

Hollenbeck, Kevin

Working Paper

Using administrative data for workforce development program evaluation

Upjohn Institute Working Paper, No. 04-103

Provided in Cooperation with:

W. E. Upjohn Institute for Employment Research, Kalamazoo, Mich.

Suggested Citation: Hollenbeck, Kevin (2004) : Using administrative data for workforce development program evaluation, Upjohn Institute Working Paper, No. 04-103, W.E. Upjohn Institute for Employment Research, Kalamazoo, MI, <https://doi.org/10.17848/wp04-103>

This Version is available at:

<https://hdl.handle.net/10419/64396>

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

**USING ADMINISTRATIVE DATA FOR
WORKFORCE DEVELOPMENT PROGRAM EVALUATION**

Upjohn Institute Staff Working Paper No. 04-103

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

August 2004

JEL Classification Codes: J24, J38, H43

Presented at the 2004 National Workforce Investment Research Colloquium, sponsored by the U.S. Department of Labor, Employment and Training Administration, Arlington, VA: May 24, 2004. Thanks to Jeff Smith, Randall Eberts, Chris O’Leary, and Timothy Bartik for review of earlier versions of the paper. Special thanks to my discussants—David Stevens and Carolyn Heinrich—for their excellent comments and suggestions at the Colloquium. Finally, I gratefully acknowledge Wei-Jang Huang and Claire Black for 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.

Abstract

This paper addresses the question of whether administrative data that are collected for performance monitoring purposes can be used for program evaluation. It argues that under certain circumstances, such data can be used. In particular, data from the state of Washington are used to examine the effectiveness of services provided to adults under the Workforce Investment Act (WIA). The general theme of an emerging literature on techniques for nonexperimental evaluations of social programs is that many different techniques have appropriate asymptotic properties. A contribution of this paper is to examine the sensitivity of net impact estimators to various estimation techniques. Virtually all of the techniques yielded estimates of positive labor market impacts for both men and women. Men had earnings gains on the order of 10 percent that resulted mainly from increased employment rates. Women had larger estimated earnings gains—on the order of 20 to 25 percent—that emanated from increased employment *and* increased wages or hours. A second purpose of the paper was to provide principles that policymakers and program administrators should apply when considering evaluation results.

The purpose of this paper is to address the question of whether performance monitoring data can be used for program evaluation purposes. It argues that under certain circumstances, such data can be used. In particular, the program performance data that are routinely gathered and monitored by administrators of many workforce development programs meet these circumstances. The chapter 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, which are referenced throughout this chapter 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 (i.e., large sample) properties. Some papers in this literature go on to speculate about which estimators seem to work best under which conditions.

Because there is no 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, some of which are quite complex, and it summarizes the net impact estimators that were generated. For the most part, the results were fairly stable across the techniques, which adds a degree of confidence to them. The final section offers guidance to policymakers and program administrators, who may not be familiar with the technical details of

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

² The approach in this chapter is very similar to work described in Mueser, Troske, and Gorislawsky (2003).

various analytical approaches about how empirical results that may appear to be complex or unstable can be used for program improvement.

INTRODUCTION

Analyses of quantitative data about workforce development programs are valuable to at least two audiences: individuals charged with administering the programs and entities that invest resources into the programs. 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. Investors (or funders) want to make sure that they are maximizing their return 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 return on investment in order to allocate or reallocate their resources. I use the term *net impact evaluation* to refer to this kind of information. The question that this chapter addresses is whether performance monitoring data can be used for net impact evaluation.

The empirical results presented in the chapter 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 chapter 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 or universities, apprenticeships, adult basic education, formal or informal on the job training, or secondary career and technical education.

PERFORMANCE MONITORING VERSUS PROGRAM EVALUATION

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 perspective 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 has become available, which I argue may be used for evaluation purposes as well.

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 program features for which they 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

program's effectiveness will need to make judgements about trends or levels in the data. Benchmarks, which are summaries of comparable indicators from other establishments or other time periods, can be used to formulate those judgements. 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.

Program Evaluation

Evaluation is intended to go beyond monitoring; its purpose is to *assess* program performance. Stufflebeam (1999) suggests that its purpose is to make judgments about worth and value. In particular, evaluation draws conclusions about whether programs are achieving their purposes or objectives. Obviously, this means that evaluators need to identify the purposes or objectives of the program, which may or may not be straightforward. In the world of workforce development programs, for example, there is sometimes tension between employment and skill development (training) goals. Moreover, once the goals have been decided, evaluation studies must find outcomes that are measurable and indicative of the outcomes. Finally, perhaps the most difficult aspect of an evaluation is the establishment of attribution i.e., determining the extent to which outcomes result from programmatic interventions. The following subsections briefly discuss two aspects of a program evaluation: process evaluation and net impact evaluation.

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

Process evaluation

An important factor in determining whether a program is achieving its goals or objectives is its operating practices. A workforce development program may be attempting to deliver a certain outcome, such as employment, but its operating practices may be impeding that goal. Perhaps individuals who were poorly matched to a job were referred to employers and, consequently, the program has lost credibility in the employer community. Perhaps the program has overpromised results for individuals who participate in particular training programs and has therefore lost credibility with trainees.

It is the function of a process analysis to observe closely program operations and attempt to identify the components of the program that are working and why, and conversely, the components of the program that are not working and why not. As with performance monitoring, the main audience for a process evaluation is program administrators. Usually the information that is collected is qualitative in nature (open-ended interviews or focus groups). Two main results occur from a process evaluation. First, program administrators are presented with recommendations about components that might be changed and about components that are working well and should not be changed. Second, a process evaluation will generate hypotheses that inform a net impact evaluation. For example, are there particular support services for clients whose accessibility seems to be highly correlated with successful outcomes? Are there particular idiosyncrasies about how program components are offered that might partially explain successful or unsuccessful outcomes?⁴

⁴ Bloom, Hill, and Riccio (2001) use process evaluation data to relate earnings outcomes to process variables such as staff caseloads and program emphases on employment versus training.

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 explains how the program has changed the lives of individuals who participated in it relative to their next best alternative. The data used to address this question are quantitative, and the evaluation should attempt to disaggregate the results because there may be systematic relationships between program outcomes and participant characteristics. The audiences for a net impact evaluation are the funding agency(ies) and 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 were given the same treatment, they may exert more or less effort in learning or applying the skills or knowledge being delivered to them. Furthermore, some individuals may not complete the treatment. 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. Having 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 into 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 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 might be for program improvement purposes.
- 2) Administrative data must be available for a group of individuals that arguably make a reasonable source of cases for a comparison group.
- 3) Outcome data must be available for both the treatment and comparison groups.
- 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.

THE NET IMPACT EVALUATION PROBLEM AND KEY ASSUMPTIONS

This section will present the problem in mathematical terms, but basically the desired information (which cannot be observed) is the difference between the outcome that occurs to an individual if they participate in the program minus the outcome that would occur if the individual did not participate. Obviously, individuals cannot simultaneously be in two states of the world, so we must *estimate* the net impacts.

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 , and 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:

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

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 $g(\cdot)$ 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 .⁵

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:

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

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

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

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

⁵ Note that Imbens (2004) shows that this condition can be slightly weakened to $\text{Pr}(W_i = 1|X_i) < 1$.

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

What is observed is the following:

$$(4) \quad Y_i = (1 - W_i)Y_i(0) + W_iY_i(1)$$

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

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

where $\hat{f}_k(X)$, $k = 1, 0$, are the outcome means for the treatment and comparison group samples, respectively, and

BIAS represents the expected difference in the $Y(0)$ outcome between the comparison group (actually observed) and the treatment group (the counterfactual.) The BIAS term may be called selection bias.

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 random whether or not the individual is a participant. This is equivalent to the assumption that participation in the treatment can be explained by X up to a random error term. 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

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

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 (or for very large samples), in practice, it may make sense to regression adjust equation (6).

Regression and Quasi-experimental Estimation of Net Impacts

Another paper in this conference addresses the use of random assignment experiments to estimate the net impacts of programs (Burtless and Greenberg 2004). Clearly, a well-conducted experiment is the “best” solution to the attribution problem because it designs in the assumption of “unconfoundedness.” 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.

For purposes of this paper, I assume that experimental data are unavailable. Instead, I assume that I have a 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 do 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 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 being in T or in U is between 0 and 1. I will call techniques that use all of U , full sample techniques.⁶ 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.⁷ This estimator essentially assumes away selection bias. It may be represented as follows:

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

⁶ Some of these techniques trim or delete observations from U , but I will still refer to them as full sample techniques.

⁷ In comments on this paper, David Stevens pointed out that its emphasis is on the traditional focus of net impact mean value estimates. David encouraged readers to not neglect analysis of outliers, an evaluation focus that has been around for decades. Stevens cited Klitgaard and Hall (1973, 1975).

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 :

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

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 “nontreated” outcome for all observations in both T and U .⁸ The following average treatment effect can then be calculated:

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

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

Equation (8) and the more general regression in the first stage of (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 such that the observations closest to the treatment observations receive the highest weights. This estimator may be written as follows (following Imbens 2004):

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

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

⁸ 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.

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

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) 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 ($= \text{Prob}(W = 1|X)$).

