TECHNICAL WORKING PAPER SERIES

PREDICTING THE EFFICACY OF FUTURE TRAINING PROGRAMS USING PAST EXPERIENCES

V. Joseph Hotz
Guido W. Imbens
Julie H. Mortimer

Technical Working Paper 238
http://www.nber.org/papers/T0238

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
May 1999

The data used in this paper are derived from files made available to the researchers by MDRC. The authors remain solely responsible for how the data have been used or interpreted. This research was partially funded by a National Science Foundation grant (SBR 9818644) to the National Bureau of Economic Research, and by a UCLA graduate division fellowship. We are grateful for comments by Joshua Angrist, Richard Blundell, David Card, Gary Chamberlain, Janet Currie and participants in seminars at the NBER, Carnegie Mellon University, University of California, Berkeley, and UCLA. Any opinions expressed are those of the author and not those of the National Bureau of Economic Research.

Predicting the Efficacy of Future Training
Programs Using Past Experience
V. Joseph Hotz, Guido W. Imbens, and Julie H. Mortimer
NBER Technical Working Paper No. 238
May 1999

ABSTRACT

We investigate the problem of predicting the average effect of a new training program using experiences with previous implementations. There are two principal complications in doing so. First, the population in which the new program will be implemented may differ from the population in which the old program was implemented. Second, the two programs may differ in the mix of their components. With sufficient detail on characteristics of the two populations and sufficient overlap in their distributions, one may be able to adjust for differences due to the first complication. Dealing with the second difficulty requires data on the exact treatments the individuals received. However, even in the presence of differences in the mix of components across training programs, comparisons of controls in both populations who were excluded from participating in any of the programs should not be affected.

To investigate the empirical importance of these issues, we compare four job training programs implemented in the mid-eighties in different parts of the U.S. We find that adjusting for pre-training earnings and individual characteristics removes most of the differences between control units, but that even after such adjustments, post-training earnings for trainees are not comparable. We surmise that differences in treatment components across training programs are the likely cause, and that more details on the specific services provided by these programs are necessary to predict the effect of future programs. We also conclude that, because of treatment effect heterogeneity, it is essential, even in experimental evaluations of training programs, to record pre-training earnings and individual characteristics in order to render the extrapolation of the results to different locations more credible.

V. Joseph Hotz
Department of Economics, UCLA
405 Hilgard Ave.
Los Angeles, CA 90095
and NBER
hotz@ucla.edu

Guido W. Imbens
Department of Economics, UCLA
405 Hilgard Ave.
Los Angeles, CA 90095
and NBER
imbens@econ.ucla.edu

Julie H. Mortimer
Department of Economics, UCLA
405 Hilgard Ave.
Los Angeles, CA 90095
hollandj@ucla.edu

1. Introduction

Consider a government contemplating the implementation of a training (or other social assistance) program. The decision to implement the program depends on the assessment of its likely effectiveness. Often policy makers have access to data from a similar program implemented in an earlier time period or in another locality. The question arises as to how these data might be used to assess the contemplated program's likely efficacy. This situation is not uncommon. For example, the U.S. federal government's primary program for job training, the Job Training Partnership Act (JTPA), is designed and administered at the local level. Thus, policy makers may wish to evaluate differences in the effectiveness of the local programs. In addition, the recent federal reforms to the U.S. welfare system have encouraged the development of state and local program diversity, both in the services clients receive, and in the target populations. Increasingly, states and local authorities who administer their own programs, seek to use information from other programs, conducted in the past or in other locations, to assess the likely impacts and cost-effectiveness of these programs.

There are two distinct steps for predicting the effectiveness of a new program using data from previous programs. First, the researcher must evaluate the effectiveness of the initial program. Estimating the average effect of the initial program, for the entire population or for subpopulations, is straightforward if assignment to treatment was random. However, if the data were not generated by a carefully designed randomized experiment, there are fundamental difficulties in estimating the average causal effects. A large literature in econometrics examines complications in program evaluation using observational (non-experimental) data (e.g., Ashenfelter and Card, 1978; Heckman and Robb, 1984; Card and Sullivan, 1988; and Friedlander and Robins, 1995). In an influential paper, Lalonde (1986) showed that many conventional econometric methods were unable to recover estimates based on experimental evaluations.[1] Recently, Dehejia and Wahba (1998), using the same data as Lalonde, find that estimates based on matching and propensity score methods developed by Rosenbaum and Rubin (1983, 1984) were more successful in replicating experimental estimates.[2] Crucial to this success was the availability of sufficiently detailed earnings histories and background characteristics.

The second step in exploiting data from previous evaluations concerns generalizing the results of the previous evaluation to a new implementation. The focus of the current paper is on this second step. The issues associated with this step have received much less attention in the literature. Meyer (1995), in his discussion of natural experiments in economics, briefly describes problems associated with what he calls the "external validity" of evaluations, following the terminology of Cook and Campbell (1979). Dehejia (1997) analyzes the decision problem faced by an individual who, informed by data from a previous experimental evaluation, is considering whether or not to enroll in a training program. The current paper can be viewed as an examination of the credibility of the Dehejia research program: if, on the basis of data from a previous randomized experiment, one cannot predict what the *average* effect of a new program will be, it will be difficult to advise *specific individuals* based on such data. Finally, Manski (1997) considers the problem of predicting the average effect of alternative assignment rules applied to the same population. In contrast, we focus on an implementation using the same assignment rule (i.e., random assignment) but applied to a different population.

At least three distinct reasons may exist for differences in the average effect of treatment

[1] See also Fraker and Maynard (1987), and Heckman and Hotz (1989).

[2] See Angrist (1998), Heckman, Ichimura, Smith and Todd (1998), Lechner (1998), and Imbens, Rubin, and Sacerdote (1999) for other economic applications of matching and propensity score methods.

between two programs or localities. First, the distribution of characteristics in the two populations may differ. For example, suppose that the first population is older on average than the second. If age is associated with program efficacy, average program effects may differ for the two programs. Second, the programs, even if nominally the same, may differ in the mixture of, and assignment rules for, their treatment components. For example, one job-training program may stress classroom training, whereas another may emphasize job search assistance. Alternatively, one training program may be better run or organized than the other, even though each contains the same population and nominally the same treatment components. Third, there may be differences in average program effects because of differences in the size of the program. The same program, implemented on similar people, may yield different average treatment effects if a larger fraction of the population receives training in one of the implementations.

In this paper, we focus on the first two of these differences in treatment effects across programs.[3] With respect to the first source of differences, one can, in principle, remove the associated biases by dividing the population into homogeneous subpopulations based on the relevant background characteristics. Within these homogeneous subpopulations, average treatment effects should be identical in both implementations of the program. By weighting these "within" estimates appropriately, one can recover the population average treatment effects. Two potential problems exist with this approach. First, not all relevant characteristics are necessarily observed. Second, even if all the relevant characteristics are observed, the distributions of some characteristics need not overlap across implementations. If there is no overlap, one cannot adjust for differences in average treatment effects that depend

[3] While potentially important, we do not address the third complication. Rather, we assume in the analyses that follow that there is no interference between trainees, and thus that the scale of the program is not a source of different average treatment effects.

on these factors.[4] For example, suppose that treatment effects vary with local labor market conditions, indexed by local unemployment rates. Further, suppose that the prevailing unemployment rate differs across location or program. Then, there is no overlap in labor market conditions across these localities/programs and the assumptions necessary for valid adjustment with background variables is not met. We refer to these violations as "macro effects."

Similar difficulties arise with the second reason for differences in average treatment effects. One can adjust for biases associated with differences in the mix of program components if the treatment components received by trainees are recorded, and if assignment to these treatments is unrelated to potential outcomes. However, predicting the efficacy of a new implementation is difficult if information concerning components received is not coded in the available data. Even if such details are recorded, typically only the assignment to the binary treatment, "training" or "no-training," is randomized. Therefore, even estimation of average effects of specific components in the initial implementation is already wrought with difficulties, and extrapolation to other programs is less likely to be credible. We refer to the latter as the "heterogeneous treatment" problem.

In this paper, we explore the empirical relevance of these two sources of differences in average program effects by analyzing data from four random assignment evaluations of job-training programs run in different localities during the 1980s. Like Lalonde (1986), we use the estimates from randomized experiments to evaluate the performance of various non-experimental methods. We focus on three comparisons. First, we compare the average outcomes for controls in pairs of locations after adjusting for various sets of background

[4] The importance of overlap in the distribution of background characteristics is discussed in Rubin (1977) and its empirical relevance investigated in Dehejia and Wahba (1998) and Heckman, Ichimura, Smith and Todd (1998).

factors. Given sufficient adjustment for background characteristics and exclusion from all training services, control groups should be comparable across sites, in the absence of macro effects. Furthermore, control groups will be comparable regardless of any potential heterogeneous treatment effects for trainees. Second, we compare the average outcomes for trainees across pairs of locations. Comparison of the average outcomes for trainees isolates the validity of the homogeneous treatment assumption. Third, we compare average treatment effects in the two locations after various degrees of adjustment. This indicates the overall success of different adjustment procedures for eliminating biases due to macro and heterogeneous treatment effects.

An important part of our empirical analyses is assessing the effectiveness of alternative sets of pre-training and aggregate variables for eliminating the above sources of bias. A growing literature documents cases in which observational control groups suffice for unbiased program evaluation, once detail on pre-training labor earnings is available (e.g., Dehejia and Wahba, 1998; Heckman, Ichimura, Todd and Smith, 1998). Often such observational control groups come from public use surveys (e.g., Lalonde, 1986) or eligible nonparticipants from the same experiment (Heckman, Ichimura, Todd, and Smith, 1998). Our non-experimental control groups are taken from experiments in other locations, and they may therefore be subject to different biases. Thus, we investigate whether and what type of pre-training variables can eliminate such biases.

In the next section we set up the inferential problem in the potential outcome notation for causal modelling. We demonstrate the close connection between the problem of predicting the effects of one program with data from an evaluation of another and the problem of evaluating programs using non-experimental data. We limit our attention in Section 2 to the case with binary, homogeneous treatments. Complications arising from the presence of macro effects are discussed in Section 3. In Section 4 we allow for heterogeneous treatments. Section 5 contains a discussion of issues related to estimation. An application of these ideas to four Work INcentive (WIN) training programs is presented in Sections 6 through 8. Section 9 concludes.

2. The Role of Unconfoundedness

A random sample of size N is drawn from a large population. Each unit i, for $i = 1, 2, \ldots, N$, is from one of two locations, indicated by $D_i \in \{0, 1\}$. For each unit there are two potential outcomes, one denoted by $Y_i(0)$, describing the outcome that would be observed if unit i received no training, and one denoted by $Y_i(1)$, describing the outcome given training. Implicit in this notation is the Stable Unit Treatment Value Assumption (SUTVA) of no interference and homogeneous treatments (Rubin, 1974, 1978). In Section 4 we shall relax this assumption to allow for heterogeneous treatments but maintain the assumption of no interference. In addition, there is, for each unit, an indicator for the treatment received, $T_i \in \{0, 1\}$ (with $T_i = 0$ corresponding to no-training, and $T_i = 1$ corresponding to training), and a set of covariates or pretreatment variables, X_i. The realized outcome for unit i is $Y_i \equiv Y_i(T_i) = T_i \cdot Y_i(1) + (1 - T_i) \cdot Y_i(0)$.

