Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Econometrics 125 (2005) 241–270
JOURNAL OF Econometrics
www.elsevier.com/locate/econbase

Predicting the efficacy of future training programs using past experiences at other locations

V. Joseph Hotz[a,*], Guido W. Imbens[b], Julie H. Mortimer[c]

[a] *Department of Economics, Bunche Hall, University of California at Los Angeles, 405 Hilgard Avenue, Los Angeles, CA 90095, USA*

[b] *Department of Economics, and Department of Agricultural and Resource Economics, 661 Evans Hall #3880, University of California at Berkeley, Berkeley, CA 94720-3880, USA*

[c] *Department of Economics, Littauer Center, Harvard University, Cambridge, MA 02138, USA*

Available online 2 July 2004

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 or nature of their components, or in their efficacy across different sub-populations. The first problem is similar to the problem of non-experimental evaluations. The ability to adjust for population differences typically depends on the availability of characteristics of the two populations and the extent of overlap in their distributions. The ability to adjust for differences in the programs themselves may require more detailed data on the exact treatments received by individuals than are typically available. This problem has received less attention, although it is equally important for the prediction of the efficacy of new programs. To investigate the empirical importance of these issues, we compare four experimental Work INcentive demonstration programs implemented in the mid-1980s in different parts of the U.S. We find that adjusting for pre-training earnings and individual characteristics removes many of the differences between control units that have some previous employment experience. Since the control treatment is the same in all locations, namely embargo from the program services, this suggests that differences in populations served can be adjusted for in this sub-population. We also find that adjusting for individual characteristics is more successful at removing differences between control group members in different locations that

*Corresponding author.

E-mail addresses: hotz@ucla.edu (V. Joseph Hotz), imbens@econ.berkeley.edu (G.W. Imbens), mortimer@harvard.edu (J.H. Mortimer).

0304-4076/$ - see front matter
doi:10.1016/j.jeconom.2004.04.009

have some employment experience in the preceding four quarters than for control group members with no previous work experience. Perhaps more surprisingly, our ability to predict the outcomes of trainees after adjusting for individual characteristics is similar. We surmise that differences in treatment components across training programs are not sufficiently large to lead to substantial differences in our ability to predict trainees' post-training earnings for many of the locations in this study. However, in the sub-population with no previous work experience there is some evidence that unobserved heterogeneity leads to difficulties in our ability to predict outcomes across locations for controls.

Keywords: Efficacy; Training programs; Heterogeneity; Prediction

1. Introduction

Consider a government contemplating the implementation of a job training (or other social assistance) program. The decision to implement the program depends on the assessment of its likely effectiveness, often based on data from a similar program implemented in an earlier time period and/or another locality. For example, the U.S. federal government's job training programs, since the passage of the Job Training Partnership Act (JTPA) in 1982, are administered at the local level. Thus, local policy makers may wish to evaluate differences in the effectiveness of different versions of this program to implement a version appropriate for their locality. The recent federal reforms to the U.S. welfare system also have encouraged the further development of state and local program diversity in programs and services, both in the services clients receive and in the target populations. The state and local authorities that administer these new programs seek to use information from other programs, conducted in different time-periods and/or locations, to assess the likely impacts and cost-effectiveness of such programs.

Two distinct steps are necessary 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 sub-populations, 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, 1985; Heckman and Robb, 1984; Lalonde, 1986; Card and Sullivan, 1988; Heckman and Hotz, 1989; Imbens and Angrist, 1994; Friedlander and Robins, 1995; Heckman et al., 1998; Hahn, 1998; Dehejia and Wahba, 1998, 1999; Angrist and Krueger, 1999; Abadie and Imbens, 2004; Hirano et al., 2003; Imbens, 2004).

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 the external validity of evaluations.[1] 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.

At least three distinct reasons may exist for differences in the average effect of treatment 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 the outcomes of interest and in particular with program efficacy, average program effects may be found to differ for the two programs. The issues in this case are similar to those encountered in non-experimental evaluations of existing programs, e.g., the need for sufficient detail in pre-treatment variables and sufficient overlap in their distribution in the two treatment states or locations. The second reason that treatment effects may differ across programs is that the programs, even if nominally the same, are heterogeneous in their components. For example, one job-training program may stress classroom training, whereas another may emphasize job search assistance. Alternatively, training programs may differ qualitatively, e.g., some are better run or organized than others, even though they target the same population and nominally consist of the same treatment components. A third reason that average program effects may differ is the presence of interactions between individuals enrolled in different sized programs. For example, a program enrolling 100% of an eligible population may have a different effect than the same program implemented for a small percentage of the eligible population. In this paper, we focus on the first two reasons for differences in average treatment effects across programs.[2]

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. Similar to Lalonde (1986), we use the estimates from randomized experiments to assess the performance of our proposed methods. We focus on two comparisons. First, we compare the average outcomes for controls in one location with the average outcomes for controls in the other locations after adjusting for various sets of background factors. Given sufficient adjustment for background characteristics and the fact that their members are excluded from all training services, control groups should be comparable across sites. Specifically, after adjustment control group outcomes will be comparable regardless of any potential differences in program activities or program management. Thus, the success of this adjustment should address the first potential source of differences in treatment effects, namely the lack of comparability in the populations across programs and/or localities. Second, we compare the

[1] Also see Cook and Campbell (1979).

[2] While potentially important, we do not address the third complication. 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. If this assumption is violated, even experimental evaluations in a random sample need not lead to valid inferences regarding the efficacy of the existing program extended to the larger population.

average outcomes for trainees in one location with the average outcomes for trainees in the other locations. Conditional on the success of controlling for background differences across the two control groups, the comparison of comparably adjusted average outcomes for trainees should isolate the effect of treatment heterogeneity across programs. In contrast, if we cannot control for background differences between the control groups, we do not know whether differences in adjusted outcomes between treatment groups are attributable to differences in the training programs, or to differences between populations in unobserved characteristics that interact with the treatment (although level effects could be eliminated by subtracting average differences in outcomes for controls).

An important part of our empirical analyses is assessing the effectiveness of alternative sets of pre-training and aggregate variables for eliminating biases due to differences in the populations. A growing literature investigates whether observational control groups suffice for unbiased program evaluation once detail on pre-training labor earnings is available, at least within some demographic groups (e.g., Dehejia and Wahba, 1999; Friedlander and Robins, 1995; Heckman et al., 1998). Often such observational control groups come from public use surveys (e.g., Lalonde, 1986) or eligible non-participants from the same experiment (Heckman et al., 1998, Smith and Todd, 2001, 2004). In an influential paper, Lalonde (1986) showed that many conventional econometric methods were unable to recover estimates based on experimental evaluations.[3] Recently, Dehejia and Wahba (1999), 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.[4] Crucial to this success was the availability of sufficiently detailed earnings histories and background characteristics and the use of flexible adjustment methods.