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

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

$$(9') \quad \tau = \frac{1}{N} \sum_{i \in T, U} [\hat{f}_1(p(X_i)) - \hat{f}_0(p(X_i))].$$

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

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 full sample into “blocks” or strata by propensity score, such 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 k th block be defined as all treatment or full comparison sample cases with values of X such that $p(X) \in [p_{1k}, p_{2k}]$. Let NT_k be the number of

treatment cases in the k th block and NU_k be the number of comparison cases from the full sample. The treatment effect with each block k is as follows:

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

and the overall estimated average treatment effect is given as follows:

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

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 comprised of the observations where individuals are most “like” the individuals comprising 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.⁹

There is a substantial and growing literature on how to sample individuals to construct the comparison sample.¹⁰ 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

⁹ This forces the analyst to use bootstrapping techniques to calculate standard errors.

¹⁰ See Heckman, LaLonde, and Smith (1999) and references cited there.

is that the cross-tabulation of data, if there are many common variables, may result in small or empty cells.¹¹

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 j th observation in T and k th observation in U , then C will be comprised 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 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_j 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.¹²

In his work on training program evaluation, Ashenfelter (1978) demonstrated that participants' pre-program 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.

¹¹ Lyndon B. Johnson School of Public Affairs (1994) used a variation of this approach.

¹² 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 implemented in Hollenbeck, King, and Schroeder (2003.)

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 score. Treatment observations are matched to 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 1-to-1 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 1-to-5 and 1-to-10 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.

EMPIRICAL ESTIMATION OF THE NET IMPACT OF WIA SERVICES

Data

The “treatment” in this section of the chapter is receipt of WIA intensive or training services by adults¹³ who exited from WIA in program year 2000 (July 2000–June 2001) in the state of Washington.¹⁴ The counterfactual that I am using to construct a comparison group is that if there were no WIA services, then individuals would receive services through the state Employment Service (Wagner-Peyser services).¹⁵ Thus the pool of observations from which we construct the comparison groups is comprised of individuals whose last reported service date in the Employment Service 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:Q2.¹⁶ The data sets used here are among the rich longitudinal data sets being used for analyses in nine states currently participating in the Administrative DATA Research and Evaluation (ADARE) Project: California, Florida, Georgia, Illinois, Maryland, Missouri, Ohio, Texas, and Washington.

The empirical analyses are intended to be illustrative in order to demonstrate the stability of the net impact estimates to various full sample or matched sample estimators. 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

¹³ Note that I am only looking at individuals served in the “adult” program, not dislocated workers or youth.

¹⁴ Note that I also estimate net program impacts for individuals who received just training services in the sensitivity analyses presented.

¹⁵ In her discussion of this paper, Carolyn Heinrich pointed out that an implicit assumption in this empirical work is that the Employment Service 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.

¹⁶ Note that in much of the analysis described in this chapter, I refer to pre-registration employment and earnings data. To construct these variables, I used wage record data starting in 1997:Q3 only. 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.

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 pre-program characteristics and outcomes. Between 2–3 percent of the comparison sample are disabled, compared to over one-fifth 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 male and one-fourth of the females that received training or intensive services from WIA were employed at 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 over \$5,000 for females. The WIA exiters' pre-program 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 pre-program characteristics. Earnings, as measured by the average quarterly earnings in the 4th quarter after leaving the program and as measured by the average quarterly earnings 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 pre-program 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 they do. Figures 2 through 5 display the data for key

outcome variables. The first two figures show quarterly earnings for male and females, respectively. Clearly the comparison sample earnings are much higher than the treatment sample. 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.

Full Sample Estimators of Net Impact

Tables 2 and 3 provide estimates of the net impact of the treatment (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 pre-program 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 shown in Table 1, the employment and earnings advantages for the comparison group outcomes were not nearly as large as the pre-program differences, so the difference-in-differences are quite large and positive.¹⁷

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 time of registration with WIA or the ES. They are as follows: age, race/ethnicity, educational attainment, veteran status, disability status,

¹⁷ All of the earnings impacts in this paper are denominated in constant 2000 dollars.

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) summarizes the employment and earnings histories of individuals using the following five variables: 1) percent of quarters employed since entering employment, 2) conditional average earnings (pre-program), 3) trend in earnings levels (constant \$), 4) variance in earnings levels, and 5) turnover. In this paper, I use these variables plus a measure of pre-program dip in earnings that may have occurred in the pre-program earnings history.¹⁸

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—propensity score and 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 Gorislawsky (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 lot of difference. With the exception of the post-program 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 used in this approach uses the full comparison sample, in principle, although we do trim some observations to guarantee full overlap. In

¹⁸ The earnings dip variable is defined as $\max [\$0, (\text{average quarterly earnings in pre-registration quarters } -3 \text{ to } -8 \text{ minus average quarterly earnings in pre-registration quarters } -1 \text{ to } -2)]$.

particular, observations are eliminated from U if their p -score $< \min (p\text{-score})$ for T and observations are eliminated from T if their p -score $> \max (p\text{-score})$ for U . We then “blocked” the file into p -score deciles, and performed an F -test to determine if 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 7th row of the table are weighted averages of the cell-by-cell treatment effects, where 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.

Matched Sample Estimators

Several different matched sample estimators were calculated. All of the approaches estimated a treatment effect by computing the average difference in outcomes for the treatment sample and the matched sample, and 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 chapter use a propensity score approach. This approach uses 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 quasi-experimental, matched sample estimation technique, it is important to try to demonstrate the “quality” of the match. Several indicators are used in this chapter. First, 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 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 fewer cases 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 $0 < \text{prob}(\text{participation} | X) < 1$) 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 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 used in this chapter are very close to 80 percent—80.9 percent for males and 83.5 percent for females. Figures 6 and 7 display the distributions of p -scores for males and females, respectively. Note that these distributions are limited to $p\text{-score} \geq 0.02$ because of the “spike” of p -scores in the comparison sample with values near 0.

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 1-to-1 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 the tables 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, I would not choose this form of matched file estimation to be my 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) *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 1-to-1, 1-to-5, and 1-to-10 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 with 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 with males.

The match quality statistics conform to expectations. Choosing more observations to match causes the mean of the p -score differences to increase. The means for the estimators using 1-to-10 are three times as great as the mean differences for the 1-to-1 estimators. Furthermore, considerably fewer comparison file observations are used uniquely in the techniques that are 1-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 pre-registration 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 radii) matching delete 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 differences 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 dramatically dropped; 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 a caliper 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 increase generally with the more binding caliper of 0.005. The average difference in p -scores tumbles by almost 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 did not vary greatly from the overall average treatment effects.

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 from WIA in its first full year of implementation, i.e., program year 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 pre-program 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. With the mean of average quarterly earnings prior to the program being approximately \$2,900, 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 that appear to mainly results from these services' impact on employment.

All of the earnings impacts for females are also positive, and of larger magnitude than the estimates for males. Consequently many of them are statistically significant at the 0.05 level.

The magnitudes range from \$391 to \$894, which 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 program year 2000 had an impact on the earnings of adult females of approximately 20 to 25 percent that result from these services' impact on employment and either wages or hours or both.

Net Impacts of Training, Separated from Intensive Services

The design of WIA calls for sequenced services for clients. All clients are eligible to receive core services. Clients who do not readily become employed after receiving core services may receive intensive services, and those who do not become employed with the intensive services may receive training services, to the extent that resources allow. The analyses in this chapter used individuals who exited from WIA who had received either intensive or training services (presumably in addition to intensive services) as the treatment group. WIA clients who received only core services were not in the analysis at all, and the comparison sample was comprised of individuals who received Wagner-Peyser services. (Any records of individuals in the comparison sample who had also received WIA services were deleted from the analyses.)

Another set of results of interest to program administrators might be the efficacy of WIA training services only. Tables 15 and 16 provide estimates of the treatment effect contrasting the case where the treatment is intensive or training with the case where the treatment is training. The top panels in each table repeat some of the prior estimates using the former treatment. As

noted, the comparison sample is the ES file. The second panel displays the estimated results using just training as the treatment, and again using the ES file as the comparison sample. The bottom panel uses training as the treatment sample, but uses individuals from the WIA program who received core or intensive services as the comparison sample.

Comparing the first two panels in both the tables shows that the estimated net impacts are quite comparable. The magnitudes of the estimates increase slightly in the second panel for both males and females, which suggests that the positive impacts for intensive or training services are slightly larger for training than for intensive services. But basically, the results seem to be quite similar. That is also true for the bottom panel, when the comparison sample is limited to WIA clients who did not receive training services.

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 chapter, I will 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 will 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 will also assume that the studies being considered have gone through a peer review process, and have achieved a level of professional adequacy. If this last assumption does not hold, the policymaker should be extremely cautious about relying on any findings.

I believe 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. 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.) 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 turns 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 have been given a credible study that has convincing evidence. However, even in this case, the decisionmaker 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 such as many of the estimates presented in this chapter, 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 equal, the validity of the estimates is likely to be increasing in 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 across 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 probably border on reasonableness and should be considered with healthy skepticism.

Principle 5: Insist on 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, the notion of stability has to be judged relative to the perturbation that has been introduced in order to compute different estimates. For example, some of the estimates in this chapter used entirely different estimation techniques and samples (for example, regression-adjusted full sample differences in means versus regression-adjusted matched sample differences in means when matching is done with replacement and selecting the 10 nearest neighbors for each treatment observation). In other cases we made minor changes, such as a trying a caliper of 0.005 instead of 0.01. Other things 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 they are stable when 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.