We are interested in the *average training effect* for the $D_i = 1$ population:

$$\tau_1 = E[Y_i(1) - Y_i(0) | D_i = 1],$$

or the average treatment effect for a subpopulation within this program,

$$\tau_1(x) = E[Y_i(1) - Y_i(0) | D_i = 1, X_i = x].$$

We wish to estimate this on the basis of N observations $(X_i, D_i, (1 - D_i) \cdot T_i, (1 - D_i) \cdot Y_i)$. That is, for units with $D_i = 0$ we observe the covariates X_i, the program indicator D_i, the treatment T_i and the actual outcome Y_i. For units in the second program with $D_i = 1$ we observe covariates X_i and the program indicator D_i but neither the treatment status nor the realized outcome.

We assume that in the $D_i = 0$ program assignment was random:

Assumption 1 (Random Assignment)

$$T_i \perp \Big(Y_i(0), Y_i(1)\Big) \Big| D_i = 0.$$

Random assignment of subjects to trainee and control status implies we can estimate the average effect of training in the initial implementation by comparing average outcomes by training status:

Lemma 1 (Identification of Average Treatment Effects Given Random Assignment)
Suppose Assumption 1 holds. Then:
(i) the population average treatment effect, $\tau_0 = E[Y_i(1) - Y_i(0)|D_i = 0]$, *is equal to* $E[Y_i|T_i = 1, D_i = 0] - E[Y_i|T_i = 0, D_i = 0]$ *which can be estimated from the population distribution of outcomes and treatment indicators.*
(ii) the average effect within subpopulations defined by covariates, $\tau_0(x) = E[Y_i(1) - Y_i(0)|X_i = x, D_i = 0]$, *is equal to* $E[Y_i|T_i = 1, D_i = 0, X_i = x] - E[Y_i|T_i = 0, D_i = 0, X_i = x]$ *which can be estimated from the population distribution of outcomes, covariates, and treatment indicators.*[5]

Proof: See Appendix A

[5] Note that in the evaluation of training programs τ_0 is typically the average effect of interest. This is so if the subpopulation subject to randomization is the target population that otherwise would be provided with some set of training services.

In other words, random assignment solves the problem of causal inference by eliminating any bias from comparisons of trainees and controls.

The simplest condition under which we can estimate the average effect for the $D_i = 1$ program is if location is random with respect to outcomes:

Assumption 2 (Random Location)
A location (or program) is said to be random if

$$D_i \perp \Big(Y_i(0), Y_i(1) \Big).$$

Unlike random assignment of treatment, the mechanism that would guarantee this assumption is not very practical. In particular, the only way to guarantee this assumption by design is to randomly assign units in the population to different locations. Note that it is *not* sufficient to randomly choose the location of the initial implementation.[6] In general, one suspects that units are not randomly assigned across locations, rather units (individuals) choose where to live.

Random location implies that the joint distribution of outcomes is the same in both subpopulations, and hence the expected average outcomes are the same in the two implementations of the program. Combined with random assignment in the initial program, this implies that the average effect in the second program is identified:

Lemma 2 (External Validity Given Random Location)
Suppose assumptions 1-2 hold. Then:

$$E[Y_i(1) - Y_i(0)|D_i = 1] = E[Y_i|T_i = 1, D_i = 0] - E[Y_i|T_i = 0, D_i = 0].$$

[6] We note that although not sufficient for identifying the average effect of a specific program, it can be of interest to randomly select locations/programs as was done in the National JTPA Study. This type of randomization, discussed in Hotz (1992), is appropriate when interest centers on obtaining an average treatment effect for a population of programs.

Proof: see Appendix A.

The random location assumption can be relaxed in the presence of pretreatment variables:

Assumption 3 (UNCONFOUNDED LOCATION)
Location of program is unconfounded given pretreatment variables X_i if

$$D_i \ \perp \ \Big(Y_i(0), Y_i(1)\Big) \ \Big| \ X_i. \tag{1}$$

In addition we require complete overlap in the covariate distributions:

Assumption 4 (COMPLETE OVERLAP)
There is complete overlap in the distribution of the pretreatment variables X_i if

$$0 < Pr(D_i = 1|X_i = x) < 1,$$

for all x in the support of X.

The complete overlap assumption implies that for all values of the covariates one can find units in both subpopulations, or, alternatively, that the support of the covariate distributions is identical in the two subpopulations. If it is not satisfied, one can redefine the estimand of interest as the conditional treatment effect in the subpopulation with common covariate support, essentially dropping units with covariates outside the common support.

Under Assumptions 1, 3, and 4, the following results holds:

Lemma 3 (EXTERNAL VALIDITY GIVEN UNCONFOUNDED LOCATION)
Suppose assumptions 1, 3, and 4 hold. Then:

$$E[Y_i(1) - Y_i(0)|D_i = 1]$$

$$= E\Big[E[Y_i|T_i = 1, D_i = 0, X_i] - E[Y_i|T_i = 0, D_i = 0, X_i]\Big|D_i = 1\Big].$$

Proof: See Appendix A.

Unconfounded location implies that within subpopulations that are homogeneous in covariates, that is, conditional on covariates, the average treatment effects in the two locations are identical. Complete overlap implies that in each subpopulation it is feasible to estimate these effects. Therefore, we obtain the population average training effect by averaging over the "within" group training effect estimates.

There are two aspects of Assumption 3, the key to identification, worthy of comment. First, note the similarity between the unconfounded location assumption and the unconfounded assignment assumption that is often made in non-experimental evaluations. The precise formulation of unconfounded treatment assignment (Rosenbaum and Rubin, 1983), is

$$T_i \ \perp \ \Big(Y_i(0), Y_i(1)\Big) \ \Big| \ X_i. \tag{2}$$

This similarity underscores the symmetry between the two parts of the prediction problem: evaluation of the initial program and generalization to the new program. In substantive terms, the two assumptions, however, are very different. Randomization of treatments within a site guarantees unconfoundedness of the assignment – it does not address the unconfoundedness of location per se. Even though the decision about where to implement the initial version of the program is typically out of the control of the populations involved, location is typically the individuals' choice. In other words, unconfoundedness of treatment assignment is often a design issue, whereas unconfoundedness of location is likely to be a modeling issue. In the absence of detailed background characteristics on the individuals, there is no compelling reason to think that the unconfounded location assumption is plausible. On the other hand, if the second implementation is in the same geographical area and occurs shortly after the initial implementation, unconfoundedness of location may well be plausible, without speaking to the plausibility of the unconfounded assignment assumption.

The second comment is that the prediction problem, under the assumptions invoked, has a missing data interpretation. Under the prediction problem, we are interested in the average difference between trainees and controls in the $D_i = 1$ subpopulation given random assignment of the treatment in this subpopulation. Our unconfounded location assumption, in the presence of random assignment of treatment, essentially amounts to assuming that the missing outcomes, for the $D_i = 1$ subpopulation, are Missing At Random in the sense of Rubin (Rubin, 1976). To see this, note that for units with $D_i = 1$, the treatment indicator T_i and the outcome Y_i are missing, that is, not observed. The Missing At Random assumption implies that

$$D_i \ \perp \ \big(Y_i, T_i\big) \ \Big| \ X_i.$$

Furthermore, the random assignment of treatment assumption implies

$$D_i \ \perp \ T_i \mid X_i.$$

Combining these two assumptions implies that

$$D_i \ \perp \ Y_i \ \Big| \ X_i, T_i.$$

Substituting for $Y_i = T_i \cdot Y_i(1) + (1 - T_i) \cdot Y_i(0)$, this implies

$$D_i \ \perp \ Y_i(0) \ \Big| \ X_i, T_i = 0, \quad \text{and} \ \ D_i \ \perp \ Y_i(1) \ \Big| \ X_i, T_i = 1.$$

Since $D_i \perp T_i | X_i$ by random assignment, it follows that conditioning on T_i is immaterial so invoking the MAR assumption amounts to assuming

$$D_i \ \perp \ Y_i(0) \ \Big| \ X_i, \quad \text{and} \ \ D_i \ \perp \ Y_i(1) \ \Big| \ X_i,$$

which is the essence of the unconfounded location assumption.

3. Macro Effects

A threat to the validity of the weighted "within" estimators of treatment effects implied by Lemma 3 is the presence of "macro effects." Differences between the two populations may be unadjustable, or the environments of the two populations may differ. Such differences can be violations of either the unconfounded location assumption or the complete overlap assumption.

First, there may be an additional covariate that is correlated with the outcomes of interest and has a distribution that differs across the two locations. If we do not observe this variable, its presence may prevent recovery of the average effects in the second location. For example, suppose the training program is more effective for younger people than for older people, and hence age is correlated with the outcome. If the average age differs between locations, the average treatment effect in the second location cannot be estimated using data from the first location unless the pre-training variables include age. We refer to this as the "unobserved covariate" interpretation, viewing the macro effects problem as one of violating the unconfounded location assumption.

A second interpretation also postulates the existence of an additional variable such that the unconfounded location assumption is satisfied conditional on this variable. If the distribution of this variable does not overlap in the two populations, even observing it does not permit estimation of average causal effects. For example, if the initial implementation is only on men, and the second implementation is only on women, differential efficacy of the program by gender may preclude accurate prediction of the average effect of the second implementation. We refer to this as the "no-overlap" interpretation, viewing the macro effects problem as one of violating the complete overlap assumption.

Although formally distinct, the two interpretations often capture the same problem. Suppose, for example, that the policy maker is concerned with differences in the demand for labor in two programs in different locations. In this case, one should adjust for individual-level differences in demand conditions. One can attempt to approximate the individual-level demand conditions by controlling for aggregate demand conditions. In this case, there may be no overlap in the (possibly within-location, degenerate) distribution of the covariate.

Unfortunately, it may not be possible to rule out macro effects by design. A typical case is one in which the initial implementation occurred in the past and the policy maker is interested in the effect of a future implementation. Given the difference in the timing of implementation, there is no guarantee that conditions are ever similar enough to allow accurate predictions. The design strategies for addressing this problem involve: (i) using initial implementations in locations that are as similar as possible in characteristics and time to the location of interest and/or (ii), collecting as much detail as possible on covariates that can proxy for differences in conditions across locations and time.

4. Heterogeneous Treatments

A complication that has been ruled out so far is heterogeneity in treatments. It is rare, even in the context of randomized evaluations of training programs, that all individuals receive exactly the same treatments in a particular program. More typically, individuals are assigned to different "tracks" based on additional screenings that occur after an initial random assignment to "training" or "no-training". Some tracks may involve classroom training, while others involve on-the-job training or job search assistance. Here we investigate the implications of treatment heterogeneity on strategies for predicting the effect of future programs.

Formally, consider a training program with $K+1$ training components. For each training component t, with $t \in \mathcal{T} = \{0, 1, \ldots, K\}$, and each unit i, with $i = 1, \ldots, N$, there is a potential outcome $Y_i(t)$. For unit i, $\tilde{T}_i \in \mathcal{T}$ is the treatment component received. The researcher only observes the binary treatment assignment, $T_i = 1\{\tilde{T}_i \geq 1\}$. The null treatment $T_i = 0$ corresponds to no training at all. Randomly selected individuals are assigned to this option in the initial location, or

$$T_i \perp \Big\{Y_i(0), \ldots, Y_i(K)\Big\} \ \Big| \ D_i = 0.$$

Different treatment components may correspond to various treatment options, e.g., combinations of classroom training and job search assistance. Conditional on getting training ($T_i = 1$), assignment to the different training components is not necessarily random in the initial location. That is

$$\tilde{T}_i \not\perp \Big\{Y_i(0), \ldots, Y_i(K)\Big\} \ \Big| \ D_i = 0.$$

To define the average outcome under training in the new location we assume there is some, possibly stochastic, rule that determines the treatment component, given assignment to training. This rule can be summarized by the conditional distribution of $\tilde{T}_i$ given training ($T_i = 1$), in each location D_i, given covariates X_i, and given potential outcomes $(Y_i(0), \ldots, Y_i(K))$.