Our non-experimental comparison groups are taken from experimental control groups in a variety of other locations, and they therefore consist of groups similarly disadvantaged and motivated. For this reason they may be subject to less severe, or at least different, biases than control groups from public use surveys or eligible non-participants. We also distinguish between two important sub-populations in our analyses: those with and without previous work experience in the previous year (which is as far back as we have information for). For a variety of reasons, previous workforce experience tends to be an important predictor of an individual's subsequent labor market success and measures of this experience are often controlled for in non-experimental evaluations of the effects of training programs. For this reason, we conduct our analyses separately for those with and without workforce experience in the four quarters prior to random assignment to assess the importance of accounting for this particular set of background

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

[4] See Angrist (1998), Heckman et al. (1998), Lechner (1999), Angrist and Krueger (1999) for discussions and economic applications of matching and propensity score methods.

characteristics when extrapolating the results of a program to a new implementation. In practice, we find that adjustments for individual characteristics work better for the sub-populations in each location consisting of people with previous work experience, regardless of program participation. The relative difficulty in predicting outcomes for the never-employed sub-population holds for both controls and trainees.

Our paper is most closely related to Friedlander and Robins (1995) who use the same data to assess the use of alternative comparison groups derived from non-experimental data and of specification tests of the validity of several estimators. They also construct non-experimental control groups from experimental control groups in other locations. Our paper provides several important extensions relative to the results reported by Friedlander and Robins. First, we restrict comparisons to the sub-populations in each location for which we have sufficient overlap. We provide evidence that average outcomes for different sub-populations may differ significantly, e.g., average outcomes are quite different for men than for women with children under 6 years of age. Without restricting our focus to the sub-populations with sufficient overlap, it is difficult to interpret predictions about average treatment effects. Second, we also distinguish between two important sub-populations in each location—people with and without previous work experience. We find that this distinction is critical for our ability to predict the results of a program using the results of a different implementation. Third, instead of constructing non-experimental control groups from other locations and comparing these to groups of trainees in the location of interest, we focus separately on both groups of controls and groups of trainees. This allows us to separately identify differences in treatment effects that are due to population differences from those that are due to program heterogeneity. Essentially, Friedlander and Robin's paper speaks to the first challenge for predicting treatment effects on the basis of previous experience (adjusting for population differences), whereas we focus not only on population differences across locations, but also on identifying differences in treatment effects that may arise from program heterogeneity. This distinction is important for policy makers when designing such programs. Fourth, we attempt to adjust for differences in macro-economic conditions across locations by incorporating aggregate covariates that proxy for such differences, e.g., we adjust for the ratio of real earnings per worker in different locations. Finally, we pool data from all other locations when predicting outcomes for a particular location of interest and use matching methods to estimate average treatment effects. Although Friedlander and Robins also make use of pooled data for some estimates, they use only small comparison groups for many of their estimates, and they use adjustment methods different from the matching approach we use.

In the next section we set up the inferential problem in the potential outcome notation for causal modeling. Complications arising from the presence of macro-effects are discussed in Section 3. An application of these ideas to four Work INcentive (WIN) training programs is presented in Sections 6–8. 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 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 or control, and $T_i = 1$ corresponding to training), and a set of covariates or pretreatment variables, X_i. The realized (observed) 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:

$$\tau_1 = E[Y_i(1) - Y_i(0) \mid D_i = 1].$$

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 in the $D_i = 0$ location 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 $D_i = 1$ location 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).

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

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.[5]

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. 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

[5] We can relax this to assuming that the selection into the treatment is based on covariates, but in order to focus on the main issues in the paper, and given the randomized assignment in the application, we do not consider this generalization here.

[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 the focus is on obtaining an average treatment effect for a population of programs.

(individuals) choose where to live. The random location assumption can be relaxed exploiting the presence of pre-treatment variables:

Assumption 2 (Unconfounded location). Location of program is unconfounded given pre-treatment variables X_i if

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

We also refer to this as the "no macro-effects" assumption. It relies on systematic differences between the locations that are correlated with the outcome of interest being captured by observed covariates. Note that in the application we include lagged outcomes (earnings and employment indicators) in the set of covariates, which is equivalent to working in differences rather than levels. Since the no-macro-effects assumption is much more plausible in differences than in levels, this increases the credibility of the analysis.

In addition we require complete overlap in the covariate distributions:

Assumption 3 (Support condition). For all x

$$\delta < \Pr(D_i = 1 \mid X_i = x) < 1 - \delta,$$

for some $\delta > 0$ and for all x in the support of X.

The support condition assumption implies that for all values of the covariates one can find units in the first sub-population.[7] Note that the "no macro-effects" assumption, detailed in the next section, also relies on the validity of this support condition.

Under Assumptions 1–3, the following results holds:

Lemma 1 (External validity given unconfounded location). *Suppose Assumptions* 1–3, *hold. Then*:

$$\begin{aligned} &E[Y_i(1) - Y_i(0) \mid D_i = 1] \\ &\quad = E[E[Y_i \mid T_i = 1, D_i = 0, X_i] - E[Y_i \mid T_i = 0, D_i = 0, X_i] \mid D_i = 1]. \end{aligned}$$

Proof. By unconfounded location, the average treatment effect conditional on the covariates does not depend on the location:

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

By random assignment in the initial location the average effect within a location conditional on the covariates is equal to the difference in average outcomes by

[7] If it is not satisfied, one can redefine the estimand of interest as the conditional treatment effect in the sub-population with common covariate support, essentially dropping units in the $D_i = 1$ location with covariates outside the common support.

treatment status and covariates:

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

Then we can average this difference estimated on the original location ($D_i = 0$) over the distribution of covariates in the location of interest ($D_i = 1$) to get the average treatment effect of interest:

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

which finishes the proof. □

The unconfounded location assumption is formally similar to the unconfounded assignment or selection on observables assumption that is often made in non-experimental evaluations, e.g.,

$$T_i \perp (Y_i(0), Y_i(1)) \mid X_i, \tag{2}$$

(Rosenbaum and Rubin, 1983). 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, however, the two assumptions are very different. For example, randomization of treatments within a site guarantees unconfoundedness of the assignment—but it does not address the unconfoundedness of location. In addition, the covariates that make Assumption 2 plausible may be very different from covariates that make unconfounded assignment to treatment plausible. Although in both cases we are concerned with covariates that are correlated with the outcomes, in the first case the covariates need also be correlated with the location decision, and in the second case they should be correlated with a very different decision, that is, the decision to participate in the program.