SUMMARY

As government resources have become scarce, more and more emphasis has been placed on accountability and demonstrated return on investment. This trend, as well as the dramatic decreases in the cost of information processing, has led to striking advances in the availability of program administrative data *and* the 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. So, a natural conclusion is that the U.S. Department of Labor should continue to support such studies within its portfolio of research and evaluation approaches.

This paper has described a number of full sample and matched sample techniques for estimating net impacts of workforce development programs. It further provided 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 appeared 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.

The U.S. Department of Labor, along with partnering states and universities, is investing resources into the ADARE consortium in order to access and link administrative data sources. This study has shown that such an investment can have a payoff by providing data that can be used for evaluating program effectiveness. A key feature of the data that the ADARE consortium has constructed is the availability of a reasonable comparison sample, which is the Employment Service (Wagner-Peyser) administrative data. Linking these data to wage record data is beneficial because of the availability of covariates to use as estimation controls and employment and earnings outcome variables from which to estimate net impacts.

A final suggestion is that these techniques might work just as well for addressing the efficacy of the nation's unemployment insurance system. Given its magnitude and importance, it would behoove the U.S. Department of Labor to investigate the feasibility of applying to that system the approaches that have been described here.

REFERENCES

- Ashenfelter, Orley. 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. Battelle Memorial Institute, Seattle, WA.
- 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.
- Bloom, Howard S., Carolyn J. Hill, and James Riccio. 2001. "Modeling the Performance of Welfare-to-Work Programs: The Effects of Program Management and Services, Economic Environment, and Client Characteristics." Working paper. New York: Manpower Demonstration Research Corporation.
- Burtless, Gary, and David Greenberg. 2004. "Evaluating Workforce Programs Using Experimental Methods." Paper presented at 2004 National Workforce Investment Colloquium held in Arlington, VA, May 24.
- Dehejia, Rajeev H., and Sadek Wahba. 1995. "Causal Effects in Non-Experimental Studies: Re-Evaluating the Evaluation of Training Programs." Working paper. Cambridge, MA: Harvard University.
- . 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 Smith. 2002. "The Performance of Performance Standards." *The Journal of Human Resources* 37(4): 778–811.
- Heckman, James J., Robert J. LaLonde, and Jeffery A. Smith. 1999. "The Economics and Econometrics of Active Labor Market Programs." In *Handbook of Labor Economics, Vol 3*, Orley Ashenfelter and David Card, eds. Amsterdam: Elsevier, pp. 1866–2097.
- Heckman, James, Hidehiko Ichimura, Jeffery Smith, and Petra Todd. 1998. "Characterizing Selection Bias Using Experimental Data." *Econometrica* 66(5): 1017–1098.
- Hollenbeck, Kevin, and Wei-Jang Huang. 2003. *Net Impact and Cost-Benefit Evaluation of Washington State's Workforce Training System: Final Report*. W.E. Upjohn Institute technical report no. TR03-018. Kalamazoo, MI: W.E. Upjohn Institute for Employment Research.

- Hollenbeck, Kevin, 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's 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." *The Review of Economics and Statistics* 86(1): 4–29.
- King, Christopher T. et al. 1994. *Texas JOBS Program Evaluation: Final Report*, Austin: Center for the Study of Human Resources, Lyndon B. Johnson School of Public Affairs, The University of Texas at Austin.
- Klitgaard, Robert E., and George R. Hall. 1973. *A Statistical Search for Unusually Effective Schools*, RAND Document 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?" *The Review of Economics and Statistics* 86(1): 156–179.
- Mohr, Lawrence B. 1992. *Impact Analysis for Program Evaluation*. Newbury Park, California: Sage.
- Mueser, Peter, Kenneth R. Troske, and Alexey Gorislavsky. 2003. "Using State Administrative Data to Measure Program Performance." Unpublished manuscript. University of Missouri, Department of Economics, Columbia, MO.
- 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*. Newbury Park, CA: Sage Publications.
- Roy, A.D. 1952. "Some Thoughts on the Distribution of Earnings." *Oxford Economic Papers* 3: 135–146.
- Stufflebeam, Daniel L. 1999. *Evaluation Models*. San Francisco: Jossey-Bass.
- 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.” *The 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
LEP	6.1	8.2	5.9	2.4	7.2	4.8
<u>Education completed</u>						
< 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**
<u>Pre-program employment</u>						
Employment rate (%)	73.2	73.1	87.7**	74.7	74.4	88.5**
Avg. earnings (\$)	2,609.1	2,908.7	6,398.1**	1,860.2	2,008.9	5,059.5**
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	1,670.2	1,388.3	671.5**	608.6**	973.1	523.6
<u>Outcomes</u>						
Earnings in quarter 4	2,746.5	2,844.0	4,235.1**	2,474.6	2,460.1	3,602.0**
Ave. earnings	4,122.7	4,176.5	6,299.3**	3,713.9	3,593.0	5,099.3**
Employment rate (%)	62.6	65.1	66.4**	64.3	66.8	67.3
Difference in earnings	-653.3	-1143.7	-2,964.3**	175.8	-26.7	-2,083.4**
Difference in avg. earnings	435.2	230.7	-920.7**	1,247.3	946.2	-596.4**
Difference in employment 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%) and a specification testing subsample (25%). Comparison samples are a random 50% 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			
	Post-program earnings (4 th qtr.)	Difference-in-differences	Post-program employment rate	Difference-in-differences
(1) Difference in means (baseline)	-1,391.2*** (194.0)	1,818.6*** (253.1)	-1.3 (2.4)	14.5*** (3.1)
<u>Regression adjustment</u>				
(2) Regression adjustment	197.9 (258.8)	314.7 (258.0)	4.3 (2.4)	5.5** (2.6)
(3) Regression adjustment (<i>p</i> -score as sole regressor)	302.8 (288.0)	166.5 (386.9)	7.1*** (2.5)	8.4*** (2.8)
<u>Kernel density estimation</u>				
(4) Bandwidth = 0.01	-31.3 (205.3)	552.6** (269.9)	6.1** (2.5)	8.7*** (3.1)
(5) Bandwidth = 0.05	-701.4*** (199.3)	1,131.0*** (264.6)	2.4 (2.4)	9.8*** (3.0)
(6) Bandwidth = 0.10	-883.6*** (204.3)	1,342.7*** (261.4)	1.9 (2.4)	11.2*** (3.0)
(7) <u>Propensity score blocking</u>	198.2 (202.0)	399.8 (262.6)	7.6*** (2.5)	8.0** (3.2)

NOTE: Table entries are estimated average treatment effects. Except as noted, regression adjustment includes the following independent variables: age, age², 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 kernel density estimates calculated by bootstrapping (100 replications).

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

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

Estimator	Outcome			
	Post-program earnings (4th qtr.)	Difference-in-differences	Post-program employment rate	Difference-in-differences
(1) Difference in means (baseline)	-1,141.9*** (163.7)	2,056.7*** (206.4)	-0.5 (2.1)	17.9*** (2.5)
<u>Regression adjustment</u>				
(2) Regression adjustment	204.5 (192.7)	419.8 (222.2)	2.1 (2.1)	5.0** (2.4)
(3) Regression adjustment (<i>p</i> -score as sole regressor)	399.2 (223.2)	486.4 (282.5)	6.2*** (2.3)	9.4*** (2.6)
<u>Kernel density estimation</u>				
(4) Bandwidth = 0.01	253.8 (166.5)	736.2*** (205.1)	7.0*** (2.3)	11.8*** (2.8)
(5) Bandwidth = 0.05	-144.3 (158.8)	1,249.0*** (188.9)	6.5*** (2.1)	15.7*** (2.8)
(6) Bandwidth = 0.10	-395.1 (158.9)	1,413.4*** (182.8)	5.2** (2.0)	16.5*** (2.7)
(7) <u>Propensity score blocking</u>	389.2** (186.3)	604.2*** (276.3)	8.0*** (2.7)	11.0*** (3.0)

NOTE: Table entries are estimated average treatment effects. Except as noted, regression adjustment includes the following independent variables: age, age², 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 kernel density estimates calculated by bootstrapping (100 replications).

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

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

Estimator	Outcome			
	Post-program earnings (4th qtr.)	Difference-in-differences	Post-program employment rate	Difference-in-differences
(1) Full sample, difference in means, regression-adjusted	197.9 (258.8)	314.7 (258.0)	4.3 (2.4)	5.5** (2.6)
<u>Mahalanobis distance matching (with replacement)</u>				
(2) Difference in means	-5.1 (256.8)	286.1 (344.2)	3.8 (3.6)	12.3*** (4.2)
(3) Regression-adjustment	473.7 (272.5)	529.4 (315.9)	7.4** (3.6)	12.3*** (4.0)
<u>Match quality</u>				
(a) Percent of comparison sample obs. that are unique		96.8		
(b) Maximum repetition		4		
(c) <i>F</i> -test, all covariates (d.f.)		3.60	(30, 360)	$p < 0.001$
(d) <i>F</i> -test, pre-registration employment and earnings (d.f.)		11.14	(6, 360)	$p < 0.001$

NOTE: Table entries are estimated average treatment effects. Except as noted, regression adjustment includes the following independent variables: age, age², 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.