Under the unconfounded location assumption, it is still true that average outcomes in the new location, conditional on no training, can be estimated without additional assumptions.

Lemma 4 (External Validity for Control Outcomes Given Unconfounded Location under Treatment Heterogeneity)

Suppose Assumptions 1, 3, and 4 hold. Then:

$$E[Y_i(0)|D_i = 1] = E\Big[E[Y_i|T_i = 0, D_i = 0, X_i]\Big|D_i = 1\Big].$$

Proof: See Appendix A

However, in general one cannot estimate the average outcomes for trainees:

$$E[Y_i(1)|D_i = 1] \neq E\Big[E[Y_i|T_i = 1, D_i = 0, X_i]\Big|D_i = 1\Big].$$

Hence only comparisons between controls in both locations would be valid and accurate predictions of causal effects cannot be obtained without additional assumptions.

Under two assumptions, estimation of the average effect of the second implementation is still feasible. The first assumption requires that the assignment rule be identical in the two locations:

Assumption 5 (Conditional Independence of Assignment and Location)

$$D_i \;\perp\; \tilde{T}_i \;\Big|\; Y_i(0), \ldots, Y_i(K), X_i.$$

The second assumption requires that assignment is independent of potential outcomes, given covariates in each location:

Assumption 6 (Unconfounded Assignment)

$$\tilde{T}_i \;\perp\; Y_i(0), \ldots, Y_i(K) \;\Big|\; X_i, D_i.$$

This second assumption implies that assignment is unconfounded in both locations. Hence, if one actually observed the treatment component, one could estimate the average effect for each component by adjusting for covariates. Even without observing the treatment component, we can now estimate the average effect of assignment to training. The first step is the implication of the two assumptions that assignment to treatment is independent of location *and* potential outcomes within subpopulations defined by covariates:

$$\tilde{T}_i \ \perp \ D_i, Y_i(0), \ldots, Y_i(K) \ \Big| \ X_i.$$

Hence:

Lemma 5 (External Validity for Trainee Outcomes Given Unconfounded Location under Treatment Heterogeneity)
Suppose assumptions 1, 3, 4, 5, and 6 hold, with Assumption 3 now assuming independence of D_i and the set of $K+1$ potential outcomes conditional on X_i. Then:

$$E[Y_i(1)|D_i = 1] = E\Big[E[Y_i|D_i = 0, T_i = 1, X_i]\Big|D_i = 1\Big].$$

Proof: See Appendix A.

Violations of Assumption 6, unconfounded assignment, are difficult to overcome. When assignment is related to potential outcomes, it is difficult to estimate the effect of different treatment components in the initial implementation, even if the treatment component is observed. In this case, there is little hope of using the results from an initial implementation, with its particular assignment rule, to predict the effect of future assignment rules. Violations of Assumption 5 do not necessarily invalidate all methods of inference. However, in order to fully adjust for differences in assignment rules, one needs to observe the actual treatment component received by each unit. With only limited information on treatments, such as the distribution of rates of participation in various program components for both locations, one may still be able to calculate bounds on the average treatment effects (e.g., Manski, 1997; Hotz, Mullin, and Sanders, 1997).

5. Evaluating Unconfoundedness Using Randomized Experiments

Suppose we have available two randomized experiments that evaluate the same training program in two different locations. Under unconfounded location, we can estimate the average outcome for controls in the second location in one of two ways. First, we can estimate the average outcome directly with the estimator

$$\widehat{E[Y_i(0)|D_i = 1]} = \sum_{i|D_i=1} Y_i \cdot (1 - T_i) \Big/ \sum_{i|D_i=1} (1 - T_i).$$

using the equality

$$E[Y_i(0)|D_i = 1] = E[Y_i|D_i = 1, T_i = 0],$$

implied by random assignment in the second experiment. Second, we can exploit the equality

$$E[Y_i(0)|D_i = 1] = E\Big[E[Y_i(0)|D_i = 1, X_i]\Big|D_i = 1\Big]$$

$$= E\Big[E[Y_i(0)|D_i = 0, X_i]\Big|D_i = 1\Big].$$

Estimators based on the second approach do not use the outcomes in the second experiment, and therefore are functionally independent of the first estimator. Under unconfounded location, the two estimators should be close and statistical tests can be based on their comparison. A similar argument can be used to construct tests based on outcomes for trainees. Finally, we can combine the two procedures to get estimates for the average causal effect in the second location. There is, of course, the possibility that biases in outcomes for trainees and controls offset each other and lead to unbiased estimates of the average causal effect. This could occur in the presence of persistent additive trends in outcomes when locations are separated by time. Transformations of the outcomes could then be used to eliminate such trends.

In practice, this procedure requires estimation of the average outcome conditional on different sets of covariates. With many covariates this may be a difficult task. However, techniques identical to those based on the propensity score in program evaluation can be used to reduce the dimension of the conditional estimation problem (e.g., Rosenbaum and Rubin, 1983, 1984). These dimension-reduction procedures rely on the predicted location given covariates, or the location score:

$$l(x) = Pr(D_i = 1|X_i = x).$$

The unconfounded location assumption given covariates implies unconfounded location given the location score:

$$D_i \;\perp\; (Y_i(0), Y_i(1)) \;\Big|\; l(X_i).$$

Hence we can modify the equalities that form the basis for the second estimator to condition solely on the location score rather than on the entire set of pre-training variables:

$$\begin{aligned} E[Y_i(0)|D_i = 0] &= E\Big[E[Y_i(0)|D_i = 0, l(X_i)]\Big|D_i = 1\Big] \\ &= E\Big[E[Y_i(0)|D_i = 0, l(X_i)]\Big|D_i = 1\Big]. \end{aligned}$$

In practice we estimate the location score $l(x)$, using a flexible parametric model such as a logistic regression model, possibly with higher-order terms and interactions by inspecting balance within blocks defined by the location score. The same strategies for searching for the specification of the standard propensity score are relevant here. See Dehejia and Wahba (1998) for more details on implementing the propensity score methodology.

6. Data

We investigate the problem of predicting the effects of future training programs from past experiences using data from four experimental evaluations of WIN (Work INcentive) demonstration programs. The programs were implemented in Arkansas, Virginia, San Diego, and Baltimore. These programs differ in timing, location, target population, funding and program activities. We briefly describe each of the four programs.[7]

The training services offered in the Arkansas WORK program consisted primarily of group job search and unpaid work experience for some trainees. It targeted AFDC applicants and recipients with children at least three years old, and the average cost of providing these services was $118 per trainee. The evaluation of this program started in 1983 and covered two counties. The training services under the Virginia Employment Services Program (ESP) included both job search assistance and some job skills training and targeted AFDC applicants and recipients with children at least six years old. It cost an average of $430 per trainee. This evaluation also began in 1983 and included five counties. The Saturation Work Initiative Model (SWIM) in San Diego targeted AFDC applicants and recipients with children at least six years old and provided job search assistance, skills training and unpaid work experience. The average cost in this program was $919 per trainee and its evaluation was begun in 1985. Finally, the Baltimore Options program provided job search, skills training, unpaid work experience and on-the-job training and targeted AFDC applicants and recipients with children at least six years old. The Baltimore program was the most expensive of the four programs, with an average cost of $953 per trainee. This evaluation began in 1982.

[7] See Gueron and Pauly (1991), Friedlander and Gueron (1992), Greenberg and Wiseman (1992), and Friedlander and Robins (1995) for more detailed discussions of each of these evaluations.

Four modifications were made to the basic data sets. First, individuals with children less than six years old were excluded from the analyses because of the severe imbalance in their distribution across programs. (Individuals with children under six were only targeted for inclusion in Arkansas.) Second, men were excluded from our analyses, as men were not part of the target population in Virginia and comprised only small fractions of the sample in the other locations (10% in Maryland, 9% in San Diego and 2% in Arkansas). Third, women without children were also excluded from the analyses. Although such households were present in all locations, they never made up more than 4% of the sample in any of the locations. Finally, we added two aggregate variables, the employment to population ratio and real earnings per worker, each measured at the county level, to account for differences in labor market conditions across the four locations.

Table 1 gives means and standard deviations for all pre-training variables and outcomes common to all four programs for the main sample used in the analyses. The individual pre-training variables fall into two categories: personal characteristics and earnings histories. We observe whether the woman has a high school diploma, whether she is non-white, whether she ever married, and whether the number of children is more than one. The earnings history variables consist of total earnings for each of the four quarters preceding randomization. We report summary statistics for the actual earnings and for an indicator of positive earnings in each of the four quarters. For the three programs other than San Diego, a t–statistic is reported for each variable corresponding to the test of the null hypothesis that the average value in that program is the same as the average in San Diego. These tests anticipate an attempt to predict average causal effects of the training program in San Diego using results from the other three locations. Finally, summary statistics are provided for the two post-training outcomes used in the analyses, employment and earnings for the first and second year respectively, as well as estimates of the average effects of the four programs.

The t-statistics show clearly that the four locations are very different in terms of the populations served. For example, in Arkansas approximately 16% of the population was working in any given quarter prior to randomization, whereas in Baltimore the incidence of working prior to randomization was as high as 30%. The percentage white ranged from 16% in Arkansas to 33% in Virginia. The percentage with a high school degree ranged from 40% in Baltimore to 55% in San Diego. Therefore, it is not surprising that post-training earnings also differ considerably by location. The estimates of the effect of the training program also differ across the four locations. In the first year, the effect of training on employment ranges from two percent in Arkansas to twelve percent in San Diego. The same effect in the second year varies from three percent in Baltimore to nine percent in San Diego, with both differences statistically significant.

7. Analyses

We focus on San Diego as the program whose results we wish to predict from the results of the other three programs. San Diego was chosen largely because the nature and cost of the SWIM program differed markedly from the other locations.[8] Also, San Diego's population had relatively high average earnings compared to the other three programs. Finally, San Diego was geographically separated from the other programs. Considering these differences, one might expect that the treatment effects for the San Diego program would be the most difficult program to predict, given data from the other three implementations.

We focus on six issues for predicting the effect of the San Diego program using data from the other programs. First, we examine the importance of restricting the samples so as to

[8] The San Diego SWIM program and the Baltimore Options program were the two most expensive programs, at $919 and $953 per trainee respectively, compared to $430 for Virginia and $118 for Arkansas.

ensure overlap in the distribution of pre-training variables. Second, we predict outcomes for controls and trainees separately to highlight the potential for heterogeneous training effects. Third, we analyze the sensitivity of the results to the choice of location. Fourth, we consider alternative methods of prediction, including least squares adjustment and propensity score methods. Fifth, we consider the importance of the individual-level pre-training variables. Finally, we investigate the importance of using aggregate, county-level information to account for labor market condition differences between San Diego and the other three locations.

One issue that requires some additional motivation is the interest in predicting average outcomes separately for controls and trainees, in addition to predicting average treatment effects. One can argue that it is sufficient to be able to predict average treatment effects and that failure to predict average outcomes for controls and trainees separately would be of little concern. However, unless we can predict the distribution of outcomes for trainees and controls in the second implementation, we cannot predict the average training effect for all transformations of any particular outcome variable. Specifically, suppose we can predict the average effect of the training on the level of earnings. Unless we can predict the distribution of earnings for both trainees and controls, in general this does not imply that we can predict the average effect of training on the logarithm of earnings or on the probability of having positive earnings. Abandoning the search for methods that allow prediction of average outcomes for trainees and controls in favor of a focus solely on differences implies tying oneself to a specific transformation of the outcome.