3. Macro-effects

A threat to the validity of the weighted "within" estimators of treatment effects implied by Lemma 1 is the presence of macro-effects, effects of variables whose values are constant within a location, or at least within sub-locations. If the large sample approximation is based on letting the number of observations within each location go to infinity, keeping the number of (sub-) locations fixed, then it is impossible to adjust for differences between the two populations unless the value of the macro-variables is exactly identical in some of the locations. Suppose for example the initial implementation is in a heavily agricultural economy, and the second implementation is in a more urban economy. The impact of the program may differ for comparable individuals because their environments are different, with demand for particular skills higher in one economy than in the other. Because such variables take on the same value within each location, they automatically fail the support condition.

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, and the potential changes in the economy over time, there is no guarantee that conditions are ever similar enough to allow accurate predictions. The design considerations for addressing this problem involve the following strategies. First, one should use initial implementations in locations that are as similar as possible in characteristics and time to the location of interest. Second, one should collect as much detail as possible on covariates that can proxy for differences in conditions across locations and time. Here past values of the outcomes of interest are particularly relevant. Even if the differences are due to location-level characteristics, their effect is likely to show up in labor market histories. Third, one should use multiple locations in order to be able to use model-based adjustment for location-level characteristics.

To be precise, let us discuss this in more detail. Suppose we wish to predict the outcomes in a location where the local unemployment rate is u_0. Suppose one is willing to assume that the outcome y as a function of the local unemployment rate u is $y_i = h(u_i;\gamma) + \varepsilon_i$ (ignoring for simplicity the presence of other covariates). A simple case would be the linear model with $y_i = \gamma_0 + \gamma_1 \cdot u_i + \varepsilon_i$. Using only the alternative locations (that is not using the location with the unemployment rate equal to u_0) we estimate γ. With only two alternative locations (and thus only two values for u_i) one can estimate at most a two-parameter model. With multiple locations, however, one may be able estimate a more realistic model, even if one cannot relax the parametric assumptions entirely: without letting the number of values of the unemployment rate get large, one cannot identify a non-parametric regression function. Given the estimated value for γ, we predict the outcome in the location of interest as $h(u_0, \hat{\gamma})$. In the application in Sections 7 and 8 we use a linear model for the macro-variables.

4. Heterogeneous treatments

A complication that has been ruled out so far is heterogeneity in treatments (as opposed to heterogeneity in treatment effects, which has been allowed for). 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 first assigned to one of two groups. Individuals in the first group, the treatment group, are to receive a variety of services, whereas individuals in the second group, the control group, are embargoed from receiving any of these services. Conditional on being assigned to the treatment group individuals are assigned to different "tracks" based on additional screenings and interviews. Some tracks may involve classroom training, while others involve on-the-job training or job search assistance. These additional screenings, and the assignment conditional on their results, often differ between locations, resulting in heterogeneity in the treatments received. Here we investigate the implications of treatment heterogeneity on strategies for predicting the effect of future programs.

The key assumption we make is that the no-training or control treatment is the same in all locations. The nature of the training, however, may well differ between locations. 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, $T_i = 1\{\tilde{T}_i \geqslant 1\}$. 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

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

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

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

The dependence arises from the assignment rules used. Typically assigned to particular training company.

Because the null treatment is randomly assigned in this scenario, it is still true that under the unconfounded location assumption, random assignment and overlap in the distributions, that average outcomes in the new location, conditional on no training, can be estimated without additional assumptions, or

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

However, in general one cannot estimate the average outcomes for trainees from the data from the other location under these assumptions:

$$E[Y_i \mid T_i = 1, D_i = 1] \neq E[E[Y_i \mid T_i = 1, D_i = 0, X_i] \mid D_i = 1].$$

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

Only in special cases is estimation of the average effect of the second implementation still feasible. For example, if the assignment rule for the different components is both identical in the two locations, and unconfounded given observed covariates in both locations, then one could predict the average outcomes in the second location. This is often implausible as assignment is often based on personal screenings.

5. Testing unconfoundedness and treatment heterogeneity

In this section, we describe the empirical strategy for evaluating the unconfoundedness and treatment homogeneity assumptions given the availability of 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 using data from the second location:

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

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

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

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, even if the treatments are heterogeneous, the two estimators should be close and statistical tests can be based on their comparison.

If treatments are homogeneous, we can follow the same procedure for the treatment groups. However, if treatments are heterogeneous, even if location is unconfounded, it is no longer true that

$$E[Y_i \mid T_i = 1, D_i = 1] = E[E[Y_i \mid T_i = 1, D_i = 0, X_i] \mid D_i = 1].$$

We therefore carry out the tests in two steps. First we test the equality for the control groups, interpreting this as a test of the location unconfoundedness assumption. Second, we carry out the same tests for the treatment groups. If the tests are rejected there—but not for the control group—we interpret this as a rejection of the homogeneous treatment effect assumption. To implement these tests we use matching methods, which have been used extensively in the evaluation literature. The specific methods we use, matching with replacement, is described below and in more detail in Abadie and Imbens (2004).

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 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.[8]

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 3 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

[8] 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.

job skills training and targeted AFDC applicants and recipients with children at least 6 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 6 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 6 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.

Four modifications were made to the basic data sets. First, individuals with children less than 6 years old were excluded from the analyses because of the severe imbalance in their distribution across programs. (Individuals with children under 6 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.[9] These modifications guarantee overlap in the marginal covariate distributions.

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 (in thousands of dollars) and for an indicator of positive earnings in each of the four quarters. 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. For each variable a t-statistic is reported for each corresponding to the test of the null hypothesis that the average value in that location is the same as the average in the other locations.

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 employed in any given quarter prior to randomization, whereas in Baltimore the incidence of employment prior to randomization was as high as 30%.

[9] These are obtained from the Regional Economic Information System (REIS) data compiled by the Bureau of Economic Analysis (BEA).