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

Estimator	Outcome			
	Post-program earnings (4th qtr.)	Difference-in-differences	Post-program employment rate	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)
<u>Mahalanobis distance matching (with replacement)</u>				
(2) Difference in means	22.4 (212.1)	837.5*** (243.8)	4.4 (2.7)	13.3*** (3.4)
(3) Regression-adjustment	784.6*** (213.4)	894.5*** (244.8)	10.8*** (3.0)	17.2*** (3.5)
<u>Match quality</u>				
(a) Percent of comparison sample obs. that are unique		90.9		
(b) Maximum repetition		13		
(c) <i>F</i> -test, all covariates (d.f.)		2.84	(31, 485)	$p < 0.001$
(d) <i>F</i> -test, pre-registration employment and earnings (d.f.)		7.99	(6, 485)	$p < 0.001$

NOTE: Table entries are estimated average treatment effects. Regression adjustment includes the following independent variables: age, age², 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.

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

Estimator	Outcome			
	Post-program earnings (4th qtr.)	Difference-in-differences	Post-program employment rate	Difference-in-differences
(1) Full sample, difference in means, regression-adjusted	197.9 (258.8)	314.7 (258.0)	4.3 (2.4)	5.5** (2.6)
<u><i>P</i>-score matching (without replacement)</u>				
(2) Difference in means	341.9 (254.3)	223.0 (330.3)	6.1 (3.2)	6.4 (3.8)
(3) Regression-adjustment	466.5 (253.3)	369.1 (309.2)	6.4 (3.4)	7.8** (3.8)
<u>Match quality</u>				
(a) Mean <i>p</i> -score difference		0.0031		
(b) Percent comparison obs. unique		100.0		
(c) Maximum repetition		1		
(d) <i>F</i> -test, all covariates (d.f.)		2.14	(30, 360)	$p < 0.001$
(e) <i>F</i> -test, pre-registration employment and earnings (d.f.)		2.23	(6, 360)	$p = 0.040$
<u><i>P</i>-score matching (with replacement)</u>				
(4) Difference in means	438.1 (263.6)	263.0 (362.8)	4.8 (3.7)	4.9 (4.2)
(5) Regression-adjustment	586.4** (247.5)	515.3 (301.8)	5.5 (3.4)	6.9 (3.8)
<u>Match quality</u>				
(a) Mean <i>p</i> -score difference		0.0011		
(b) Percent comparison obs. unique		92.2		
(c) Maximum repetition		3		
(d) <i>F</i> -test, all covariates (d.f.)		2.19	(30, 360)	$p < 0.001$
(e) <i>F</i> -test, pre-registration employment and earnings (d.f.)		2.58	(6, 360)	$p = 0.018$

NOTE: Table entries are estimated average treatment effects. Regression adjustment includes the following independent variables: age, age², 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 0.05 level.

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

Estimator	Outcome			
	Post-program earnings (4th qtr.)	Difference-in-differences	Post-program employment rate	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)
<u><i>P</i>-score matching (without replacement)</u>				
(2) Difference in means	310.4 (171.1)	546.9** (241.4)	7.4*** (2.2)	10.1*** (3.2)
(3) Regression-adjustment	398.4 (204.5)	400.7 (258.6)	7.1** (2.9)	11.3*** (3.3)
<u>Match quality</u>				
(a) Mean <i>p</i> -score difference		0.0439		
(b) Percent comparison obs. unique		100.0		
(c) Maximum repetition		1		
(d) <i>F</i> -test, all covariates (d.f.)		1.59	(31, 485)	<i>p</i> = 0.025
(e) <i>F</i> -test, pre-registration employment and earnings (d.f.)		0.97	(6, 485)	<i>p</i> = 0.446
<u><i>P</i>-score matching (with replacement)</u>				
(4) Difference in means	421.0** (200.7)	484.5 (237.6)	10.1*** (2.6)	14.5*** (3.6)
(5) Regression-adjustment	512.0** (202.3)	531.3** (235.0)	10.6*** (2.9)	15.5*** (3.3)
<u>Match quality</u>				
(a) Mean <i>p</i> -score difference		0.0025		
(b) Percent comparison obs. unique		88.9		
(c) Maximum repetition		13		
(d) <i>F</i> -test, all covariates (d.f.)		1.80	(31, 485)	<i>p</i> = 0.006
(e) <i>F</i> -test, pre-registration employment and earnings (d.f.)		1.57	(6, 485)	<i>p</i> = 0.154

NOTE: Table entries are estimated average treatment effects. Regression adjustment includes the following independent variables: age, age², 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.

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			
	Post-program earnings (4th qtr.)	Difference-in-differences	Post-program employment rate	Difference-in-differences
(1) Full sample, difference in means, regression-adjusted	197.9 (258.8)	314.7 (258.0)	4.3 (2.4)	5.5** (2.6)
<u><i>P</i>-score matching (with replacement, 1-to-1)</u>				
(2) Difference in means	438.1 (263.6)	263.0 (362.8)	4.8 (3.7)	4.9 (4.2)
(3) Regression-adjustment	586.4** (247.5)	515.3 (301.8)	5.5 (3.4)	6.9 (3.8)
<u>Match quality</u>				
(a) Mean <i>p</i> -score difference		0.0011		
(b) Percent comparison obs. unique		92.2		
(c) Maximum repetition		3		
(d) <i>F</i> -test, all covariates (d.f.)		2.19	(30, 360)	<i>p</i> < 0.001
(e) <i>F</i> -test, pre-registration employment and earnings (d.f.)		2.58	(6, 360)	<i>p</i> = 0.018
<u><i>P</i>-score matching (with replacement, 1-to-5)</u>				
(4) Difference in means	271.0 (223.3)	207.0 (289.3)	6.4** (2.6)	6.5** (3.4)
(5) Regression-adjustment	369.9 (193.5)	226.5 (233.7)	6.7** (2.6)	8.6*** (2.9)
<u>Match quality</u>				
(a) Mean <i>p</i> -score difference		0.0011		
(b) Percent comparison obs. unique		85.5		
(c) Maximum repetition		7		
(d) <i>F</i> -test, all covariates (d.f.)		2.08	(30, 1528)	<i>p</i> < 0.001
(e) <i>F</i> -test, pre-registration employment and earnings (d.f.)		2.25	(6, 1528)	<i>p</i> = 0.036
<u><i>P</i>-score matching (with replacement, 1-to-10)</u>				
(6) Difference in means	262.8 (218.9)	181.5 (282.0)	6.1** (2.5)	6.1** (3.2)
(7) Regression-adjustment	348.0 (183.8)	252.2 (217.1)	6.7*** (2.4)	8.1*** (2.7)
<u>Match quality</u>				
(a) Mean <i>p</i> -score difference		0.0034		
(b) Percent comparison obs. unique		81.9		
(c) Maximum repetition		11		
(d) <i>F</i> -test, all covariates (d.f.)		1.96	(30, 2988)	<i>p</i> = 0.001
(e) <i>F</i> -test, pre-registration employment and earnings (d.f.)		2.34	(6, 2988)	<i>p</i> = 0.029

NOTE: Table entries are estimated average treatment effects. Regression adjustment includes the following independent variables: age, age², 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.

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			
	Post-program earnings (4 th qtr.)	Difference-in-differences	Post-program employment rate	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)
<u><i>P</i>-score matching (with replacement, 1-to-1)</u>				
(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)
<u>Match quality</u>				
(a) Mean <i>p</i> -score difference		0.0025		
(b) Percent comparison obs. unique		88.9		
(c) Maximum repetition		13		
(d) <i>F</i> -test, all covariates (d.f.)		1.80	(31, 485)	<i>p</i> = 0.006
(e) <i>F</i> -test, pre-registration employment and earnings (d.f.)		1.57	(6, 485)	<i>p</i> = 0.154
<u><i>P</i>-score matching (with replacement, 1-to-5)</u>				
(4) Difference in means	421.3*** (162.4)	666.0*** (213.0)	8.2*** (2.2)	10.3*** (2.9)
(5) Regression-adjustment	494.4*** (140.1)	599.4*** (162.3)	8.8*** (2.3)	11.6*** (2.5)
<u>Match quality</u>				
(a) Mean <i>p</i> -score difference		0.0047		
(b) Percent comparison obs. unique		82.1		
(c) Maximum repetition		29		
(d) <i>F</i> -test, all covariates (d.f.)		2.11	(31, 2049)	<i>p</i> < 0.001
(e) <i>F</i> -test, pre-registration employment and earnings (d.f.)		0.82	(6, 2049)	<i>p</i> = 0.552
<u><i>P</i>-score matching (with replacement, 1-to-10)</u>				
(6) Difference in means	419.5*** (158.6)	701.0*** (209.9)	8.5*** (2.0)	11.2*** (2.6)
(7) Regression-adjustment	501.1*** (131.1)	604.1*** (152.4)	8.8*** (2.2)	12.6*** (2.4)
<u>Match quality</u>				
(a) Mean <i>p</i> -score difference		0.0081		
(b) Percent comparison obs. unique		75.8		
(c) Maximum repetition		44		
(d) <i>F</i> -test, all covariates (d.f.)		1.98	(31, 4004)	<i>p</i> = 0.001
(e) <i>F</i> -test, pre-registration employment and earnings (d.f.)		0.94	(6, 4004)	<i>p</i> = 0.467

NOTE: Table entries are estimated average treatment effects. Regression adjustment includes the following independent variables: age, age², 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.