7.1 Outcomes

We consider the following four outcomes: an indicator for employment and total earnings, each measured in the first and second years after randomization. For each of these outcomes, we make four comparisons: San Diego with all three other locations together and San Diego with each of the three other locations separately. For each pair of locations and for each outcome, we then make predictions of average outcomes for controls, for trainees, and for average program effects. Finally, for each comparison we use a number of different procedures for estimating the difference, as described below.

7.2 Methods

We predict outcomes in San Diego using four methods. First, we predict the outcomes in San Diego using the average outcome in the other locations. This "level" prediction does not adjust for any of the differences between San Diego and the other locations. The second method predicts the gain (the change in earnings or employment relative to the pre-randomization year) in San Diego using the gain in the other locations. Third, we use least squares methods to predict the outcomes in San Diego using some of the pre-training variables. Fourth, we predict San Diego outcomes using propensity score methods (Rosenbaum and Rubin, 1983, 1984; Dehejia and Wahba, 1998). For details on the implementation of each of these methods, see Appendix B.

For the least squares and propensity score methods, we also must choose pre-training variables for the adjustment procedures. The pre-training variables included are the four personal characteristic variables (indicators for high school degree, non-white, never married, one child), four quarters of pretreatment earnings, and four dummies for whether these quarterly pre-training earnings are positive. In addition, other pre-training variables are defined at a more aggregate level, such as the unemployment rate or price levels. Ideally we would like to compare individuals faced with the same local unemployment rate and price level. This is less likely to be feasible, however, given the few locations for which we have data. Hence, adjustments for differences in these aggregate pre-training variables by necessity rely more on modelling and smoothing assumptions.

In the analyses below we incorporate two aggregate measures of labor market conditions: the ratio of employment to population, and average real earnings per worker. Both are measured at the county level. We use these in two different ways. First, we include both variables as regressors in least squares adjustments. Second, we deflate the individual-level earnings measures by the ratio of real earnings in San Diego to the average real earnings in the alternative location. Specifically, if individual i is from Arkansas, we modify earnings in the first quarter prior to randomization by the ratio of real earnings in the year prior to randomization in San Diego to real earnings in the year prior to randomization in Arkansas. These ratios may differ from because of differences in location and in the calendar years in which the various evaluations were conducted. The use of these adjustments are an attempt to make the earnings measures more comparable. Without such transformations it is difficult to believe that earnings histories for individuals in Arkansas are informative about individuals in San Diego with identical earnings histories.

8. Results

8.1 The Importance of Overlap

For most comparisons, we construct a basic data set by discarding observations with little or no overlap in their distributions across locations, as described in Section 6. Tables 2 and 3 provide some detail on these discarded observations and the motivation for discarding them. In Table 2, we present the sample means of variables for the discarded observations in each location. We test the null hypothesis that the average for these discarded observations is the same as the average for the basic data set in San Diego. Note that these t-statistics are not much larger than those in Table 1, which correspond to a test of mean differences using the observations we do not discard. This suggests that discarding these observations may not affect our results.

To further investigate the importance of overlap, we present more detailed tests on the discarded observations for the Baltimore program in Table 3. The discarded observations for Baltimore consist of men, women with children under six, and women with no children. Table 3 shows that insignificant differences between discarded Baltimore observations and the San Diego sample is often the result of offsetting differences. For example, consider earnings in quarter −4. Earnings for men in Baltimore are significantly higher than earnings in the basic San Diego sample (women with children over six years old). However, earnings for women with children under six in Baltimore are much lower than earnings in the basic San Diego sample. Combining the two discarded groups leads to a sample that is, on average, not so different from the basic San Diego sample. However, it is difficult to believe that combining men and women with young children provides a suitable control group for women with older children.

8.2 Predicting Average Outcomes for Controls and Trainees versus Predicting Average Training Effects

Table 4A presents the prediction errors from predicting the first outcome, average earnings in the first post-randomization year, for controls. In the first column, average outcomes are predicted for San Diego using all three other locations. The first row uses the average of first-year earnings in the other locations as a predictor for average first-year earnings in San Diego. It shows that earnings in the first post-randomization year are $240 higher in San Diego than in the other three locations. This difference of $240 is statistically not significantly different from zero, with a t-statistic of 1.8. In the second row, the average *gain* in earnings for San Diego is predicted using the average gain in the other three locations. Now the discrepancy is only $130. Rows 3 and 4 predict the average first-year earnings in

San Diego using least squares and propensity score methods. Both predictions underestimate average earnings by \$140. The full set of individual background characteristics, as well as aggregate variables, are used for this prediction. Tables 5A-7A report results for the other outcomes. Tables 4B-7B report results for trainees, and Tables 4C-7C predict average training effects.

For the two earnings levels outcomes, earnings in the first and second year following randomization, there is a very clear result. We can predict the average outcomes for the controls fairly accurately, but cannot predict the outcomes for trainees very well. Consider earnings in the first year. Using data from all three alternative locations, both least squares and propensity score predictions for average earnings in San Diego are off by approximately \$140 for controls. Testing the null hypothesis that the prediction is equal to the average outcome in San Diego leads to t-statistics of 1.2 and 1.1 for least squares and propensity score estimates, respectively. For trainees, the difference between predicted and actual averages are \$290 and \$320, significantly different from zero with t-statistics of 2.9. The second year earnings results are equally clear. For controls, the predictions, using least squares and propensity score methods, are off by \$20 and \$50 respectively, and are not significantly different from zero. For trainees, the corresponding prediction errors are \$180 and \$250, with t-statistics of 1.3 and 1.6. In both cases, we cannot predict the average treatment effect in San Diego because we fail to predict one of the components.

For the employment indicators, the results are not quite as clear. In the first year after randomization, we clearly predict average outcomes better for controls than for trainees. However, in the second year, outcomes are more accurately predicted for the trainees than for the controls. In this respect, it is useful to consider the patterns of employment in the different locations. In the year prior to randomization, the fraction with positive earnings in

San Diego was 0.390, higher than Arkansas at 0.260 and Virginia at 0.365, but lower than Baltimore at 0.442. During the first year after randomization, the proportion of controls with positive earnings in San Diego was 0.401, compared to 0.267 in Arkansas, 0.426 in Baltimore and 0.454 in Virginia. In the second post-randomization year, the proportions are 0.398 in San Diego, 0.271 in Arkansas, 0.465 in Baltimore, and 0.461 in Virginia. Thus, it appears that employment trends changed substantially during the second year of the Baltimore evaluation. Given the similarities between Baltimore and San Diego in earlier outcomes, this contributes heavily to the failure to predict average employment rates in the second post-randomization year.

8.3 Single Locations versus Combined Locations

The four locations differ considerably in background characteristics and labor market histories. For example, 56% of San Diego observations have a high school diploma, compared to only 40% in Baltimore and 43% in Virginia. In the quarter prior to randomization, 27% of the sample from San Diego had positive earnings, compared to only 17% in Arkansas, 29% in Baltimore and 25% in Virginia. This suggests that a single comparison group might not work very well in predicting outcomes in San Diego. This hypothesis is supported by the data. Consider the first-year earnings for controls. We predict average earnings in San Diego using all three alternative locations (with a prediction error of $140), but the prediction error is $770 using only Arkansas data. Similarly, we predict employment in the first year accurately using the combined control group (a prediction error of 3%), but the prediction performs quite poorly using only the Virginia data (a prediction error of 8%). Only the Baltimore control group performs consistently as well as the combined control group, but not better. Therefore, using information from several alternative locations rather than a single location improves prediction considerably, presumably because combining the locations increases the degree of overlap with the San Diego sample.

8.4 Least Squares versus Propensity Score Methods

In all cases, the least squares and propensity score estimates are fairly close, both in point estimates and in estimated precision. For example, in Table 4A, the average prediction error for controls using the three alternative locations is \$140 for least squares, with a standard error of \$120, and \$140 with a standard error of \$130 for the propensity score estimates. This suggests that the distribution of covariates is reasonably similar between San Diego and the other three locations and that the linearity assumptions in the least squares estimates do not affect the estimates too much. See Imbens, Rubin and Sacerdote (1999) for more discussion on this issue. More evidence supporting this interpretation is the fact that larger differences between least squares and propensity score estimates occur in comparisons between San Diego and single alternative locations, such as Arkansas. In general, however, least squares estimates perform fairly well.

8.5 Choosing Pre-training Variables

Least squares and propensity score estimates generally perform better than predictions based on simple differences in levels or gains, especially when prediction is based on a single alternative location. Consider, for example, the estimates based on the average outcome in Arkansas in Table 4A. This simple "level" prediction leads to an average prediction error of \$1,070, with a t-statistic of 6.7. The least squares estimate is much smaller at \$570 with a t-statistic of 2.5. Using the Baltimore data to predict the average gain in San Diego has a prediction error of \$270, with a t-statistic of 1.8. The least squares error again is much smaller at negative \$30.

However, results in Tables 4-7 compare estimates using no pre-training variables and estimates using the full set of available pre-training variables. Tables 8-11 examine these comparisons in more detail for controls, using least squares adjustment. In the first three rows of Tables 8-11, we estimate the average outcome in San Diego with the average outcome in the other locations, as in the "level" row in Tables 4-7. The first row uses the full data set. Results in the second row discard observations for men, women without children and women with children under six years old. In the third row, we adjust earnings using the county-level real earnings measure. In the fourth row, we use least squares methods to adjust for the four personal characteristics (indicator for high school degree, non-white, never married, and one child). In the next row, we also adjust for the level of earnings and indicators for positive earnings in each of the four quarters preceding randomization. The last row adds the two aggregate variables to the least squares regression.

Restricting the data set and adjusting earnings appears to be more important than the exact choice of pre-training variables. This general finding differs somewhat from previous work in the non-experimental evaluation literature, where it is typically very important to adjust for detailed labor market histories. See for example, Ashenfelter and Card (1985), Card and Sullivan (1988), Lalonde (1986), and Dehejia and Wahba (1998). One reason for this difference may be the nature of the non-experimental control groups. In many of these studies, e.g., Lalonde (1986), the control groups are constructed from data sets collected outside the scope of the training programs, such as the Current Population Survey (CPS) and the Panel Study of Income Dynamics (PSID). Therefore, it is not surprising that such comparison groups differ substantially from the target population for these programs. In our case, the control groups are from experiments on relatively similar target populations, namely individuals eligible for training programs. After removing any non-overlapping groups in these samples, the groups appear to be quite similar so that simple adjustments are adequate to remove remaining biases.

8.6 Aggregate Confounders

We use the aggregate variables in two ways. First, we include them as pre-training variables in the regression adjustment procedures. We do not include them in the specification of the propensity score because there is little overlap in their distributions across locations. In the propensity score methods, we use the aggregate variables in the within-block least squares adjustment only. Second, we use the real earnings measure to adjust earnings, making them more comparable across time and location. Both aspects appear to contribute to the ability to predict control outcomes.

9. Conclusion

Using training programs from three very different locations (Baltimore, Arkansas and Virginia), we attempt to predict the effect of a training program in San Diego. We find that we are able to predict the average outcomes for non-trainees fairly accurately, thus eliminating selection bias. Important for achieving this results is the inclusion of pre-training earnings, some personal characteristics, and some measures of aggregate differences across locations. Thus, we find that we require less detail on background information to predict control outcomes relative to other studies which use non-experimental control groups from large surveys such as the CPS. For example, Dehejia and Wahba (1998) require two years of earnings information to estimate training effects using controls from the PSID and CPS as substitutes for experimental controls. Heckman, Ichimura, Smith and Todd (1998) also find that adjustment for detailed pre-training differences (including earnings) is required for removing most of the average bias. Using control groups from other experimental evaluations appears to lead to more suitable comparison groups in our analyses, even though these experiments are conducted in very different locations and for different training programs. In contrast to these results for non-trainees, however, we cannot predict outcomes for trainees accurately.

