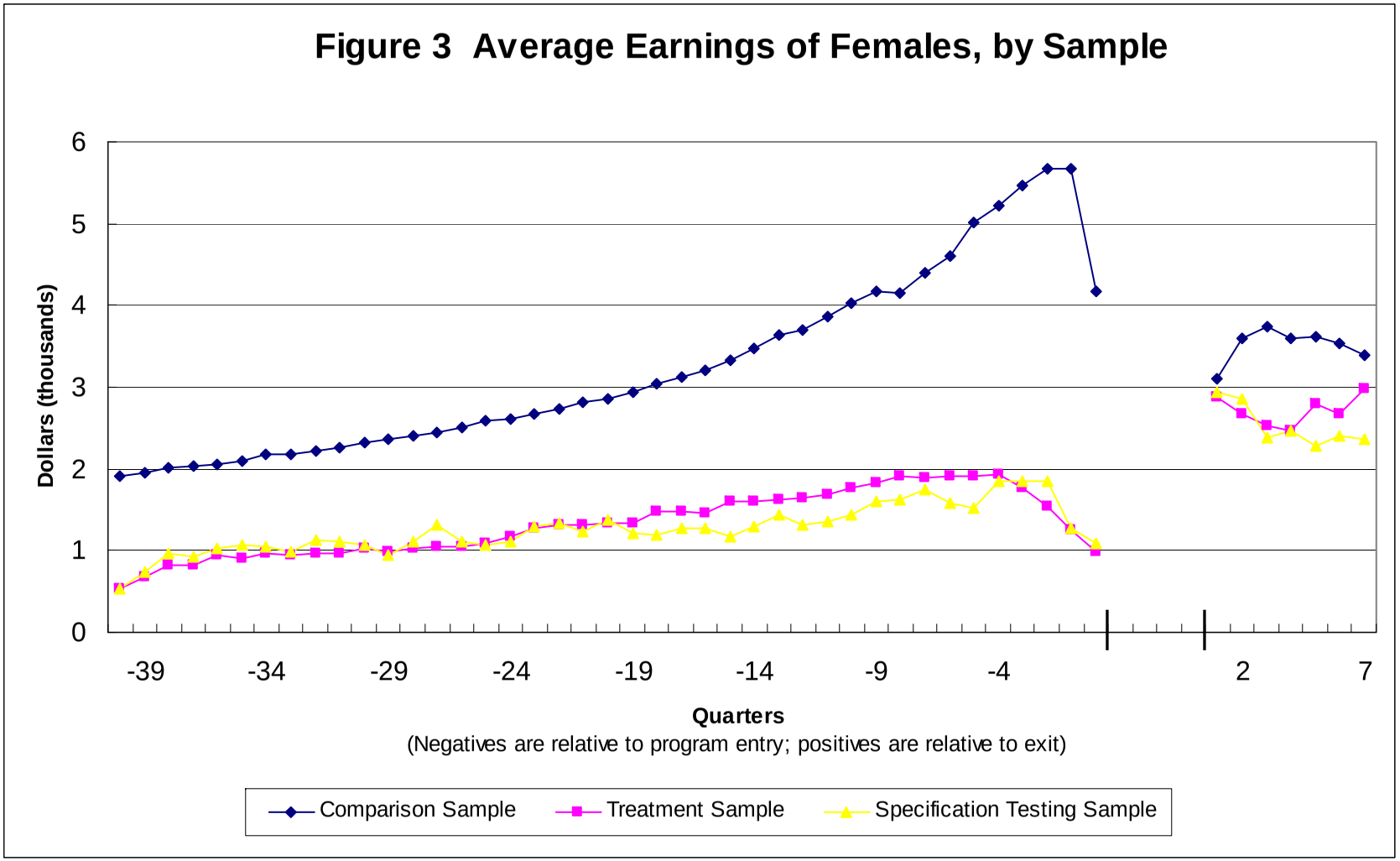


Figure 3 Average Earnings of Females, by Sample
Dollars (thousands)
6
5
4
3
2
1
0
-39
-34
-29
-24
-19
-14
-9
-4
2
7
Quarters
(Negatives are relative to program entry; positives are relative to exit)