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

Estimator	Outcome			
	Post-program earnings (4th qtr.)	Difference-in differences	Post-program employment rate	Difference-in differences
(1) Full sample, difference in means, regression-adjusted	197.9 (258.8)	314.7 (258.0)	4.3 (2.4)	5.5** (2.6)
<u><i>P</i>-score matching (with replacement)</u>				
(2) Difference in means	438.1 (263.6)	263.0 (362.8)	4.8 (3.7)	4.9 (4.2)
(3) Regression-adjustment	586.4** (247.5)	515.3 (301.8)	5.5 (3.4)	6.9 (3.8)
<u>Match quality</u>				
(a) Mean <i>p</i> -score difference		0.0011		
(b) Percent comparison obs. unique		92.2		
(c) Maximum repetition		3		
(d) <i>F</i> -test, all covariates (d.f.)		2.19	(30, 360)	<i>p</i> < 0.001
(e) <i>F</i> -test, pre-registration employment and earnings (d.f.)		2.58	(6, 360)	<i>p</i> = 0.018
<u><i>P</i>-score matching (with replacement, caliper = 0.01)</u>				
(4) Difference in means	423.2 (268.3)	305.4 (354.9)	4.6 (3.7)	5.6 (4.3)
(5) Regression-adjustment	601.9** (251.5)	550.9 (307.5)	5.6 (3.4)	6.8 (3.9)
<u>Match quality (deleted 8 matches)</u>				
(a) Mean <i>p</i> -score difference		0.0002		
(b) Percent comparison obs. unique		92.8		
(c) Maximum repetition		3		
(d) <i>F</i> -test, all covariates (d.f.)		2.12	(30, 352)	<i>p</i> = 0.001
(e) <i>F</i> -test, pre-registration employment and earnings (d.f.)		2.67	(6, 352)	<i>p</i> = 0.015
<u><i>P</i>-score matching (with replacement, caliper = 0.005)</u>				
(6) Difference in means	437.6 (270.6)	323.6 (368.2)	4.5 (3.7)	5.6 (4.4)
(7) Regression-adjustment	609.8** (251.7)	560.8 (307.9)	5.5 (3.4)	6.4 (3.9)
<u>Match quality (deleted 10 matches)</u>				
(a) Mean <i>p</i> -score difference		0.0002		
(b) Percent comparison obs. unique		92.7		
(c) Maximum repetition		3		
(d) <i>F</i> -test, all covariates (d.f.)		2.09	(30, 350)	<i>p</i> < 0.001
(e) <i>F</i> -test, pre-registration employment and earnings (d.f.)		2.65	(6, 350)	<i>p</i> = 0.016

NOTE: Table entries are estimated average treatment effects. Regression adjustment includes the following independent variables: age, age², 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 0.05 level.

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

Estimator	Outcome			
	Post-program earnings (4th qtr.)	Difference-in-differences	Post-program employment rate	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)
<u><i>P</i>-score matching (with replacement)</u>				
(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)
<u>Match quality</u>				
(a) Mean <i>p</i> -score difference		0.0025		
(b) Percent comparison obs. unique		88.9		
(c) Maximum repetition		13		
(d) <i>F</i> -test, all covariates (d.f.)		1.80	(31, 485)	<i>p</i> = 0.006
(e) <i>F</i> -test, pre-registration employment and earnings (d.f.)		1.57	(6, 485)	<i>p</i> = 0.154
<u><i>P</i>-score matching (with replacement, caliper = 0.01)</u>				
(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)
<u>Match quality (deleted 27 matches)</u>				
(a) Mean <i>p</i> -score difference		0.0005		
(b) Percent comparison obs. unique		89.8		
(c) Maximum repetition		5		
(d) <i>F</i> -test, all covariates (d.f.)		1.75	(31, 458)	<i>p</i> = 0.009
(e) <i>F</i> -test, pre-registration employment and earnings (d.f.)		1.51	(6, 458)	<i>p</i> = 0.174
<u><i>P</i>-score matching (with replacement, caliper = 0.005)</u>				
(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)
<u>Match quality (deleted 37 matches)</u>				
(a) Mean <i>p</i> -score difference		0.0003		
(b) Percent comparison obs. unique		91.0		
(c) Maximum repetition		4		
(d) <i>F</i> -test, all covariates (d.f.)		1.76	(31, 448)	<i>p</i> = 0.008
(e) <i>F</i> -test, pre-registration employment and earnings (d.f.)		1.44	(6, 448)	<i>p</i> = 0.197

NOTE: Table entries are estimated average treatment effects. Regression adjustment includes the following independent variables: age, age², 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.

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			
	Post-program earnings (4th qtr.)	Difference-in-differences	Post-program employment rate	Difference-in-differences
(1) Full sample, difference in means, regression-adjusted	197.9 (258.8)	314.7 (258.0)	4.3 (2.4)	5.5** (2.6)
<u><i>P</i>-score matching (without replacement)</u>				
(2) Difference in means	341.9 (254.3)	223.0 (330.3)	6.1 (3.2)	6.4 (3.8)
(3) Regression-adjustment	466.5 (253.3)	369.1 (309.2)	6.4 (3.4)	7.8** (3.8)
<u>Match quality</u>				
(a) Mean <i>p</i> -score difference		0.0031		
(b) Percent comparison obs. unique		100.0		
(c) Maximum repetition		1		
(d) <i>F</i> -test, all covariates (d.f.)		2.14	(30, 360)	<i>p</i> < 0.001
(e) <i>F</i> -test, pre-registration employment and earnings (d.f.)		2.23	(6, 360)	<i>p</i> = 0.040
<u><i>P</i>-score matching (without replacement, caliper = 0.01)</u>				
(4) Difference in means	360.2 (259.3)	311.9 (342.5)	5.9 (3.3)	7.1 (3.8)
(5) Regression-adjustment	502.5 (257.7)	425.2 (316.9)	6.6 (3.4)	7.3 (3.9)
<u>Match quality (deleted 9 matches)</u>				
(a) Mean <i>p</i> -score difference		0.0003		
(b) Percent comparison obs. unique		100.0		
(c) Maximum repetition		1		
(d) <i>F</i> -test, all covariates (d.f.)		2.09	(30, 351)	<i>p</i> = 0.001
(e) <i>F</i> -test, pre-registration employment and earnings (d.f.)		2.30	(6, 351)	<i>p</i> = 0.034
<u><i>P</i>-score matching (without replacement, caliper = 0.005)</u>				
(6) Difference in means	305.5 (259.4)	237.3 (342.4)	5.7 (3.3)	6.9 (3.9)
(7) Regression-adjustment	460.3 (258.4)	370.0 (316.0)	6.3 (3.4)	7.3 (3.9)
<u>Match quality (deleted 15 matches)</u>				
(a) Mean <i>p</i> -score difference		0.0002		
(b) Percent comparison obs. unique		100.0		
(c) Maximum repetition		1		
(d) <i>F</i> -test, all covariates (d.f.)		2.08	(30, 354)	<i>p</i> = 0.001
(e) <i>F</i> -test, pre-registration employment and earnings (d.f.)		2.44	(6, 354)	<i>p</i> = 0.026

NOTE: Table entries are estimated average treatment effects. Regression adjustment includes the following independent variables: age, age², 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 0.05 level.

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			
	Post-program earnings (4th qtr.)	Difference-in-differences	Post-program employment rate	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)
<u><i>P</i>-score matching (without replacement)</u>				
(2) Difference in means	310.4 (171.1)	546.9** (241.1)	7.4*** (2.2)	10.1*** (3.2)
(3) Regression-adjustment	398.4 (204.5)	400.7 (238.6)	7.1** (2.9)	11.3*** (3.3)
<u>Match quality</u>				
(a) Mean <i>p</i> -score difference		0.0439		
(b) Percent comparison obs. unique		100.0		
(c) Maximum repetition		1		
(d) <i>F</i> -test, all covariates (d.f.)		1.59	(31, 485)	<i>p</i> = 0.025
(e) <i>F</i> -test, pre-registration employment and earnings (d.f.)		0.97	(6, 485)	<i>p</i> = 0.446
<u><i>P</i>-score matching (without replacement, caliper = 0.01)</u>				
(4) Difference in means	278.1 (186.1)	673.4*** (261.9)	7.3*** (2.4)	11.3*** (3.4)
(5) Regression-adjustment	318.5 (226.9)	377.2 (263.3)	5.8 (3.2)	10.6*** (3.6)
<u>Match quality (deleted 59 matches)</u>				
(a) Mean <i>p</i> -score difference		0.0004		
(b) Percent comparison obs. unique		100.0		
(c) Maximum repetition		1		
(d) <i>F</i> -test, all covariates (d.f.)		1.86	(31, 426)	<i>p</i> = 0.004
(e) <i>F</i> -test, pre-registration employment and earnings (d.f.)		0.56	(6, 426)	<i>p</i> = 0.760
<u><i>P</i>-score matching (without replacement, caliper = 0.005)</u>				
(6) Difference in means	271.8 (182.3)	657.7** (253.7)	7.0*** (2.4)	11.2*** (3.3)
(7) Regression-adjustment	356.0 (223.6)	416.5 (259.0)	6.2** (3.1)	10.9*** (3.5)
<u>Match quality (deleted 66 matches)</u>				
(a) Mean <i>p</i> -score difference		0.0002		
(b) Percent comparison obs. unique		100.0		
(c) Maximum repetition		1		
(d) <i>F</i> -test, all covariates (d.f.)		2.12	(31, 419)	<i>p</i> < 0.001
(e) <i>F</i> -test, pre-registration employment and earnings (d.f.)		0.51	(6, 419)	<i>p</i> = 0.797

NOTE: Table entries are estimated average treatment effects. Regression adjustment includes the following independent variables: age, age², 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.