Table 1
Summary statistics and *t*-statistics for difference with pooled data

Cost per head	Pooled (7916)		San Diego (2603)		Arkansas (480)		Baltimore (2080)		Virginia (2753)	
	Mean	s.d.	Mean $919	*t*-stat	Mean $118	*t*-stat	Mean $953	*t*-stat	Mean $430	*t*-stat
Personal char.										
High school dipl.	0.47	(0.50)	0.56	[11.2]	0.50	[1.3]	0.40	[−7.2]	0.43	[−5.1]
Non-white	0.69	(0.46)	0.69	[−0.7]	0.83	[8.3]	0.70	[1.0]	0.67	[−3.6]
Never married	0.30	(0.46)	0.26	[−5.6]	0.35	[2.2]	0.38	[8.2]	0.28	[−3.6]
One child	0.47	(0.50)	0.48	[1.4]	0.42	[−2.3]	0.48	[1.4]	0.46	[−1.4]
More than one child	0.53	(0.50)	0.52	[−1.4]	0.58	[2.3]	0.52	[−1.4]	0.54	[1.4]
Pre-training earnings										
Earn Q-1	0.36	(0.89)	0.40	[2.4]	0.18	[−7.4]	0.42	[3.4]	0.31	[−3.7]
Earn Q-2	0.37	(0.90)	0.40	[2.4]	0.17	[−8.0]	0.42	[3.2]	0.32	[−3.3]
Earn Q-3	0.36	(0.90)	0.38	[1.6]	0.19	[−6.6]	0.44	[4.6]	0.30	[−4.1]
Earn Q-4	0.34	(0.87)	0.37	[2.1]	0.18	[−6.6]	0.42	[5.0]	0.28	[−5.1]
Earn Q-1 pos.	0.26	(0.44)	0.27	[0.5]	0.17	[−5.5]	0.29	[3.6]	0.25	[−1.4]
Earn Q-2 pos.	0.24	(0.43)	0.25	[0.5]	0.16	[−4.7]	0.31	[7.6]	0.20	[−5.9]
Earn Q-3 pos.	0.25	(0.43)	0.25	[0.7]	0.14	[−6.7]	0.30	[5.5]	0.23	[−3.2]
Earn Q-4 pos.	0.25	(0.44)	0.25	[0.1]	0.17	[−4.9]	0.29	[4.1]	0.24	[−1.7]
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	1.71	(3.24)	2.08	[4.4]	0.84	[−7.2]	1.65	[−0.7]	1.59	[−2.1]
Year 1 earn contr	1.63	(3.50)	1.77	[1.8]	0.71	[−7.5]	1.75	[1.3]	1.52	[−1.1]
Ave treat eff [*t*-stat]	0.08	[1.0]	0.30	[2.0]	0.13	[0.8]	−0.10	[−0.7]	0.07	[0.6]
Year 2 earn train	2.50	(4.46)	2.86	[3.0]	1.28	[−7.0]	2.54	[0.3]	2.39	[−1.5]
Year 2 earn contr	2.22	(4.34)	2.28	[0.6]	1.10	[−6.5]	2.49	[2.5]	2.12	[−0.9]
Ave treat eff [*t*-stat]	0.28	[2.8]	0.57	[2.8]	0.18	[0.8]	0.05	[0.3]	0.27	[1.7]
Post-training employment										
Year 1 emp train	0.49	(0.50)	0.52	[3.1]	0.29	[−6.7]	0.47	[−1.0]	0.49	[0.7]
Year 1 emp contr	0.41	(0.49)	0.40	[−1.0]	0.27	[−5.4]	0.43	[1.0]	0.45	[2.9]
Ave treat eff [*t*-tstat]	0.07	[6.9]	0.12	[6.2]	0.02	[0.5]	0.05	[2.1]	0.04	[1.9]

Table 1 (*continued*)

Cost per head	Pooled (7916)		San Diego (2603)		Arkansas (480)		Baltimore (2080)		Virginia (2753)	
	Mean	s.d.	Mean $919	*t*-stat	Mean $118	*t*-stat	Mean $953	*t*-stat	Mean $430	*t*-stat
Year 2 emp train	0.49	(0.50)	0.49	[−0.4]	0.31	[−6.1]	0.49	[0.0]	0.52	[2.9]
Year 2 emp contr	0.43	(0.49)	0.40	[−2.5]	0.27	[−5.7]	0.47	[3.1]	0.46	[2.5]
Ave treat eff [*t*-stat]	0.07	[6.8]	0.09	[4.6]	0.04	[0.9]	0.03	[1.2]	0.06	[2.8]
No. receiving training	4406		1297		229		1016		1864	
No. receiving no training	3510		1306		251		1064		889	

The percentage white ranged from 17% in Arkansas to 33% in Virginia. The percentage with a high school degree ranged from 40% in Baltimore to 55% in San Diego. Given all these differences, 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 3% in Baltimore to 9% in San Diego, with both differences statistically significant. In all four locations the employment to population ratios increase over the duration of the programs, although the levels are somewhat different.

7. Analyses

We focus on five issues for predicting the effect of each program using data from the other programs. First, we examine the importance of restricting the samples by discarding men and women without children at least 6 years old, so as to ensure overlap in the distribution of pre-training variables. Second, after restricting the sample, we predict outcomes separately for two sub-populations: women with no previous employment history, and women with some previous work experience. Third, we predict outcomes for controls and trainees separately to highlight the potential for heterogeneous training effects. Fourth, we analyze the sensitivity of the results to the choice of a single control location versus combining the control groups from the other three locations. Fifth, we examine the importance of different sets of pre-training covariates for predicting results in other locations.

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 compare each location with all three other locations together, for controls and for trainees separately. Then we compare each location with each

other location separately. Finally, we investigate the sensitivity to excluding sets of covariates. For all these comparisons we analyze separately those with some and those with no employment in the four quarters prior to randomization.

Below we discuss the methods used to predict the average outcomes in the target area. As an example we will focus on the prediction of average earnings in the first post-randomization year for the control group with some prior earnings in San Diego using data from the other three locations. Using matching methods described in detail in Abadie and Imbens (2004), we predict the average outcomes for controls, $\mu_{\mathrm{SD},c,1} = E[Y \mid D_i = \mathrm{SD}, T_i = 0, E_i = 1]$, using both pre- and post-randomization data from Arkansas, Baltimore and Virginia, but only pre-randomization data from San Diego. We then compare this indirect estimate to the average outcome for controls in San Diego,

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

where $N_{d,w,e}$ is the number of observations with treatment status $w \in \{c, t\}$, location $d \in \{\mathrm{SD}, \mathrm{AK}, \mathrm{VA}, \mathrm{MD}\}$ and prior employment history $e \in \{0, 1\}$ (not employed at all in previous four quarters or some employment in previous four quarters). We report the differences between the direct and indirect estimates, and the t-statistic associated with the test of the null hypothesis that the estimands are equal.

Now we describe the matching methods in more detail. We estimate the average outcome in San Diego by matching each control observation in this group to the closest control in the three other locations in terms of the covariates. Let x be the vector of covariates for a control observation in San Diego. Then we search for the control observation with covariate vector z in the three other locations that minimizes $(x - z)'(x - z)$. Before searching for the nearest control we normalize the covariates so they have zero mean and unit variance. The twelve covariates included in the vector of covariates x are four dummy variables for personal characteristics (high school diploma, non-white, married, and more than one child), the level of earnings in each of the four quarters preceding randomization, indicators for positive earnings in those four quarters. In contrast to many matching procedures (see Rosenbaum, 1995, for a survey) the matching is done with replacement so that the order in which the San Diego observations are matched does not matter. As a result of the matching we have 519 pairs, one for each San Diego control observation with some prior employment. The matches may be used more than once, so the pairs are not necessarily independent.