Our interpretation for our differential success in predicting the outcomes of controls versus trainees across locations is as follows. Although the populations differ considerably across the four locations, enough detail exists on pre-training variables to predict earnings and employment gains within subpopulations *in the absence of training.* These findings are consistent with recent studies in the literature on non-experimental evaluations of training programs. This literature does not speak, however, to the difficulties in predicting outcomes for trainees using data from different implementations of the program. Here we find that none of the adjustment procedures succeeds in removing biases for trainees. The most plausible interpretation is that differences in the exact nature of the programs, such as the mix of components, are responsible for this failure. In that case, it appears to be difficult to make accurate predictions of the effects of future implementations of these training programs without more detail on the exact nature of the programs at the individual level.

The results support the importance of collecting detailed pre-training data, even in the context of randomized experiments, for extrapolation to other populations. They also suggest that more detail on the exact nature of the programs than is typically provided is necessary for experimental or non-experimental evaluations of existing programs to be useful guides for policy makers. How much and what type of detail about program structure and training components received is needed is a subject for future investigation.

Appendix A: Proofs

Proof of Lemma 1:

$$E[Y_i(1) - Y_i(0)|D_i = 0] = E[Y_i(1)|D_i = 0] - E[Y_i(0)|D_i = 0].$$

By Assumption 1 this is equal to

$$E[Y_i(1)|, T_i = 1, D_i = 0] - E[Y_i(0)|T_i = 0, D_i = 0],$$

which proves the first part. The same argument, conditional on $X_i = x$ proves the second part. □

Proof of Lemma 2:

We can write

$$E[Y_i(1) - Y_i(0)|D_i = 1] = E[Y_i(1)|D_i = 1] - E[Y_i(0)|D_i = 1].$$

By random allocation this is equal to

$$E[Y_i(1)|D_i = 0] - E[Y_i(0)|D_i = 0],$$

and by random assignment in the initial location this is equal to

$$E[Y_i(1)|, T_i = 1, D_i = 0] - E[Y_i(0)|T_i = 0, D_i = 0]$$

$$= E[Y_i|T_i = 1, D_i = 0] - E[Y_i|T_i = 0, D_i = 0],$$

which completes the proof. □

Proof of Lemma 3:

Using the same argument as in the proof for Lemma 2, we can show that

$$E[Y_i(1) - Y_i(0)|X_i = x, D_i = 1]$$

$$= E[Y_i|X_i = x, T_i = 1, D_i = 0] - E[Y_i|X_i = x, T_i = 0, D_i = 0].$$

Then

$$E[Y_i(1) - Y_i(0)|D_i = 1] = E\Big[E[Y_i(1) - Y_i(0)|X_i = x, D_i = 1]|D_i = 1\Big],$$

which finishes the proof. □

Proof of Lemma 4:

This proof is based on the same argument as the proof for Lemma 3:

$$E[Y_i(0)|X_i = x, D_i = 1] = E[Y_i|X_i = x, T_i = 0, D_i = 0],$$

and

$$E[Y_i(0)|D_i = 1] = E\Big[E[Y_i(0)|X_i = x, D_i = 1]|D_i = 1\Big],$$

which finishes the proof. □

Proof of Lemma 5

We can write

$$E[Y_i|D_i = 1, T_i = 1, X_i = x] = E[\sum_{t=1}^{K} Y_i(t) \cdot 1\{\tilde{T}_i = t\}|D_i = 1, T_i = 1, X_i = x]$$

$$= \sum_{t=1}^{K} E[Y_i(t)|\tilde{T}_i = t, D_i = 1, T_i = 1, X_i = x] \cdot Pr(T_i = t|D_i = 1, T_i = 1, X_i = x).$$

By unconfounded assignment,

$$E[Y_i(t)|\tilde{T}_i = t, D_i = 1, T_i = 1, X_i = x] = E[Y_i(t)|D_i = 1, T_i = 1, X_i = x],$$

and by unconfounded location this is equal to

$$E[Y_i(t)|D_i = 0, T_i = 1, X_i = x].$$

By conditional independence of assignment and location

$$Pr(T_i = t|D_i = 1, T_i = 1, X_i = x) = Pr(T_i = t|D_i = 0, T_i = 1, X_i = x).$$

Putting the two parts together:

$$\sum_{t=1}^{K} E[Y_i(t)|\tilde{T}_i = t, D_i = 1, T_i = 1, X_i = x] \cdot Pr(T_i = t|D_i = 1, T_i = 1, X_i = x)$$

$$= \sum_{t=1}^{K} E[Y_i(t)|D_i = 0, T_i = 1, X_i = x] \cdot Pr(T_i = t|D_i = 0, T_i = 1, X_i = x)$$

$$= \sum_{t=1}^{K} E[Y_i(t)|\tilde{T}_i = t, D_i = 0, T_i = 1, X_i = x] \cdot Pr(T_i = t|D_i = 0, T_i = 1, X_i = x)$$

$$= E[\sum_{t=1}^{K} Y_i(t) \cdot 1\{\tilde{T}_i = t\}|D_i = 0, T_i = 1, X_i = x]$$

$$= E[Y_i|D_i = 0, T_i = 1, X_i = x].$$

Averaging this over the distribution of X_i given $D_i = 1$ gives the desired result. □

Appendix B: Estimation Methods

In this appendix we describe the exact implementation of the estimators employed in the paper. In all the tables we use San Diego as the target program, and attempt to predict the average outcome for controls or trainees or the average program effect using data from all the other three programs combined, or from one of the other three programs. We report the difference between the actual average of the outcome variable or treatment effect directly estimated from San Diego data and the predicted value using outcomes from the other locations and pre-training variables from San Diego, as well as t-statistics for the null hypothesis that this difference is zero on average.

We discuss in detail the estimators in the first column of Table 4A. In this column the data used to predict the results for San Diego are from the other three locations combined. We have three sets of pre-training variables. The first, personal characteristics consists of four dummy variables, indicating whether the woman has a high school diploma, whether she is white, whether she has ever been married, and whether she has more than one child. The second set of pre-training variables consists of earnings for each of the four quarters prior to the job-training program, and four dummy variables, indicating positive earnings in each of those four quarters. The third set consists of the aggregate measures, the ratio of employment to population and real earnings per worker, both measured at the county level.

The four outcome variables are: (i) earnings in the first four quarters following the randomization, (ii) earnings in the fifth to eighth quarters following the randomization, (iii) an indicator for positive earnings in any of the first four quarters following the randomization, (iv) an indicator for positive earnings in any of the fifth to eighth quarter following randomization. In Tables 4A-C only the first outcome variable, earnings in the first four quarters following randomization is used. Tables 5A-C analyze earnings in the second year, Tables 6A-C analyze the employment indicator for the first post-randomization year and Tables 7A-C analyze the employment indicator for the second year.

The different rows in Table 4A refer to the different estimators for the average outcome for controls in San Diego. That is, in all entries in this table we attempt to estimate average earnings for control units in San Diego:

$$\hat{\mu}_{c,SD} = \sum_{d_i=1,t_i=0} y_i/N_{10},$$

where y_i, t_i, and d_i, are the outcome for unit i, her training status, and her location, respectively, and N_{dt} is the number of observations with $d_i = d$ and $t_i = t$. According to Table 1, $\hat{\mu}_{c,SD} = 1.77.$, or \$1,770.

First consider the "level" estimates. Here we predict the average outcome for controls in San Diego using the average outcome in the other three locations combined. The estimator for $\mu_{c,SD}$ in the "level" row is

$$\hat{\mu}^{level}_{c,SD} = \sum_{d_i=0,t_i=0} y_i/N_{00}.$$

In addition to the difference between the direct and indirect estimates $\hat{\mu}_{c,SD} - \hat{\mu}^{level}_{c,SD}$, we report the t-statistic testing the null hypothesis that the estimand, $\hat{\mu}^{level}_{c,SD}$, is identical to the average outcome for controls in San Diego, $\hat{\mu}_{c,SD}$.

The second row uses a single pre-training variable, the sum of earnings in the four quarters preceeding randomization to redefine the target outcome as the gain in earnings or employment status in San Diego. Let y_{i0} denote the pre-randomization value of the outcome for individual i, and let

$$\hat{\delta}_{c,SD} = \sum_{d_i=1,t_i=0} (y_i - y_{i0})/N_{10},$$

be the direct estimate of the average change in outcome for San Diego. This value can be calculated from Table 1, which indicates that $\hat{\delta}_{c,SD} = 0.22$. The estimator used in the second row of Table 4A is:

$$\hat{\delta}^{gain}_{c,SD} = \sum_{d_i=0,t_i=0} (y_i - y_{i0})/N_{00}.$$

Again we report the error, $\hat{\delta}_{c,SD} - \hat{\delta}^{gain}_{c,SD}$, and the t-statistic comparing the direct and indirect estimates.

The third row uses a linear regression to predict the change for a woman with pre-training characteristics x_i. First we regress the outcome, the change in earnings, on the characteristics using the observations with $d_i = 0$:

$$\hat{\beta} = \Big(\sum_{i|d_i=0,t_i=0} x_i x_i' \Big)^{-1} \Big(\sum_{i|d_i=0,t_i=0} x_i (y_i - y_{i0}) \Big).$$

Then the predicted value for the average change in San Diego is

$$\hat{\delta}^{ols}_{c,SD} = \sum_{d_i=1,t_i=0} x_i'\hat{\beta}/N_{10}.$$

The covariates in this regression include fourteen variables: four dummy variables for personal characteristics, the level of earnings in each of the four quarters preceding randomization, indicators for positive earnings in those four quarters, and the two aggregate measures. In the estimates based on a single comparison location we drop the aggregate measures that lead to perfect collinearity.

The fourth row reports results from propensity score procedures. First we divide the sample into the subsample of women with no pre-training earnings (approximately sixty percent of our sample) and the subsample of women with positive pre-training earnings in at least one of the four quarters.

For the first group, there are sixteen possible combinations of the four dummies describing personal characteristics. Twelve of the sixteen categories contain controls and trainees from all four locations. In the remaining four categories, women with one child and women with more than one child were combined to create two additional cells. The two cells are white women who never married, with and without a high-school diploma. For each San Diego control observation in these fourteen cells we predict $(y_i - y_{i0})$ by the average value of this change for observations with the exact same values for x in the other locations. That is, for all San Diego observations with $d_i = 1$, $t_i = 0$, $y_{i0} = 0$, and covariates x_i, we estimate the gain by least squares estimates using only the control units in the alternative locations with the exact same values for the individual level covariates, and as regressors the dummy for San Diego and the aggregate variables.

For the second group, women with positive earnings in at least one of the four quarters prior to randomization, we estimate a location score as a logistic function of a set of pre-training variables. Formally, we fit a logistic model for the indicator for San Diego on a set of regressors consisting of the four personal characteristics, four earnings levels, the four positive earnings indicators, and all squared terms and interactions, excluding only those that lead to perfect collinearity. Let $\hat{l}_i$ be the predicted value of the logistic regression for individuals in this group. We then calculate the quintiles of the distribution of $\hat{l}_i$ for control observations in San Diego with positive pre-training earnings ($d_i = 1$, $t_i = 0$, and $y_{i0} > 0$). Based on these quintiles we split the controls with positive pretraining earnings into five groups. Within these five blocks we estimate the average outcome in San Diego using least squares with a dummy variable for San Diego and the same pre-training variables as before (four personal characteristics, the four earnings levels, four positive earnings indicators, and two aggregate variables). Finally we weight these fourteen within-cell plus five within-block estimates by the proportion of San Diego controls in all nineteen groups to get an overall estimate of the average change for controls in San Diego.