Table 14
Summary of Net Impact Estimates

Estimator	Male		Female	
	Earnings (difference-in- differences)	Employment (difference-in- differences)	Earnings (difference-in- differences)	Employment (difference-in- differences)
(1) Full sample, difference in means, unadjusted	1818.6*** (253.1)	14.5*** (3.1)	2,056.7*** (206.4)	17.9*** (2.5)
(2) Full sample, regression adjusted	314.7 (258.0)	5.5** (2.6)	419.8 (222.2)	5.0** (2.4)
(3) Full sample, regression adjusted (<i>p</i> -score only)	166.5 (386.9)	8.4*** (2.8)	486.4 (282.5)	9.4*** (2.6)
(4) Full sample, kernel density, bandwidth = 0.01	552.6** (269.9)	8.7*** (3.1)	736.2*** (205.1)	11.8*** (2.8)
(5) <i>p</i> -score blocking	399.8 (262.6)	8.0** (3.2)	604.2*** (226.3)	11.0*** (3.0)
(6) Characteristics matching (Mahalanobis metric), regression adjusted	529.4 (315.9)	12.3*** (4.0)	894.5*** (244.8)	17.2*** (3.5)
(7) <i>p</i> -score matching, w/o replacement, regression adjusted	369.1 (309.2)	7.8** (3.8)	400.7 (258.6)	11.3*** (3.3)
(8) <i>p</i> -score matching, w/o replacement, 0.01 caliper, regression adjusted	370.0 (316.0)	7.3 (3.9)	416.5 (259.0)	10.9*** (3.5)
(9) <i>p</i> -score matching, w/replacement, regression adjusted	515.3 (301.8)	6.9 (3.8)	531.3** (235.0)	15.5*** (3.3)
(10) <i>p</i> -score matching, w/replacement, 1-to-5, regression adjusted	226.5 (233.7)	8.6*** (2.9)	592.4*** (162.3)	11.6*** (2.5)
(11) <i>p</i> -score matching, w/replacement, 0.01 caliper, regression adjusted	550.9 (307.5)	6.8 (3.9)	391.1 (245.3)	12.7*** (3.3)
(12) Treatment sample mean levels at time of program registration	2,908.7	73.1	2,008.9	74.4

NOTE: Table entries are estimated average treatment effects. Regression adjustment includes the following independent variables: age, age², 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 row (4) calculated by bootstrapping (100 replications).

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

Table 15
Net Impact Estimates Contrasting Treatments and Comparison Samples, Males

Treatment/Comparison sample	Outcome			
	Post-program earnings (4th qtr.)	Difference-in-differences	Post-program employment rate	Difference-in-differences
Treatment = Intensive/Training Comparison sample = ES				
(1) Full sample, simple difference in means	-1,391.2*** (194.0)	1,818.6*** (253.1)	-1.3 (2.4)	14.5*** (3.1)
(2) Full sample, regression adjusted	197.9 (258.8)	314.7 (258.6)	4.3 (2.4)	5.5** (2.6)
(3) Matched sample, <i>p</i> -score, w/replacement	438.1 (267.6)	263.0 (362.8)	4.8 (3.7)	4.9 (4.2)
(4) Matched sample, <i>p</i> -score, w/replacement, regression adjusted	586.4** (247.5)	515.3 (301.8)	5.5 (3.4)	6.9 (3.8)
Treatment = Training Comparison group = ES				
(5) Full sample, simple difference in means	-1,331.8*** (197.4)	2,125.3*** (246.6)	-0.7 (2.5)	14.4*** (3.2)
(6) Full sample, regression adjusted	251.1 (267.4)	583.8 (313.7)	4.2 (2.5)	4.9 (2.7)
(7) Matched sample, <i>p</i> -score, w/replacement	228.5 (314.3)	333.6 (387.5)	9.6** (4.0)	10.1** (4.6)
(8) Matched sample, <i>p</i> -score, w/replacement, regression adjusted	275.1 (273.6)	233.8 (307.5)	8.7** (3.5)	10.3** (4.0)
Treatment = Training Comparison group = Core/Intensive				
(9) Full sample, simple difference in means	343.1 (337.3)	545.6 (450.8)	3.7 (4.2)	1.4 (5.1)
(10) Full sample, regression adjusted	642.1 (357.3)	707.9 (431.4)	3.7 (4.8)	-0.4 (5.4)

NOTE: Table entries are estimated average treatment effects. Regression adjustment includes the following independent variables: age, age², 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.

Table 16
Net Impact Estimates Contrasting Treatments and Comparison Samples, Females

Treatment/Comparison sample	Outcome			
	Post-program earnings (4th qtr.)	Difference-in-differences	Post-program employment rate	Difference-in-differences
Treatment = Intensive/Training Comparison sample = ES				
(1) Full sample, simple difference in means	-1,141.9*** (163.7)	2,056.7*** (206.4)	-0.5 (2.1)	17.9** (2.5)
(2) Full sample, regression adjusted	204.5 (192.7)	419.8 (222.2)	2.1 (2.1)	5.6** (2.4)
(3) Matched sample, <i>p</i> -score, w/replacement	421.0** (200.7)	484.5** (237.6)	10.1*** (2.6)	14.5*** (3.6)
(4) Matched sample, <i>p</i> -score, w/replacement, regression adjusted	512.0** (202.3)	531.3** (235.0)	10.6*** (2.9)	15.5*** (3.3)
Treatment = Training Comparison group = ES				
(5) Full sample, simple difference in means	-1,006.8*** (146.0)	2,268.6*** (188.6)	2.0 (2.1)	20.8*** (2.6)
(6) Full sample, regression adjusted	231.4 (201.8)	383.0 (231.2)	2.9 (2.3)	6.4*** (2.5)
(7) Matched sample, <i>p</i> -score, w/replacement	757.1*** (189.4)	578.1** (247.0)	14.7*** (2.9)	13.4*** (3.8)
(8) Matched sample, <i>p</i> -score, w/replacement, regression adjusted	710.4*** (188.4)	692.8*** (222.7)	13.3*** (2.9)	14.0*** (3.4)
Treatment = Training Comparison group = Core/Intensive				
(9) Full sample, simple difference in means	570.4** (287.0)	557.1 (341.2)	9.7*** (3.7)	11.1*** (4.5)
(10) Full sample, regression adjusted	383.8 (282.5)	319.1 (322.2)	7.8** (3.8)	9.8** (4.5)

NOTE: Table entries are estimated average treatment effects. Regression adjustment includes the following independent variables: age, age², 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.

Figure 1 Treatment Sample and Full Sample from Which Matched Comparison Sample May Be Drawn