After matching all the San Diego controls in this group there may still be bias left, as the matching is not exact and differences between the covariate vectors within the matched pairs remain. We attempt to remove some of the remaining bias by regressing the outcomes for the matched sample on the same 12 covariates plus the two aggregate variables (the employment-to-population ratio and real earnings per worker). That is, we take the 519 matched observations, which are the observations from the three other locations matched to the 519 San Diego observations. We regress these 519 observations (none from San Diego) on the set of fourteen covariates. Given the estimated regression coefficients we adjust the matched

outcomes for differences in the matched covariate values. For example, suppose the regression coefficients are $\hat{\beta}$, the San Diego observation has covariates x (including possibly aggregate variables), and its match has covariates z, (presumably close to x for variables other than the macro variables, but not identical). Then the raw outcome for the match, say y, is adjusted to $y + \hat{\beta}(x - z)$. With z close to x this bias adjustment should not be large, but in simulations in Abadie and Imbens (2004) it leads to a substantial reduction in the bias of the average treatment effect. Note that because the lagged outcomes are included in this regression, the results are identical to those we would obtain if the outcome was redefined as the change in earnings.

In an additional use of the aggregate variables beyond that in the regression adjustment, we also deflate the individual-level earnings measures in Arkansas, Baltimore and Virginia by the ratio of real earnings in San Diego to the average real earnings in Arkansas, Baltimore and Virginia, respectively. The use of these adjustments are an attempt to make the earnings measures more comparable across different locations.

8. Results

8.1. The importance of overlap and prior employment status

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. 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 included observations in each location. To investigate the importance of overlap, we break down the discarded observations into men, women with children under six, and women with no children. The insignificant differences between discarded and included observation in each location are often the result of offsetting differences. For example, earnings for women without children are significantly lower than the earnings of included women (who have children at least 6 years of age), while earnings for men are significantly higher. Combining the two discarded groups leads to a sample that is, on average, not significantly different from the included observations. However, it is difficult to believe that combining men and women with young children provides a suitable control group for women with older children.

Table 3 provides summary statistics by individuals' prior employment status in the preceding four quarters. The two groups have similar demographic characteristics, with the exception that women who had some prior work experience were more likely to have a high school diploma. The effect of the training program differs considerably between the two groups. For those with no prior work experience, the programs are generally effective in raising employment rates and earnings in all four locations. There is much less evidence of this for those with prior work experience. We therefore carry out the predictions separately for those with and without prior work experience.

Table 2
Summary statistics and *t*-statistics for difference with included data

	Included		All discards		Men		Child < 6		No kids	
	Mean	s.d.	Mean	*t*-stat	Mean	*t*-stat	Mean	*t*-stat	Mean	*t*-stat
Panel A: *San Diego personal char.*										
High school dipl.	0.56	(0.50)	0.59	[1.3]	0.52	[−1.1]	0.72	[5.5]	0.33	[−3.9]
Non-white	0.69	(0.46)	0.63	[−2.6]	0.60	[−2.7]	0.63	[−1.7]	0.73	[0.8]
Never married	0.26	(0.44)	0.46	[9.1]	0.38	[4.0]	0.44	[5.7]	0.85	[13.2]
One child	0.48	(0.50)	0.37	[−5.1]	0.37	[−3.4]	0.45	[−0.8]	0	[−48.9]
More than one child	0.52	(0.50)	0.40	[−5.4]	0.36	[−5.3]	0.55	[0.8]	0	[−53.2]
Earn Q1–Q4	1.56	(3.52)	1.62	[0.4]	2.26	[2.5]	1.22	[−1.9]	0.49	[−5.4]
Earn Q1–Q4 pos.	0.39	(0.49)	0.40	[0.6]	0.47	[2.5]	0.38	[−0.4]	0.25	[−2.5]
Observations	2603		608		278		263		67	
Panel B: *Arkansas personal char.*										
High school dipl.	0.50	(0.50)	0.50	[0.1]	0.19	[−4.4]	0.51	[0.6]	0.50	[0.0]
Non-white	0.83	(0.37)	0.89	[2.7]	0.86	[0.5]	0.89	[2.7]	1	[9.79]
Never married	0.35	(0.48)	0.60	[8.5]	0.70	[4.5]	0.59	[8.3]	0.63	[2.2]
One child	0.42	(0.49)	0.38	[−1.3]	0.08	[−6.6]	0.40	[0.6]	0	[−18.5]
More than one child	0.58	(0.49)	0.58	[−0.3]	0.24	[−4.5]	0.60	[0.6]	0	[−25.9]
Earn Q1–Q4	0.72	(1.9)	0.60	[−1.1]	0.04	[−7.3]	0.64	[−0.7]	0.22	[−3.0]
Earn Q1–Q4 pos.	0.26	(0.44)	0.23	[−1.2]	0.03	[−6.9]	0.24	[−0.6]	0.13	[−1.5]
Observations	480		647		37		604		16	
Panel C: *Baltimore personal char.*										
High school dipl.	0.40	(0.49)	0.54	[6.5]	0.39	[−0.3]	0.67	[10.2]	0.39	[−0.1]
Non-white	0.70	(0.46)	0.70	[0.6]	0.52	[−5.9]	0.86	[7.3]	0.83	[2.1]
Never married	0.38	(0.48)	0.49	[5.3]	0.19	[−7.2]	0.72	[13.2]	0.56	[2.1]
One child	0.48	(0.50)	0.43	[−2.3]	0.41	[−2.3]	0.49	[0.3]	0	[−43.8]
More than one child	0.52	(0.50)	0.48	[−1.7]	0.51	[−0.4]	0.51	[−0.3]	0	[−47.4]
Earn Q1–Q4	1.71	(3.32)	1.31	[−3.1]	2.04	[1.6]	0.76	[−7.7]	1.22	[−1.3]
Earn Q1–Q4 pos.	0.44	(0.50)	0.44	[0.1]	0.59	[4.8]	0.34	[−3.9]	0.36	[−1.0]
Observations	2080		677		279		362		36	
Panel D: *Virginia personal char.*										
High school dipl.	0.43	(0.50)	0.50	[2.7]	—		0.53	[3.5]	0.51	[1.2]
Non-white	0.67	(0.47)	0.70	[1.4]	—		0.55	[−4.0]	0.71	[0.7]
Never married	0.28	(0.45)	0.54	[10.1]	—		0.19	[−3.8]	0.24	[−0.7]
One child	0.46	(0.50)	0.41	[−1.7]	—		0.43	[−1.0]	0.40	[−0.8]
More than one child	0.54	(0.50)	0.44	[−3.9]	—		0.56	[0.5]	0.58	[0.6]
Earn Q1–Q4	1.21	(2.69)	0.89	[−2.7]	—		1.05	[−1.1]	0.94	[−0.8]
Earn Q1–Q4 pos.	0.37	(0.48)	0.40	[1.3]	—		0.33	[−1.1]	0.31	[−0.9]
Observations	2753		397		0		338		55	