The other columns in Table 4A are constructed exactly the same way using only the controls from one alternative location at a time, rather than the controls from all three alternative locations at the same time. Table 4B is constructed exactly the same way, using the trainees instead of the controls. Table 4C compares estimates for average treatment effects for San Diego. In the first two rows (level and gain) we estimate separately the average outcome for the control and trainee units, exactly as in Tables 4A-B, and then take the difference. For the OLS row in Table 4C, we predict the average treatment effect by predicting first for each unit in San Diego, trainee or control, the unit-specific treatment effect, and then average these over all units. This implies that the prediction error in this row is not necessarily identical to the difference in prediction errors in the OLS rows in Tables 4A and 4B, Two modifications are made for the propensity score estimates in Table 4C. First, the propensity, or location score $\hat{l}_i$ is estimated using both trainee and control units. Second, we use only one block for women with positive pre-training earnings, rather than five blocks, for predicting San Diego outcomes when using the Arkansas location. This is due to the smaller sample size in Arkansas. Tables 5-7 are constructed exactly the same way, using different outcome measures.

Table 1: Summary Statistics and T-statistics for Difference with San Diego

	San Diego (2603)		Arkansas (480)			Baltimore (2080)			Virginia (2753)		
	mean	s.d.	mean	s.d.	t-stat	mean	s.d.	t-stat	mean	s.d.	t-stat
Personal Char.											
High School Dipl.	0.56	(0.50)	0.50	(0.50)	[-2.4]	0.40	(0.49)	[-10.7]	0.43	(0.49)	[-9.5]
Nonwhite	0.69	(0.46)	0.83	(0.37)	[7.5]	0.70	(0.46)	[1.1]	0.67	(0.47)	[-1.6]
Never Married	0.26	(0.44)	0.35	(0.48)	[3.7]	0.38	(0.48)	[8.3]	0.28	(0.45)	[1.2]
One Child	0.48	(0.50)	0.42	(0.49)	[-2.5]	0.48	(0.50)	[0.1]	0.46	(0.50)	[-1.7]
More Than One Child	0.52	(0.50)	0.58	(0.49)	[2.5]	0.52	(0.50)	[-0.1]	0.54	(0.50)	[1.7]
Pre-training Earnings											
Earn Q-1	0.40	(1.02)	0.18	(0.51)	[-7.0]	0.42	(0.94)	[0.8]	0.31	(0.75)	[-3.4]
Earn Q-2	0.40	(1.03)	0.17	(0.52)	[-7.4]	0.42	(0.93)	[0.7]	0.32	(0.79)	[-3.2]
Earn Q-3	0.38	(0.98)	0.19	(0.56)	[-6.1]	0.44	(0.96)	[2.0]	0.30	(0.82)	[-3.3]
Earn Q-4	0.37	(0.99)	0.18	(0.52)	[-6.2]	0.42	(0.90)	[1.9]	0.28	(0.74)	[-4.0]
Earn Q-1 Pos.	0.27	(0.44)	0.17	(0.38)	[-5.0]	0.29	(0.46)	[2.0]	0.25	(0.44)	[-1.1]
Earn Q-2 Pos.	0.25	(0.43)	0.16	(0.37)	[-4.3]	0.31	(0.46)	[4.6]	0.20	(0.40)	[-3.7]
Earn Q-3 Pos.	0.25	(0.44)	0.14	(0.35)	[-6.1]	0.30	(0.46)	[3.2]	0.23	(0.42)	[-2.2]
Earn Q-4 Pos.	0.25	(0.44)	0.17	(0.38)	[-4.4]	0.29	(0.45)	[2.6]	0.24	(0.43)	[-1.1]
Aggregate Variables											
Emp./Pop.											
Pre-Randomization	0.53		0.54			0.48			0.49		
Year 1	0.55		0.55			0.48			0.50		
Year 2	0.56		0.57			0.49			0.52		
Real Inc. (Thousands)											
Pre-Randomization	17.8		16.2			18.3			16.6		
Year 1	18.1		16.8			17.4			17.5		
Year 2	18.7		17.0			17.6			17.8		
Post-training Earnings											
Year 1 Earn Train	2.08	(3.83)	0.84	(1.77)	[-7.8]	1.65	(3.14)	[-2.9]	1.60	(2.93)	[-3.8]
Year 1 Earn Contr	1.77	(3.95)	0.71	(1.83)	[-6.7]	1.75	(3.60)	[-0.1]	1.52	(2.95)	[-1.7]
Ave Treat Eff (s.e.)	0.30	(0.15)	0.13	(0.16)	[-0.8]	-0.10	(0.15)	[-1.9]	0.07	(0.12)	[-1.2]
Year 2 Earn Train	2.86	(5.45)	1.28	(2.58)	[-6.9]	2.54	(4.23)	[-1.6]	2.39	(3.95)	[-2.7]
Year 2 Earn Contr	2.28	(4.81)	1.10	(2.67)	[-5.5]	2.49	(4.35)	[1.1]	2.12	(3.92)	[-0.9]
Ave Treat Eff (s.e.)	0.57	(0.20)	0.18	(0.24)	[-1.3]	0.05	(0.19)	[-1.9]	0.27	(0.16)	[-1.2]
Post-training Employment											
Year 1 Emp Train	0.52	(0.50)	0.29	(0.45)	[-7.1]	0.47	(0.50)	[-2.4]	0.49	(0.50)	[-1.6]
Year 1 Emp Contr	0.40	(0.49)	0.27	(0.44)	[-4.3]	0.43	(0.49)	[1.2]	0.45	(0.50)	[2.5]
Ave Treat Eff (s.e.)	0.12	(0.02)	0.02	(0.04)	[-2.2]	0.05	(0.02)	[-2.5]	0.04	(0.02)	[-2.9]
Year 2 Emp Train	0.49	(0.50)	0.31	(0.46)	[-5.3]	0.49	(0.50)	[0.2]	0.52	(0.50)	[1.7]
Year 2 Emp Contr	0.40	(0.49)	0.27	(0.45)	[-4.1]	0.47	(0.50)	[3.3]	0.46	(0.50)	[2.9]
Ave Treat Eff (s.e.)	0.09	(0.02)	0.04	(0.04)	[-1.1]	0.03	(0.02)	[-2.1]	0.06	(0.02)	[-1.2]

Table 2: Summary Statistics and T-statistics for Difference with San Diego, Using Only Discarded Observations in Comparison Programs

	San Diego (2603)		Arkansas (647)			Baltimore (677)			Virginia (397)		
	mean	s.d.	mean	s.d.	t-stat	mean	s.d.	t-stat	mean	s.d.	t-stat
Personal Char.											
High School Dipl.	0.56	(0.50)	0.50	(0.50)	[-2.6]	0.54	(0.50)	[-0.6]	0.50	(0.50)	[-2.1]
Nonwhite	0.69	(0.46)	0.89	(0.31)	[13.2]	0.71	(0.45)	[1.4]	0.70	(0.46)	[0.6]
Never Married	0.26	(0.44)	0.60	(0.49)	[15.9]	0.49	(0.50)	[10.9]	0.54	(0.50)	[10.6]
One Child	0.48	(0.50)	0.38	(0.49)	[-4.7]	0.43	(0.50)	[-2.3]	0.41	(0.49)	[-2.6]
More Than One Child	0.52	(0.50)	0.57	(0.49)	[2.5]	0.48	(0.50)	[-1.8]	0.44	(0.50)	[-3.0]
Pre-training Earnings											
Earn Q-1	0.40	(1.02)	0.14	(0.46)	[-9.5]	0.28	(0.72)	[-3.4]	0.28	(0.68)	[-3.0]
Earn Q-2	0.40	(1.03)	0.15	(0.48)	[-9.1]	0.31	(0.80)	[-2.4]	0.26	(0.67)	[-3.8]
Earn Q-3	0.38	(0.98)	0.15	(0.50)	[-8.3]	0.33	(0.85)	[-1.3]	0.18	(0.62)	[-5.4]
Earn Q-4	0.37	(0.99)	0.16	(0.49)	[-7.8]	0.38	(0.90)	[0.3]	0.16	(0.59)	[-5.8]
Earn Q-1 Pos.	0.27	(0.44)	0.14	(0.35)	[-7.7]	0.25	(0.43)	[-1.2]	0.29	(0.45)	[0.8]
Earn Q-2 Pos.	0.25	(0.43)	0.14	(0.35)	[-6.7]	0.27	(0.44)	[1.2]	0.17	(0.38)	[-3.5]
Earn Q-3 Pos.	0.25	(0.44)	0.14	(0.35)	[-7.1]	0.26	(0.44)	[0.1]	0.18	(0.38)	[-3.7]
Earn Q-4 Pos.	0.25	(0.44)	0.15	(0.36)	[-6.0]	0.25	(0.43)	[-0.2]	0.24	(0.43)	[-0.5]
Post-training Earnings											
Year 1 Earn Train	2.08	(3.83)	0.74	(1.83)	[-9.2]	2.50	(3.97)	[1.8]	1.80	(3.50)	[-1.1]
Year 1 Earn Contr	1.77	(3.95)	0.59	(1.68)	[-8.2]	1.94	(3.81)	[0.7]	1.67	(2.65)	[-0.4]
Ave Treat Eff (s.e.)	0.30	(0.15)	0.15	(0.14)	[-0.7]	0.56	(0.30)	[0.8]	0.14	(0.31)	[-0.5]
Year 2 Earn Train	2.86	(5.45)	1.30	(2.77)	[-7.2]	3.98	(5.95)	[3.2]	2.45	(4.47)	[-1.3]
Year 2 Earn Contr	2.28	(4.81)	0.94	(2.49)	[-7.0]	3.00	(4.89)	[2.4]	2.17	(3.62)	[-0.4]
Ave Treat Eff (s.e.)	0.57	(0.20)	0.36	(0.21)	[-0.7]	0.98	(0.42)	[0.9]	0.29	(0.41)	[-0.6]
Post-training Employment											
Year 1 Emp Train	0.52	(0.50)	0.30	(0.46)	[-7.9]	0.60	(0.49)	[2.5]	0.48	(0.50)	[-1.3]
Year 1 Emp Contr	0.40	(0.49)	0.24	(0.43)	[-6.0]	0.51	(0.50)	[3.5]	0.54	(0.50)	[3.2]
Ave Treat Eff (s.e.)	0.12	(0.02)	0.06	(0.03)	[-1.6]	0.09	(0.04)	[-0.8]	-0.06	(0.05)	[-3.3]
Year 2 Emp Train	0.49	(0.50)	0.32	(0.47)	[-5.8]	0.64	(0.48)	[5.1]	0.53	(0.50)	[1.2]
Year 2 Emp Contr	0.40	(0.49)	0.26	(0.44)	[-4.9]	0.61	(0.49)	[7.0]	0.51	(0.50)	[2.6]
Ave Treat Eff (s.e.)	0.09	(0.02)	0.06	(0.04)	[-0.8]	0.03	(0.04)	[-1.4]	0.02	(0.05)	[-1.3]

Table 3: Summary Statistics and T-statistics for Difference with San Diego, Using Only Discarded Observations in Baltimore