Comparison Sample
Treatment Sample
Specification Testing Sample

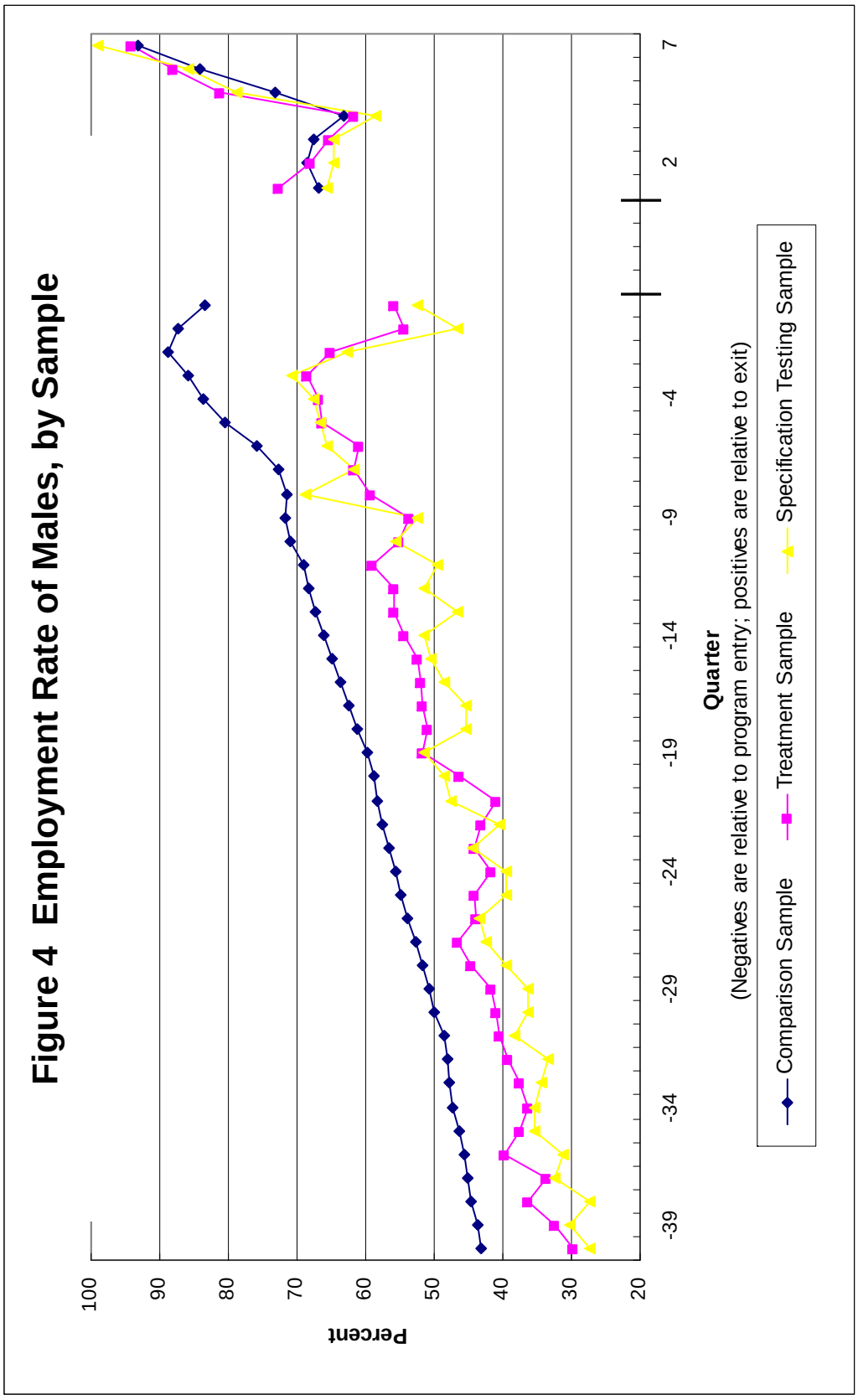
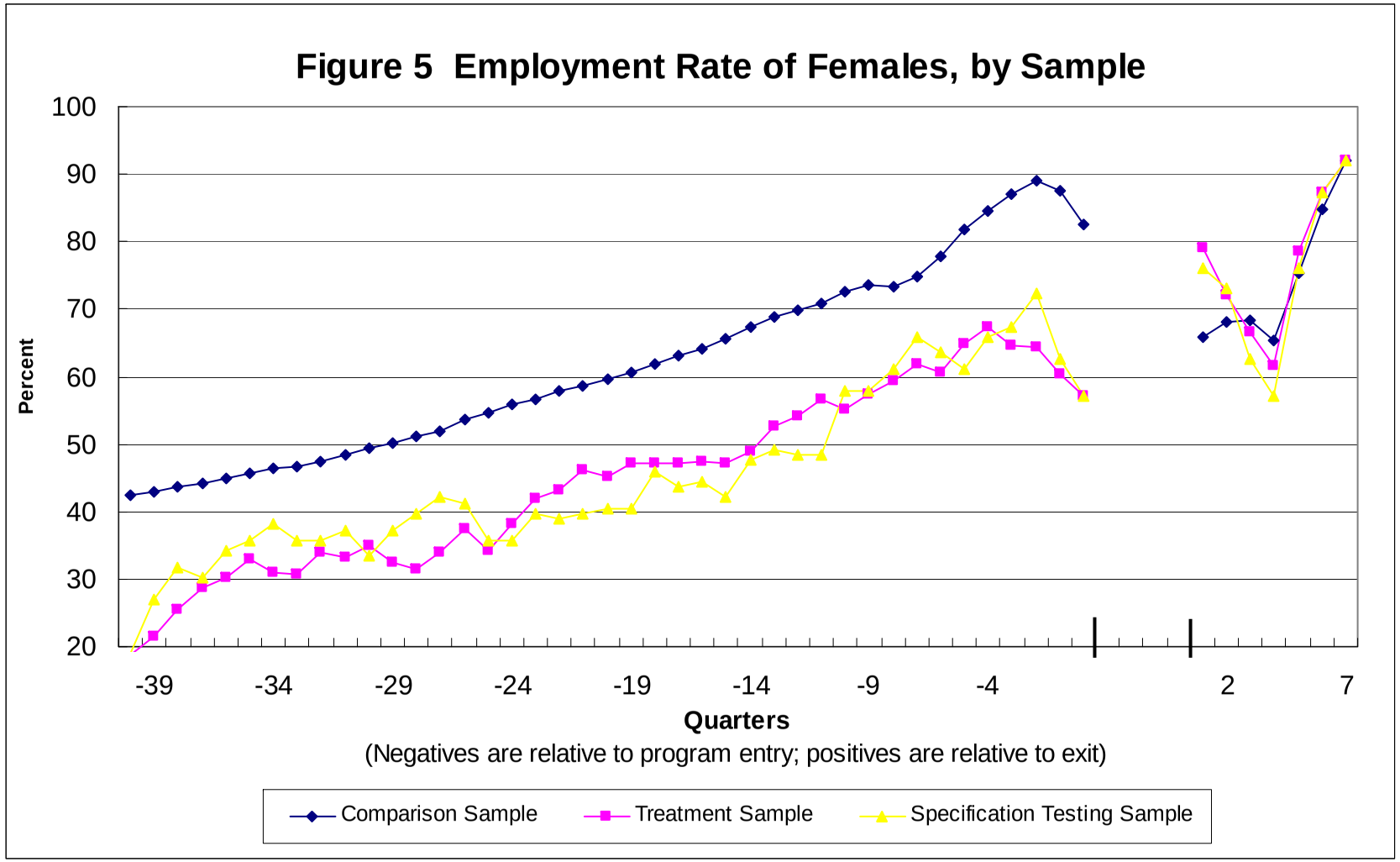


Figure 4 Employment Rate of Males, by Sample
Percent
100
90
80
70
60
50
40
30
20
-39
-34
-29
-24
-19
-14
-9
-4
2
7
Quarter
(Negatives are relative to program entry; positives are relative to exit)
Comparison Sample
Treatment Sample
Specification Testing Sample



Note: Employment is defined as quarterly earnings exceeding $100.