Table 3
Summary statistics by prior employment status

	San Diego (1587)		Arkansas (355)		Baltimore (1161)		Virginia (1748)	
	Mean	s.d.	Mean	s.d	Mean	s.d.	Mean	s.d.
Previously not employed								
Personal char.								
High school dipl.	0.50	(0.50)	0.47	(0.50)	0.32	(0.47)	0.37	(0.48)
Non-white	0.70	(0.46)	0.83	(0.38)	0.68	(0.47)	0.68	(0.47)
Never married	0.27	(0.44)	0.33	(0.47)	0.38	(0.49)	0.28	(0.44)
One child	0.45	(0.50)	0.42	(0.49)	0.46	(0.50)	0.43	(0.49)
More than one child	0.54	(0.50)	0.58	(0.49)	0.54	(0.50)	0.57	(0.49)
Pre-training earnings	All zero							
Year 1 earn	1.02	(3.10)	0.31	(1.10)	0.63	(1.81)	0.86	(2.09)
Year 2 earn	1.57	(4.29)	0.56	(1.89)	1.21	(2.90)	01.46	(3.12)
Year 1 empl	0.30	(0.46)	0.13	(0.34)	0.27	(0.44)	0.34	(0.47)
Year 2 empl	0.32	(0.47)	0.15	(0.36)	0.31	(0.46)	0.38	(0.49)
Ave treat eff Y1 earn [*t*-stat]	0.36	[2.3]	0.33	[2.8]	0.19	[1.7]	0.07	[0.5]
Ave treat eff Y2 earn [*t*-stat]	0.54	[2.5]	0.42	[2.1]	0.41	[2.4]	0.22	[1.4]
Ave treat eff Y1 empl [*t*-stat]	0.16	[6.9]	0.10	[2.7]	0.07	[2.9]	0.05	[1.9]
Ave treat eff Y2 empl [*t*-stat]	0.11	[4.6]	0.08	[2.1]	0.07	[2.6]	0.04	[2.3]
	San Diego (1016)		Arkansas (125)		Baltimore (919)		Virginia (1005)	
	Mean	s.d.	Mean	s.d.	Mean	s.d	Mean	s.d.
Previously employed								
Personal char.								
High school dipl.	0.64	(0.48)	0.57	(0.50)	0.50	(0.50)	0.52	(0.49)
Non-white	0.67	(0.47)	0.84	(0.37)	0.73	(0.44)	0.65	(0.48)
Never married	0.25	(0.43)	0.40	(0.49)	0.37	(0.48)	0.28	(0.45)
One child	0.51	(0.50)	0.41	(0.50)	0.51	(0.50)	0.50	(0.50)
More than one child	0.49	(0.50)	0.59	(0.49)	0.49	(0.50)	0.49	(0.50)
Pre-training earnings								
Earn Q-1	1.02	(1.42)	0.71	(0.79)	0.95	(1.22)	0.86	(1.04)
Earn Q-2	1.04	(1.43)	0.67	(0.83)	0.96	(1.20)	0.89	(1.11)
Earn Q-3	0.98	(1.37)	0.71	(0.90)	0.99	(1.25)	0.82	(1.18)
Earn Q-4	0.95	(1.41)	0.69	(0.84)	0.96	(1.15)	0.75	(1.07)
Earn Q-1 pos.	0.68	(0.47)	0.66	(0.48)	0.66	(0.47)	0.70	(0.46)
Earn Q-2 pos.	0.65	(0.48)	0.66	(0.48)	0.65	(0.48)	0.66	(0.47)
Earn Q-3 pos.	0.65	(0.48)	0.55	(0.50)	0.67	(0.46)	0.56	(0.50)
Earn Q-4 pos.	0.63	(0.48)	0.63	(0.48)	0.69	(0.46)	0.67	(0.47)
Year 1 earn	3.34	(4.53)	2.08	(2.59)	3.05	(4.30)	2.81	(3.69)
Year 2 earn	4.13	(5.91)	2.98	(3.47)	4.17	(5.11)	3.76	(4.72)
Year 1 empl	0.71	(0.45)	0.70	(0.46)	0.68	(0.47)	0.73	(0.44)
Year 2 empl	0.63	(0.48)	0.69	(0.47)	0.69	(0.46)	0.70	(0.46)
Ave treat eff Y1 earn [*t*-stat]	0.29	[1.0]	−0.21	[−0.5]	−0.25	[−0.9]	0.16	[0.5]
Ave treat eff Y2 earn [*t*-stat]	0.72	[1.9]	−0.23	[−0.4]	−0.17	[−0.5]	0.46	[1.5]
Ave treat eff Y1 empl [*t*-stat]	0.08	[2.8]	−0.13	[−1.6]	0.05	[1.5]	0.04	[1.4]
Ave treat eff Y2 empl [*t*-stat]	0.07	[2.4]	−0.02	[−0.2]	0.00	[0.0]	0.07	[2.2]

8.2. Predicting average outcomes for individuals with prior employment versus individuals with no prior employment

Table 4 reports results separately for individuals with and without previous employment in the preceding four quarters. The accuracy with which we predict average outcomes differs between these groups. For those with some previous employment in the preceding four quarters the predictions are generally closer to the

Table 4
Differences between target location and all other locations, adjusted for all covariates

	SD/others		AR/others		Balt./others		VA/others	
	Dif	*t*-stat	Dif	*t*-stat	Dif	*t*-stat	Dif	*t*-stat
Differences for controls								
Panel A: Some previous employment								
Earnings (000's)								
Year 1	0.181	[0.6]	−0.741	[−1.4]	−0.108	[−0.3]	0.130	[0.4]
Year 2	−0.417	[−1.0]	0.529	[0.7]	0.385	[0.9]	0.294	[0.7]
Employment								
Year 1	−0.040	[−1.1]	0.394	[5.1]	−0.015	[−0.4]	0.024	[0.6]
Year 2	−0.120	[−3.1]	0.246	[2.9]	0.011	[0.3]	−0.014	[−0.3]
Panel B: No previous Employment								
Earnings (000's)								
Year 1	0.285	[2.8]	−0.678	[−12.9]	−0.523	[−8.3]	0.478	[5.8]
Year 2	0.292	[2.3]	−0.956	[−9.0]	−0.670	[−6.8]	0.594	[5.2]
Employment								
Year 1	−0.016	[−1.1]	−0.178	[−11.6]	−0.096	[−6.4]	0.163	[10.2]
Year 2	0.006	[0.4]	−0.207	[−12.1]	−0.097	[−6.1]	0.162	[9.6]
Differences for trainees								
Panel A: Some previous employment								
Earnings								
Year 1	0.460	[1.7]	0.527	[0.9]	−0.102	[−0.4]	0.907	[3.6]
Year 2	0.613	[1.8]	−0.132	[−0.2]	−0.450	[−1.2]	0.656	[2.0]
Employment								
Year 1	0.033	[1.1]	−0.325	[−4.0]	0.003	[0.1]	0.190	[5.8]
Year 2	−0.040	[−1.2]	0.010	[0.1]	−0.027	[−0.8]	0.124	[4.2]
Panel B: No previous employment								
Earnings								
Year 1	0.326	[3.5]	−0.333	[−4.2]	−0.492	[−7.3]	0.289	[4.9]
Year 2	0.268	[1.9]	−0.554	[−4.3]	−0.469	[−4.2]	0.619	[6.8]
Employment								
Year 1	0.048	[3.2]	−0.153	[−7.2]	−0.131	[−8.5]	0.124	[9.3]
Year 2	0.010	[0.7]	−0.176	[−8.0]	−0.106	[−6.6]	0.169	[12.2]