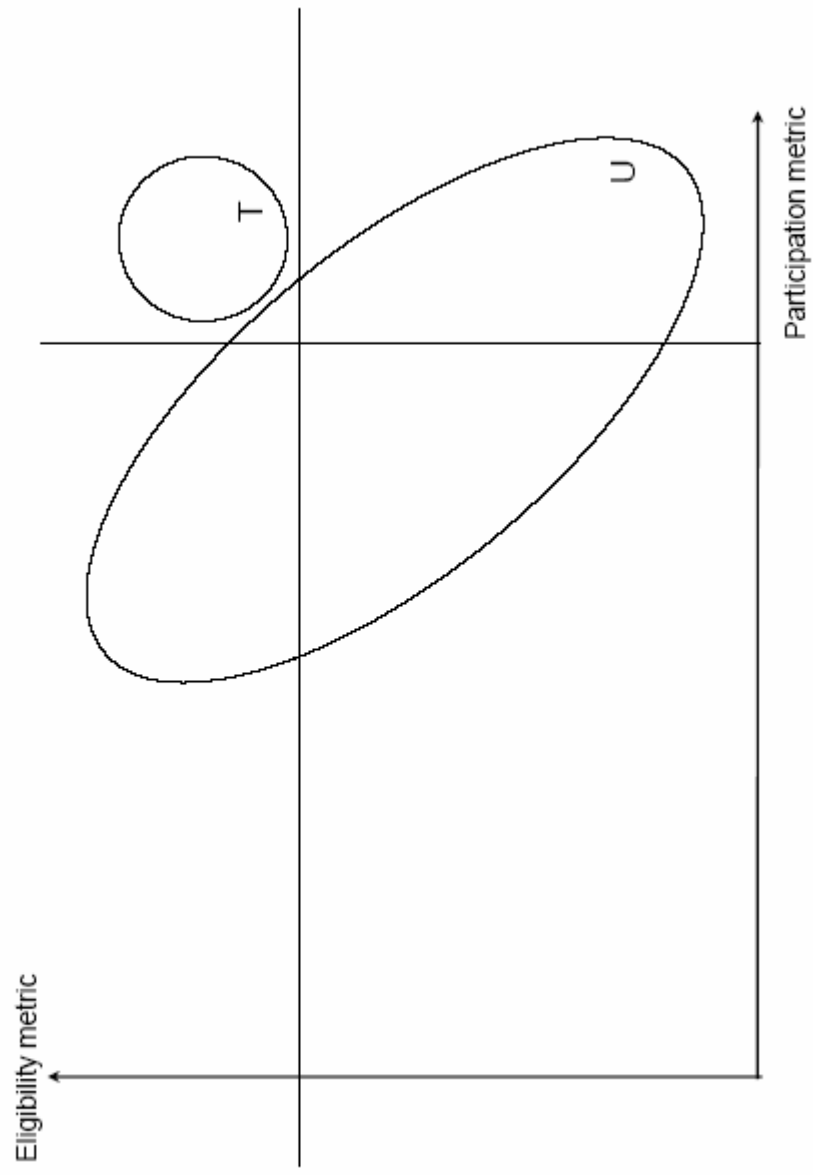


Figure 2 **Average Earnings, Male**
Comparison Sample: Full

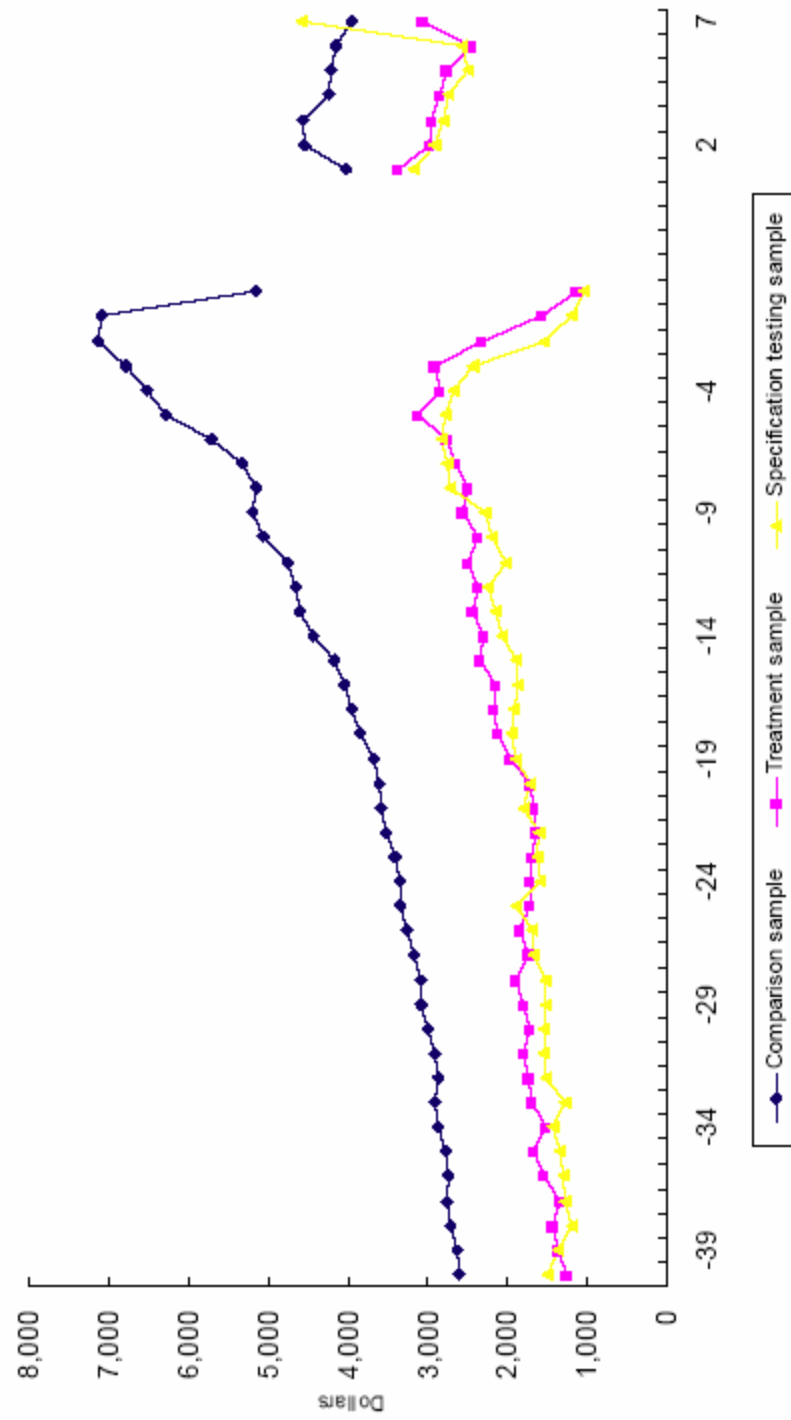


Figure 3 **Average Earnings, Female**
Comparison Sample: Full

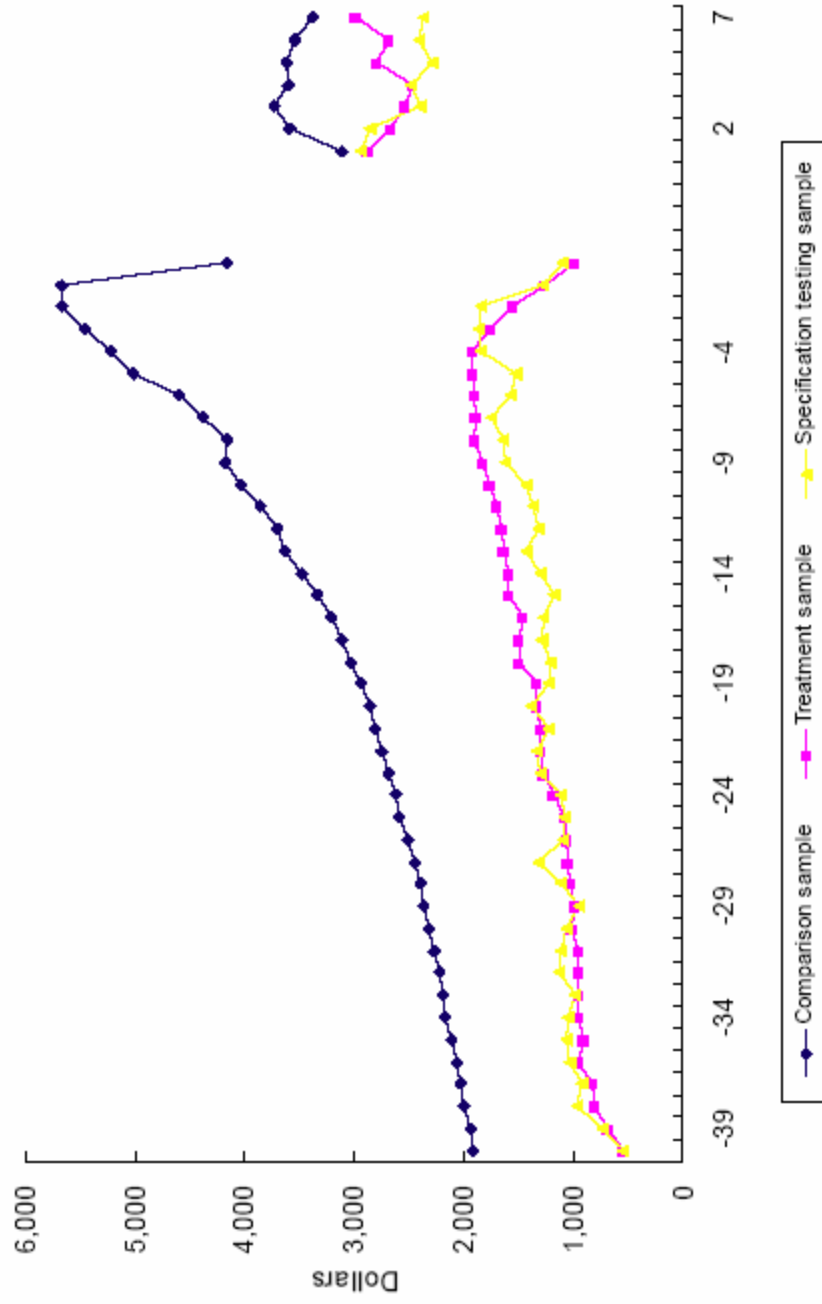
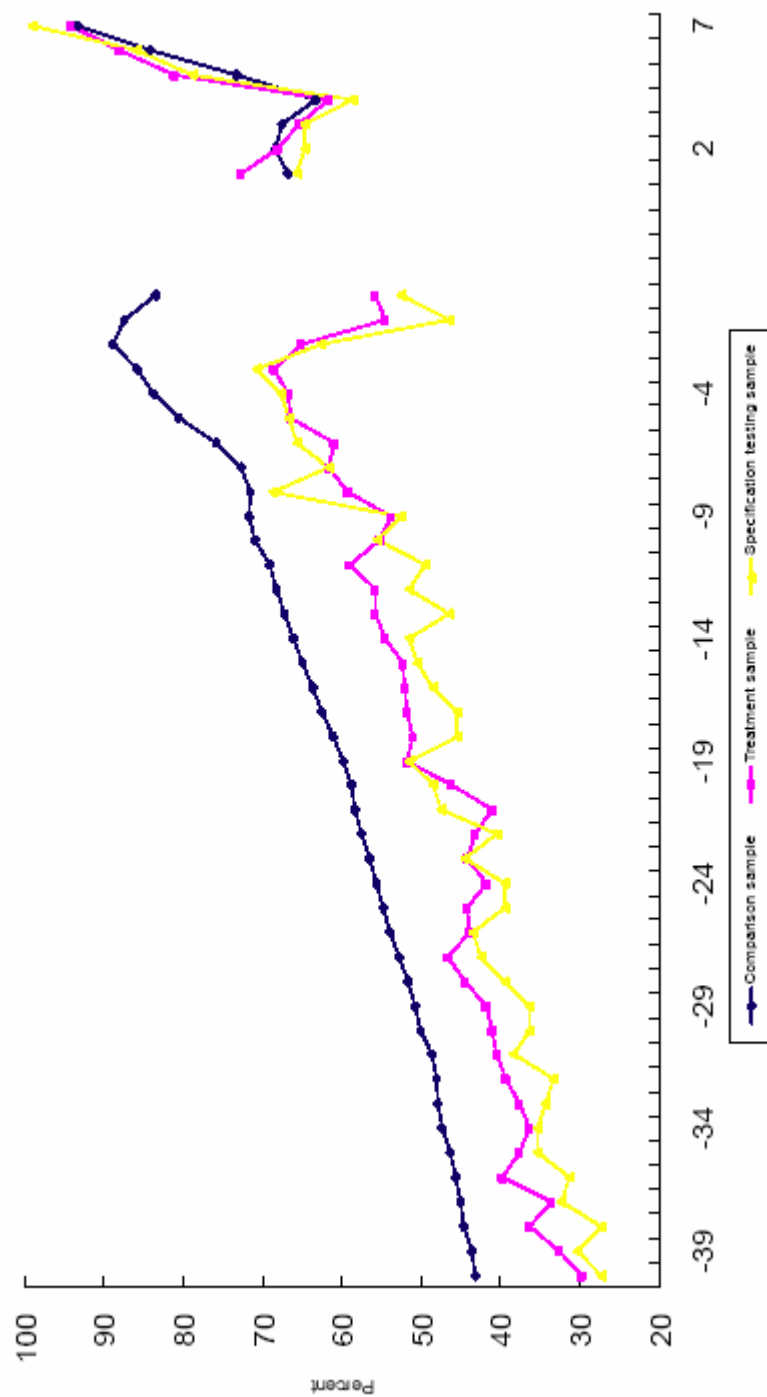