	San Diego (2603)		All Discards (677)		Men (279)		Child<6 (362)		No Kids (36)	
	mean	s.d.	mean	t-stat	mean	t-stat	mean	t-stat	mean	t-stat
Personal Char.										
High School Dipl.	0.56	(0.50)	0.54	[-0.6]	0.39	[-5.3]	0.67	[4.5]	0.39	[-2.0]
Nonwhite	0.69	(0.46)	0.71	[1.4]	0.52	[-5.5]	0.86	[8.2]	0.83	[2.3]
Never Married	0.26	(0.44)	0.49	[10.9]	0.19	[-2.9]	0.72	[18.1]	0.56	[3.5]
One Child	0.48	(0.50)	0.43	[-2.3]	0.41	[2.3]	0.49	[0.4]	0	[-48.9]
More Than One Child	0.52	(0.50)	0.48	[-1.8]	0.51	[-0.5]	0.51	[-0.4]	0	[-53.2]
Pre-training Earnings										
Earn Q-1	0.40	(1.02)	0.28	[-3.4]	0.44	[0.7]	0.17	[-7.1]	0.26	[-1.5]
Earn Q-2	0.40	(1.03)	0.31	[-2.4]	0.50	[1.4]	0.17	[-7.2]	0.32	[-0.7]
Earn Q-3	0.38	(0.98)	0.33	[-1.3]	0.48	[1.5]	0.22	[-4.0]	0.29	[-0.8]
Earn Q-4	0.37	(0.99)	0.38	[0.3]	0.62	[3.4]	0.20	[-4.8]	0.35	[-0.2]
Earn Q-1 Pos.	0.27	(0.44)	0.25	[-1.2]	0.32	[1.8]	0.19	[-3.7]	0.28	[0.1]
Earn Q-2 Pos.	0.25	(0.43)	0.27	[1.2]	0.35	[3.6]	0.20	[-1.8]	0.25	[0.1]
Earn Q-3 Pos.	0.25	(0.44)	0.26	[0.1]	0.35	[3.2]	0.19	[-2.7]	0.17	[-1.4]
Earn Q-4 Pos.	0.25	(0.44)	0.25	[-0.2]	0.34	[2.8]	0.19	[-3.1]	0.25	[-0.1]
Post-training Earnings										
Year 1 Earn Train	2.08	(3.83)	2.50	[1.8]	3.72	[3.8]	1.75	[-1.3]	0.97	[-2.6]
Year 1 Earn Contr	1.77	(3.95)	1.94	[0.7]	3.09	[3.0]	1.11	[-3.4]	0.83	[-1.6]
Ave Treat Eff (s.e.)	0.30	(0.15)	0.56	[0.8]	0.64	[0.5]	0.64	[1.1]	0.14	[-0.2]
Year 2 Earn Train	2.86	(5.45)	3.98	[3.2]	5.96	[4.7]	2.86	[-0.0]	0.81	[-5.9]
Year 2 Earn Contr	2.28	(4.81)	3.00	[2.4]	4.24	[3.6]	2.20	[-0.3]	0.73	[-3.6]
Ave Treat Eff (s.e.)	0.57	(0.20)	0.98	[0.9]	1.72	[1.3]	0.65	[0.2]	0.09	[-0.9]
Post-training Employment										
Year 1 Emp Train	0.52	(0.50)	0.60	[2.5]	0.71	[4.4]	0.54	[0.4]	0.38	[-1.3]
Year 1 Emp Contr	0.40	(0.49)	0.51	[3.5]	0.59	[4.2]	0.47	[1.6]	0.27	[-1.1]
Ave Treat Eff (s.e.)	0.12	(0.02)	0.09	[-0.8]	0.12	[-0.0]	0.07	[-0.9]	0.11	[-0.0]
Year 2 Emp Train	0.49	(0.50)	0.64	[5.1]	0.70	[5.1]	0.60	[2.8]	0.57	[0.8]
Year 2 Emp Contr	0.40	(0.49)	0.61	[7.0]	0.61	[5.0]	0.62	[5.7]	0.40	[0.0]
Ave Treat Eff (s.e.)	0.09	(0.02)	0.03	[-1.4]	0.08	[-0.1]	-0.02	[-2.0]	0.17	[0.5]

Table 4.A: Adjusted Differences between San Diego and Other Locations in First Year Post-training Earnings for Controls in Thousands of Dollars

	SD/All		SD/AR		SD/MD		SD/VA	
	dif	t-stat	dif	t-stat	dif	t-stat	dif	t-stat
Level	0.24	[1.8]	1.07	[6.7]	0.03	[0.2]	0.25	[1.7]
Gain	0.13	[0.9]	0.16	[1.0]	0.27	[1.8]	-0.06	[-0.4]
OLS(Agg., Adj. Earn.)	0.14	[1.2]	0.57	[2.5]	-0.03	[-0.2]	0.02	[0.1]
PS(Agg., Adj. Earn.)	0.14	[1.1]	0.77	[2.2]	0.02	[0.1]	0.07	[0.4]

Table 4.B: Adjusted Differences between San Diego and Other Locations in First Year Post-training Earnings for Trainees in Thousands of Dollars

	SD/All		SD/AR		SD/MD		SD/VA	
	dif	t-stat	dif	t-stat	dif	t-stat	dif	t-stat
Level	0.52	[4.4]	1.24	[7.8]	0.42	[2.9]	0.48	[3.8]
Gain	0.28	[2.4]	0.48	[2.5]	0.47	[3.3]	0.16	[1.2]
OLS(Agg., Adj. Earn.)	0.29	[2.9]	0.71	[3.1]	0.21	[1.5]	0.16	[1.3]
PS(Agg., Adj. Earn.)	0.32	[2.9]	0.68	[2.0]	0.43	[2.6]	0.21	[1.7]

Table 4.C: Adjusted Differences between San Diego and Other Locations in Experimental and Predicted Average Treatment Effect for First Year Post-training Earnings in Thousands of Dollars

	SD/All		SD/AR		SD/MD		SD/VA	
	dif	t-stat	dif	t-stat	dif	t-stat	dif	t-stat
Level	0.28	[1.6]	0.17	[0.8]	0.40	[1.9]	0.23	[1.2]
Gain	0.15	[0.9]	0.31	[1.3]	0.19	[0.9]	0.21	[1.1]
OLS(Agg., Adj. Earn.)	0.18	[1.0]	0.13	[0.5]	0.30	[1.5]	0.10	[0.5]
PS(Agg., Adj. Earn.)	0.16	[0.9]	0.05	[0.2]	0.20	[0.9]	0.25	[1.2]

Table 5.A: Adjusted Differences between San Diego and Other Locations in Second Year Post-training Earnings for Controls in Thousands of Dollars

	SD/All		SD/AR		SD/MD		SD/VA	
	dif	t-stat	dif	t-stat	dif	t-stat	dif	t-stat
Level	0.10	[0.6]	1.18	[5.5]	-0.21	[-1.1]	0.17	[0.9]
Gain	-0.01	[-0.1]	0.28	[1.3]	0.04	[0.2]	-0.15	[-0.7]
OLS(Agg., Adj. Earn.)	0.02	[0.1]	0.57	[1.9]	-0.33	[-1.8]	-0.09	[-0.4]
PS(Agg., Adj. Earn.)	0.05	[0.3]	0.94	[2.2]	-0.25	[-1.2]	-0.13	[-0.6]

Table 5.B: Adjusted Differences between San Diego and Other Locations in Second Year Post-training Earnings for Trainees in Thousands of Dollars

	SD/All		SD/AR		SD/MD		SD/VA	
	dif	t-stat	dif	t-stat	dif	t-stat	dif	t-stat
Level	0.50	[3.0]	1.57	[6.9]	0.32	[1.6]	0.47	[2.7]
Gain	0.26	[1.7]	0.82	[3.6]	0.36	[1.9]	0.14	[0.9]
OLS(Agg., Adj. Earn.)	0.18	[1.3]	0.87	[2.6]	-0.13	[-0.7]	0.02	[0.1]
PS(Agg., Adj. Earn.)	0.25	[1.6]	1.03	[2.1]	0.08	[0.3]	0.13	[0.7]

Table 5.C: Adjusted Differences between San Diego and Other Locations in Experimental and Predicted Average Treatment Effect for Second Year Post-training Earnings in Thousands of Dollars

	SD/All		SD/AR		SD/MD		SD/VA	
	dif	t-stat	dif	t-stat	dif	t-stat	dif	t-stat
Level	0.40	[1.7]	0.39	[1.3]	0.52	[1.9]	0.30	[1.2]
Gain	0.27	[1.2]	0.54	[1.7]	0.32	[1.2]	0.28	[1.1]
OLS(Agg., Adj. Earn.)	0.21	[0.9]	0.33	[0.9]	0.26	[0.9]	0.13	[0.4]
PS(Agg., Adj. Earn.)	0.15	[0.6]	0.22	[0.6]	0.25	[0.8]	0.28	[1.0]

Table 6.A: Adjusted Differences between San Diego and Other Locations in First Year Post-training Employment for Controls

	SD/All		SD/AR		SD/MD		SD/VA	
	dif	t-stat	dif	t-stat	dif	t-stat	dif	t-stat
Level	-0.018	[-1.0]	0.134	[4.3]	-0.025	[-1.2]	-0.053	[-2.5]
Gain	-0.011	[-0.6]	0.012	[0.4]	0.038	[1.8]	-0.075	[-3.2]
OLS(Agg., Adj. Earn.)	-0.014	[-0.9]	0.070	[2.4]	-0.002	[-0.1]	-0.098	[-4.4]
PS(Agg., Adj. Earn.)	-0.025	[-1.5]	0.105	[2.4]	0.011	[0.6]	-0.081	[-3.8]

Table 6.B: Adjusted Differences between San Diego and Other Locations in First Year Post-training Employment for Trainees

	SD/All		SD/AR		SD/MD		SD/VA	
	dif	t-stat	dif	t-stat	dif	t-stat	dif	t-stat
Level	0.051	[3.1]	0.234	[7.1]	0.050	[2.4]	0.030	[1.6]
Gain	0.040	[2.2]	0.095	[2.7]	0.090	[3.9]	0.006	[0.3]
OLS(Agg., Adj. Earn.)	0.041	[2.6]	0.171	[5.2]	0.050	[2.5]	-0.005	[-0.3]
PS(Agg., Adj. Earn.)	0.036	[2.2]	0.171	[3.9]	0.071	[3.1]	0.010	[0.6]

Table 6.C: Adjusted Differences between San Diego and Other Locations in Experimental and Predicted Average Treatment Effect for First Year Post-training Employment

	SD/All		SD/AR		SD/MD		SD/VA	
	dif	t-stat	dif	t-stat	dif	t-stat	dif	t-stat
Level	0.069	[2.9]	0.099	[2.2]	0.074	[2.5]	0.083	[2.9]
Gain	0.051	[2.0]	0.083	[1.9]	0.052	[1.6]	0.081	[2.7]
OLS(Agg., Adj. Earn.)	0.054	[2.2]	0.117	[2.5]	0.046	[1.6]	0.095	[3.1]
PS(Agg., Adj. Earn.)	0.063	[2.7]	0.130	[2.9]	0.062	[2.0]	0.097	[3.4]

Table 7.A: Adjusted Differences between San Diego and Other Locations in Second Year Post-training Employment for Controls

	SD/All		SD/AR		SD/MD		SD/VA	
	dif	t-stat	dif	t-stat	dif	t-stat	dif	t-stat
Level	-0.043	[-2.5]	0.127	[4.1]	-0.067	[-3.3]	-0.063	[-2.9]
Gain	-0.036	[-1.8]	0.005	[0.2]	-0.005	[-0.2]	-0.085	[-3.4]
OLS(Agg., Adj. Earn.)	-0.027	[-1.6]	0.084	[2.7]	-0.047	[-2.4]	-0.076	[-3.3]
PS(Agg., Adj. Earn.)	-0.041	[-2.3]	0.058	[1.3]	-0.034	[-1.6]	-0.096	[-4.3]