actual averages. For example, the adjusted difference for the earnings of controls between San Diego and the other locations for the first year after randomization is \$181, with a t-statistic of 0.6. For the group of controls with no prior employment the adjusted difference is \$285, with a t-statistic of 2.8. Note that the average post-randomization earnings for the group with some prior earnings is much higher at \$3,340, versus \$1,020 for those with no prior earnings (see Table 3), so that the difference in percentage terms is considerably higher, 5% versus 28%. In Baltimore, for the group with prior employment the difference in average earnings of controls in the first year is −\$108, with a t-statistic of 0.3, and for those with no prior employment the difference is −\$523 with a t-statistic of 8.3. This pattern is consistent across locations and outcomes. With some exceptions—in particular Arkansas in the first-year employment outcome of controls—predictions for those with prior employment tend to be more accurate, especially in relative terms, than those with no recent prior employment, and the latter tend to be statistically much more significant, partly given their lower variances. These conclusions hold, to a somewhat lesser extent, for the treated individuals.

8.3. Predicting average outcomes for controls versus trainees

To investigate whether heterogeneity in the four different programs affected our ability to predict outcomes, we assessed the accuracy of predictions for the treatment group. These results are reported in the second panel of Table 4. We find that the level of predictive accuracy for the trainees is similar to that for the controls once we separate out the groups with and without recent prior employment. In comparison with the results for the controls this is somewhat dependent on the increased precision in predicting future outcomes for those with no employment experience. This in turn is not due to differences in sample size but smaller variances for the group with no prior employment experience as most of those have subsequently zero earnings. Overall, however, the predicted errors for trainees with prior employment are not significantly different from zero, whereas they generally are for trainees with zero prior earnings.

8.4. 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, as reported in Tables 5–8. Consider the first-year earnings for controls. We predict average earnings in San Diego using all three alternative locations fairly well (with a prediction error of \$181), but the prediction error is \$1,278 using only Arkansas data. Similarly, we predict employment in the first year accurately using the

Table 5
Differences between San Diego and other locations, adjusted for all covariates

	SD/others		SD/AR		SD/Balt.		SD/VA	
	Dif	*t*-stat	Dif	*t*-stat	Dif	*t*-stat	Dif	*t*-stat
Differences for controls								
Panel A: Some previous employment								
Earnings								
Year 1	0.181	[0.6]	1.278	[1.8]	−0.029	[0.1]	0.465	[1.4]
Year 2	−0.417	[−1.0]	0.871	[1.0]	−0.639	[−1.3]	0.117	[0.3]
Employment								
Year 1	−0.040	[−1.1]	−0.139	[−1.9]	0.046	[1.1]	−0.073	[−1.8]
Year 2	−0.120	[−3.1]	−0.085	[−1.0]	−0.056	[−1.3]	−0.110	[−2.5]
Panel B: No previous employment								
Earnings								
Year 1	0.284	[2.8]	0.698	[3.4]	0.267	[2.0]	−0.025	[−0.2]
Year 2	0.292	[2.3]	0.916	[3.6]	0.214	[1.3]	−0.237	[−1.5]
Employment								
Year 1	−0.016	[−1.1]	0.149	[5.4]	−0.003	[−0.2]	−0.148	[−8.5]
Year 2	0.006	[0.4]	0.153	[5.2]	−0.009	[−0.5]	−0.098	[−5.3]
Differences for trainees								
Panel A: Some previous employment								
Earnings								
Year 1	0.460	[1.7]	1.529	[2.5]	0.899	[2.9]	0.343	[1.2]
Year 2	0.612	[1.8]	0.812	[0.9]	0.137	[0.3]	0.658	[1.8]
Employment								
Year 1	0.033	[1.1]	0.157	[2.0]	0.068	[1.8]	0.010	[0.3]
Year 2	−0.040	[−1.2]	−0.023	[−0.3]	−0.026	[−0.6]	−0.027	[−0.7]
Panel B: No previous employment								
Earnings								
Year 1	0.326	[3.5]	0.650	[3.2]	0.379	[2.7]	0.145	[1.4]
Year 2	0.268	[1.9]	1.032	[3.3]	0.185	[0.9]	0.020	[0.1]
Employment								
Year 1	0.048	[3.2]	0.206	[6.2]	0.081	[3.6]	−0.011	[−0.7]
Year 2	0.010	[0.7]	0.176	[5.2]	0.019	[0.8]	−0.041	[−2.5]

combined control group (a prediction error of 4.0%), but the prediction performs quite poorly using only the Virginia data (a prediction error of 7.3%). 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 generally improves prediction, presumably because combining the locations increases the degree of overlap with the San Diego

Table 6
Differences between Arkansas and other locations, adjusted for all covariates

	AR/others		AR/SD		AR/Balt.		AR/VA	
	Dif	*t*-stat	Dif	*t*-stat	Dif	*t*-stat	Dif	*t*-stat
Differences for controls								
Panel A: Some previous employment								
Earnings								
Year 1	−0.741	[−1.4]	−0.559	[−0.9]	0.284	[0.6]	−1.495	[−2.6]
Year 2	0.529	[0.7]	−0.573	[−0.7]	−0.822	[−0.9]	−1.649	[−2.4]
Employment								
Year 1	0.394	[5.4]	0.155	[1.9]	0.116	[1.4]	−0.012	[−0.1]
Year 2	0.246	[2.9]	0.094	[1.0]	−0.041	[−0.502]	−0.082	[−1.0]
Panel B: No previous employment								
Earnings								
Year 1	−0.678	[−12.9]	−0.665	[−10.8]	−0.433	[−7.9]	−0.453	[−7.7]
Year 2	−0.956	[−9.0]	−0.869	[−7.3]	−0.656	[−5.5]	−0.722	[−6.2]
Employment								
Year 1	−0.177	[−11.6]	−0.143	[−8.5]	−0.147	[−8.3]	−0.177	[−10.4]
Year 2	−0.207	[−12.1]	−0.163	[−8.7]	−0.160	[−8.2]	−0.205	[−10.5]
Differences for trainees								
Panel A: Some previous employment								
Earnings								
Year 1	0.527	[0.9]	0.596	[−1.1]	−0.973	[−1.3]	0.161	[0.3]
Year 2	−0.132	[−0.2]	0.430	[0.6]	−1.745	[−1.7]	−1.001	[−1.4]
Employment								
Year 1	−0.325	[−4.0]	−0.127	[−1.3]	0.029	[0.3]	−0.020	[−0.2]
Year 2	0.010	[0.1]	0.126	[1.3]	0.008	[0.1]	0.020	[0.2]
Panel B: No previous employment								
Earnings								
Year 1	−0.333	[−4.2]	−0.507	[−5.8]	−0.281	[−3.1]	−0.187	[−2.2]
Year 2	−0.554	[−4.3]	−0.774	[−5.4]	−0.778	[−5.2]	−0.383	[−2.9]
Employment								
Year 1	−0.153	[−7.2]	−0.183	[−7.8]	−0.127	[−5.2]	−0.126	[−5.7]
Year 2	−0.176	[−8.0]	−0.169	[−7.1]	−0.179	[−6.8]	−0.170	[−7.5]