Figure 4 **Percent Employed, Male**
Comparison Sample: Full



**Figure 5 Percent Employed, Female
Comparison Sample: Full**

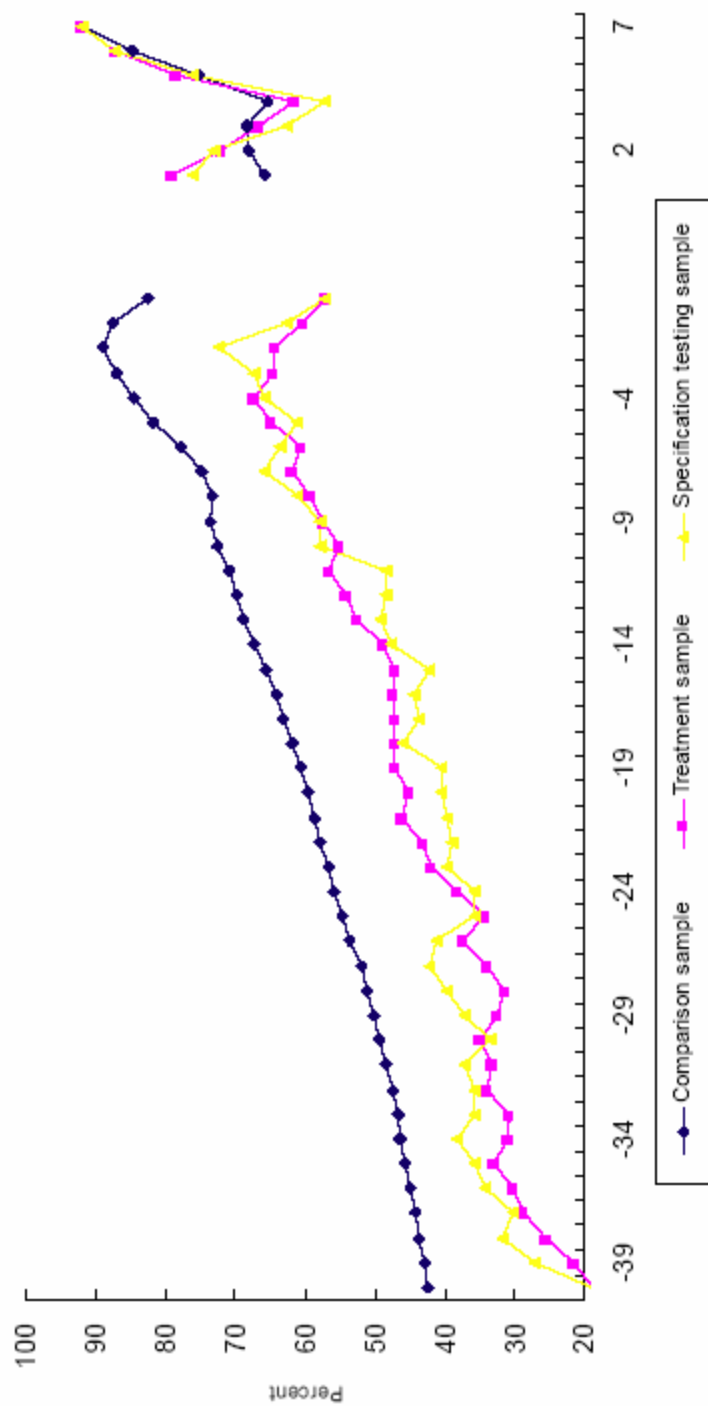


Figure 6 Distribution of Propensity Score

Male

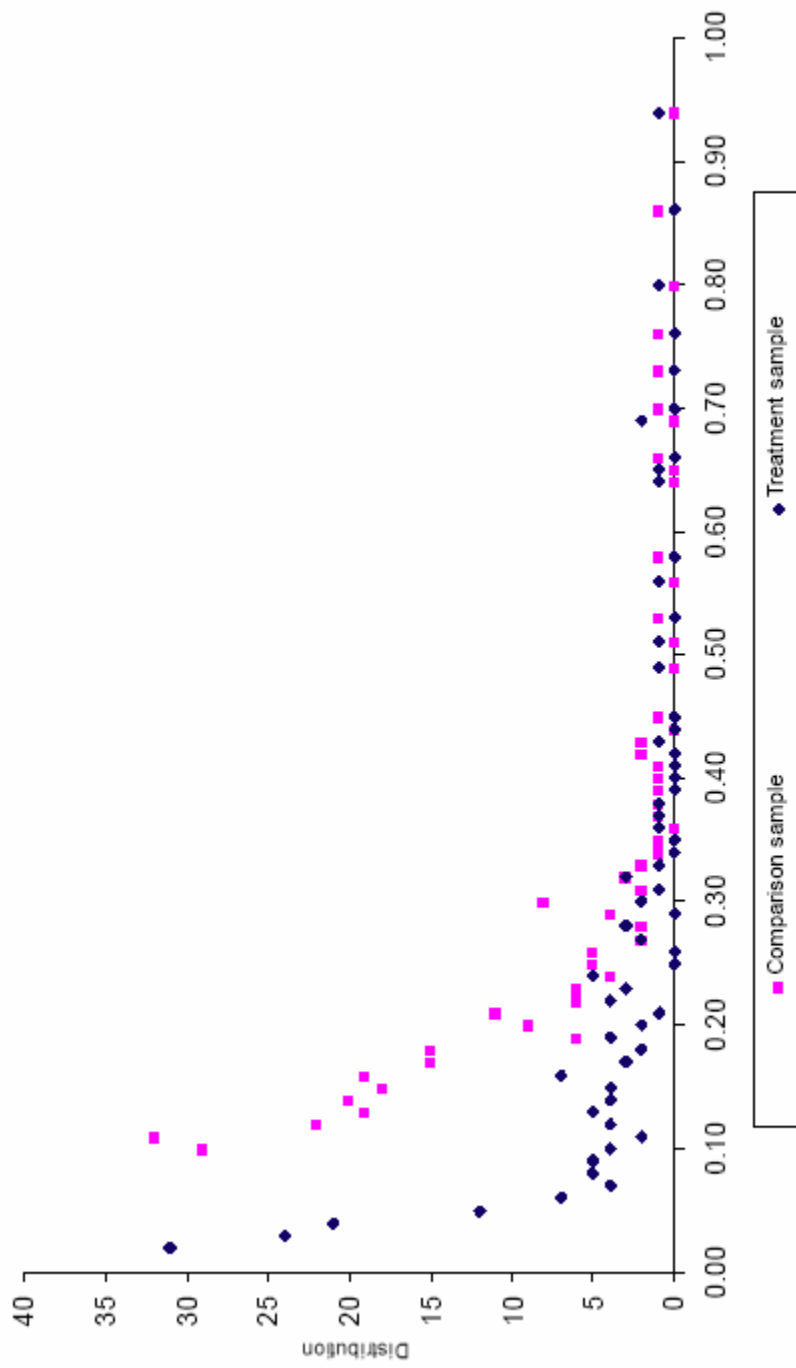


Figure 7 Distribution of Propensity Score
Female