Table 7.B: Adjusted Differences between San Diego and Other Locations in Second Year Post-training Employment for Trainees

	SD/All		SD/AR		SD/MD		SD/VA	
	dif	t-stat	dif	t-stat	dif	t-stat	dif	t-stat
Level	-0.006	[-0.4]	0.177	[5.3]	-0.005	[-0.2]	-0.030	[-1.7]
Gain	-0.018	[-0.9]	0.039	[1.1]	0.035	[1.4]	-0.053	[-2.5]
OLS(Agg., Adj. Earn.)	-0.010	[-0.6]	0.123	[3.6]	-0.012	[-0.6]	-0.057	[-3.0]
PS(Agg., Adj. Earn.)	-0.016	[-1.0]	0.140	[3.0]	-0.003	[-0.1]	-0.045	[-2.4]

Table 7.C: Adjusted Differences between San Diego and Other Locations in Experimental and Predicted Average Treatment Effect for Second Year Post-training Employment

	SD/All		SD/AR		SD/MD		SD/VA	
	dif	t-stat	dif	t-stat	dif	t-stat	dif	t-stat
Level	0.037	[1.5]	0.050	[1.1]	0.062	[2.1]	0.033	[1.2]
Gain	0.019	[0.7]	0.034	[0.7]	0.040	[1.2]	0.032	[1.0]
OLS(Agg., Adj. Earn.)	0.018	[0.7]	0.040	[0.8]	0.030	[1.0]	0.024	[0.8]
PS(Agg., Adj. Earn.)	0.015	[0.6]	0.047	[0.9]	0.028	[0.9]	0.042	[1.4]

Table 8: Least Squares Adjusted Differences between San Diego and Other Locations in First Year Post-training Earnings for Controls, Using Different Comparison Groups and Control Variables

		SD/All		SD/AR		SD/MD		SD/VA	
Control Var.	Data	dif	t-stat	dif	t-stat	dif	t-stat	dif	t-stat
None	All	0.20	[1.8]	1.05	[8.8]	-0.11	[-0.8]	0.14	[1.1]
None	Restr.	0.24	[1.8]	1.07	[6.7]	0.03	[0.2]	0.25	[1.7]
None	Restr. (Adj. Earn.)	0.18	[1.4]	1.01	[6.1]	-0.04	[-0.3]	0.20	[1.4]
Pers.	Restr. (Adj. Earn.)	0.04	[0.3]	0.93	[3.6]	-0.24	[-1.5]	0.08	[0.5]
Pers., Earn.	Restr. (Adj. Earn.)	0.05	[0.4]	0.57	[2.5]	-0.03	[-0.2]	0.02	[0.1]
Pers., Earn., Aggr.	Restr. (Adj. Earn.)	0.14	[1.2]	0.57	[2.5]	-0.03	[-0.2]	0.02	[0.1]

Table 9: Least Squares Adjusted Differences between San Diego and Other Locations in Second Year Post-training Earnings for Controls, Using Different Comparison Groups and Control Variables

		SD/All		SD/AR		SD/MD		SD/VA	
Control Var.	Data	dif	t-stat	dif	t-stat	dif	t-stat	dif	t-stat
None	All	0.12	[0.9]	1.25	[7.8]	-0.35	[-2.1]	0.14	[0.8]
None	Restr.	0.10	[0.6]	1.18	[5.5]	-0.21	[-1.2]	0.17	[0.9]
None	Restr. (Adj. Earn.)	-0.04	[-0.2]	1.06	[4.6]	-0.37	[-1.9]	0.06	[0.3]
Pers.	Restr. (Adj. Earn.)	-0.19	[-1.2]	0.94	[3.0]	-0.58	[-3.0]	-0.09	[-0.5]
Pers., Earn.	Restr. (Adj. Earn.)	-0.15	[-1.0]	0.61	[2.1]	-0.33	[-1.8]	-0.12	[-0.7]
Pers., Earn., Aggr.	Restr. (Adj. Earn.)	0.02	[0.1]	0.57	[1.9]	-0.33	[-1.8]	-0.09	[-0.4]

Table 10: Least Squares Adjusted Differences between San Diego and Other Locations in First Year Post-training Employment for Controls, Using Different Comparison Groups and Control Variables

		SD/All		SD/AR		SD/MD		SD/VA	
Control Var.	Data	dif	t-stat	dif	t-stat	dif	t-stat	dif	t-stat
None	All	-0.011	[-0.7]	0.154	[7.0]	-0.041	[-2.2]	-0.062	[-3.1]
None	Restr.	-0.018	[-1.0]	0.134	[4.3]	-0.025	[-1.2]	-0.053	[-2.5]
None	Restr. (Adj. Earn.)	-0.018	[-1.0]	0.134	[4.3]	-0.025	[-1.2]	-0.053	[-2.5]
Pers.	Restr. (Adj. Earn.)	-0.035	[-2.0]	0.128	[3.9]	-0.041	[-2.0]	-0.072	[-3.4]
Pers., Earn.	Restr. (Adj. Earn.)	-0.025	[-1.6]	0.072	[2.5]	-0.002	[-0.1]	-0.077	[-4.0]
Pers., Earn., Aggr.	Restr. (Adj. Earn.)	-0.014	[-0.9]	0.070	[2.4]	-0.002	[-0.1]	-0.098	[-4.4]

Table 11: Least Squares Adjusted Differences between San Diego and Other Locations in Second Year Post-training Employment for Controls, Using Different Comparison Groups and Control Variables

		SD/All		SD/AR		SD/MD		SD/VA	
Control Var.	Data	dif	t-stat	dif	t-stat	dif	t-stat	dif	t-stat
None	All	-0.042	[-2.8]	0.137	[6.2]	-0.097	[-5.3]	-0.067	[-3.4]
None	Restr.	-0.043	[-2.5]	0.127	[4.1]	-0.067	[-3.3]	-0.063	[-2.9]
None	Restr. (Adj. Earn.)	-0.043	[-2.5]	0.127	[4.1]	-0.067	[-3.3]	-0.063	[-2.9]
Pers.	Restr. (Adj. Earn.)	-0.056	[-3.2]	0.124	[3.7]	-0.080	[-3.9]	-0.076	[-3.5]
Pers., Earn.	Restr. (Adj. Earn.)	-0.047	[-2.9]	0.084	[2.7]	-0.047	[-2.4]	-0.081	[-4.0]
Pers., Earn., Aggr.	Restr. (Adj. Earn.)	-0.027	[-1.6]	0.084	[2.7]	-0.047	[-2.4]	-0.076	[-3.3]

References

Angrist, J., (1998), "Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants", *Econometrica*, Vol. 66, No. 2, 249–288.

Ashenfelter, O., and D. Card, (1985), "Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs", *Review of Economics and Statistics*, 67, 648–660.

Card, D., and Sullivan, (1988), "Measuring the Effect of Subsidized Training Programs on Movements In and Out of Employment", *Econometrica*, vol. 56, no. 3 497–530.

Cook, T., and D. Campbell, (1979), *Quasi-experimentation: Design and Analysis Issues for Field Settings*, Rand McNally, Chicago.

Dehejia, R., (1997) "A Decision-theoretic Approach to Program Evaluation", Chapter 2, Ph.D. Dissertation, Department of Economics, Harvard University.

Dehejia, R., and S. Wahba, (1998), "Causal Effects in Non–experimental Studies: Re–evaluating the Evaluation of Training Programs" NBER working paper #6829.

Fraker, T., and R. Maynard, (1987), "The Adequacy of Comparison Group Designs for Evaluations of Employment-Related Programs", *Journal of Human Resources*, Vol. 22, No. 2, p 194–227.

Friedlander, D., and J. Gueron, (1992) "Are High-Cost Services More Effective than Low-Cost Services?",

Friedlander, D., and P. Robins, (1995), "Evaluating Program Evaluations: New Evidence on Commonly Used Nonexperimental Methods", *American Economic Review*, Vol. 85, p 923–937.

Greenberg, D., and M. Wiseman, (1992) "What Did the OBRA Demonstrations Do?" in *Evaluating Welfare and Training Programs*, ed. by Irwin Garfinkel and Charles Manski, Cambridge, MA: Harvard University Press.

GUERON, J., (1990), "Work and Welfare: Lessons on Employment Programs", *Journal of Economic Perspectives*, Vol. 4, No. 1, 79-98.

GUERON, J.,AND E. PAULY (1991), *From Welfare to Work*, Russell Sage Foundation, New York.

HECKMAN, J., AND R. ROBB, (1984), "Alternative Methods for Evaluating the Impact of Interventions," in Heckman and Singer (eds.), *Longitudinal Analysis of Labor Market Data*, Cambridge, Cambridge University Press.

HECKMAN, J., AND J. HOTZ, (1989), "Alternative Methods for Evaluating the Impact of Training Programs", (with discussion), *Journal of the American Statistical Association.*

HECKMAN, J., H. ICHIMURA, J. SMITH, AND P. TODD, (1998), "Characterizing Selection Bias Using Experimental Data", *Econometrica*, Vol. 66, No. 5, p 1017–1098.

HOTZ, V. J. (1992), "Recent Experience in Designing Evaluations of Social Programs: The Case of the National JTPA Study", in *Evaluating Welfare and Training Programs*, ed. by Irwin Garfinkel and Charles Manski, Cambridge, MA: Harvard University Press, 76-114.

HOTZ, V. J., C. MULLIN, AND S. SANDERS (1997), "Bounding Causal Effects Using Data from a Contaminated Natural Experiment: Analyzing the Effects of Teenage Childbearing", *Review of Economic Studies*, Vol. 64, No. 4, 576-603.

IMBENS, G., D. RUBIN, AND B. SACERDOTE, (1999), "Estimating the Effect of Unearned Income on Labor Supply, Earnings, Savings and Consumption: Evidence from a Survey of Lottery Players", Department of Economics, UCLA.

LALONDE, R. (1986), "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," *American Economic Review*, Vol. 76, No. 4, p 604-620.

LECHNER, M, (1998), "Earnings and Employment Effects of Continuous Off-the-job Training in East Germany After Unification," *Journal of Business and Economic Statistics.*

Manski, C., (1997), "The Mixing Problem in Programme Evaluation", *Review of Economic Studies*, Vol 64, No. 4, 537-554.

Meyer, B., (1995) "Natural and Quasi-Experiments in Economics" *Journal of Business and Economic Statistics*, Vol 13, No 2, 151-161.

Rosenbaum, P., (1987), "The role of a second control group in an observational study", *Statistical Science*, (with discussion), Vol 2., No. 3, 292–316.

Rosenbaum, P., and D. Rubin, (1983), "The central role of the propensity score in observational studies for causal effects", *Biometrika*, 70, 1, 41–55.

Rosenbaum, P., and D. Rubin, (1984), "Reducing bias in observational studies using subclassification on the propensity score," *Journal of the American Statistical Association*, Vol 79, 516–524.

Rubin, D. (1974), "Estimating Causal Effects of Treatments in Randomized and Non-randomized Studies," *Journal of Educational Psychology*, 66, 688-701.

Rubin, D. B., (1976), "Inference and Missing Data," *Biometrika* 63, 581–592.

Rubin, D. B., (1977), "Assignment to a Treatment Group on the Basis of a Covariate", *Journal of Educational Statistics*, 2, 1-26.

Rubin, D., (1978), "Bayesian inference for causal effects: The Role of Randomization", *Annals of Statistics*, 6:34–58.