sample. Again, we are better able to predict outcomes for those with previous employment versus those with no previous employment, although the use of multiple locations is helpful for prediction in both groups. For example, employment in the first and second year for controls with no prior employment is predicted quite well using all control groups (1.6% and 0.6% prediction error, respectively). However,

Table 7
Differences between Baltimore and other locations, adjusted for all covariates

	Balt./others		Balt./AR		Balt./SD		Balt./VA	
	Dif	*t*-stat	Dif	*t*-stat	Dif	*t*-stat	Dif	*t*-stat
Differences for controls								
Panel A: Some previous employment								
Earnings								
Year 1	−0.107	[−0.3]	1.318	[2.1]	0.232	[0.6]	0.760	[2.0]
Year 2	0.385	[0.9]	1.390	[1.9]	1.031	[2.2]	1.108	[2.3]
Employment								
Year 1	−0.015	[−0.4]	−0.154	[−2.2]	−0.013	[−0.3]	0.033	[0.7]
Year 2	0.011	[0.3]	0.029	[0.4]	0.082	[1.8]	−0.004	[−0.1]
Panel B: No previous employment								
Earnings								
Year 1	−0.523	[−8.3]	0.387	[3.0]	−0.198	[−2.7]	0.546	[−7.1]
Year 2	−0.670	[−6.8]	0.576	[2.9]	−0.142	[−1.2]	−0.802	[−6.6]
Employment								
Year 1	−0.096	[−6.4]	0.144	[4.7]	0.014	[0.8]	−0.198	[−10.6]
Year 2	−0.097	[−6.1]	0.136	[4.2]	0.002	[0.1]	−0.149	[−7.5]
Differences for trainees								
Panel A: Some previous employment								
Earnings								
Year 1	−0.102	[−0.4]	1.534	[2.7]	−0.140	[−0.4]	−0.007	[−0.0]
Year 2	−0.450	[−1.2]	1.312	[1.9]	0.394	[0.9]	0.754	[1.9]
Employment								
Year 1	0.003	[0.1]	0.224	[3.1]	−0.005	[−0.1]	0.002	[0.1]
Year 2	−0.026	[−0.8]	0.093	[1.2]	0.046	[1.1]	0.018	[0.5]
Panel B: No previous employment								
Earnings								
Year 1	−0.492	[−7.3]	0.281	[1.9]	−0.306	[−3.6]	−0.453	[−6.2]
Year 2	−0.469	[−4.2]	0.837	[3.4]	0.061	[0.4]	−0.673	[−5.5]
Employment								
Year 1	−0.131	[−8.5]	0.119	[3.4]	−0.060	[−3.1]	−0.132	[−7.9]
Year 2	−0.106	[−6.6]	0.174	[4.8]	0.002	[0.1]	−0.154	[−8.8]

using only Arkansas data, the prediction errors are 14.9% and 15.3% for the same group.

The lower half of Table 5 reports the same calculations for program trainees in the comparison of San Diego to the other three separate locations. The prediction patterns are similar as those for controls, although the relative performance of the

Table 8
Differences between Virginia and other locations, adjusted for all covariates

	VA/others		VA/AR		VA/Balt.		VA/SD	
	Dif	*t*-stat	Dif	*t*-stat	Dif	*t*-stat	Dif	*t*-stat
Differences for controls								
Panel A: Some previous employment								
Earnings								
Year 1	0.129	[0.4]	0.906	[1.6]	−0.219	[−0.6]	−0.044	[−0.1]
Year 2	0.294	[0.7]	0.616	[0.8]	0.007	[0.015]	0.188	[0.4]
Employment								
Year 1	0.024	[0.6]	−0.105	[−1.4]	0.072	[1.6]	0.057	[1.3]
Year 2	−0.014	[−0.3]	−0.018	[−0.2]	0.015	[0.3]	0.090	[1.9]
Panel B: No previous employment								
Earnings								
Year 1	0.478	[5.8]	0.681	[4.4]	0.284	[2.8]	0.054	[0.6]
Year 2	0.594	[5.2]	0.857	[4.0]	0.328	[2.3]	0.168	[1.3]
Employment								
Year 1	0.163	[10.2]	0.212	[6.9]	0.073	[3.6]	0.088	[4.8]
Year 2	0.162	[9.6]	0.189	[5.9]	0.063	[2.9]	0.078	[4.0]
Differences for trainees								
Panel A: Some previous employment								
Earnings								
Year 1	0.907	[3.6]	1.274	[2.6]	0.536	[1.9]	−0.392	[−1.4]
Year 2	0.656	[2.0]	0.968	[1.3]	−0.058	[−0.1]	−0.316	[−0.8]
Employment								
Year 1	0.190	[5.8]	0.291	[3.8]	0.067	[1.8]	−0.005	[−0.2]
Year 2	0.124	[4.2]	0.093	[1.2]	0.044	[1.2]	0.059	[1.7]
Panel B: No previous employment								
Earnings								
Year 1	0.289	[4.9]	0.427	[3.2]	0.144	[1.8]	−0.187	[−2.6]
Year 2	0.619	[6.8]	0.887	[4.2]	0.077	[0.6]	0.025	[0.3]
Employment								
Year 1	0.124	[9.3]	0.177	[5.7]	0.056	[3.1]	−0.014	[−0.9]
Year 2	0.169	[12.2]	0.217	[7.0]	0.053	[2.8]	0.040	[2.4]

separate locations such as Baltimore and Virginia are different. Virginia performs relatively better than Baltimore for predicting outcomes of trainees, but the use of the combined sample still reduces the prediction error compared to using only individual locations. The results in Tables 6–8 repeat the exercise, but predict outcomes in Arkansas, Baltimore and Virginia, respectively. The results in these tables confirm the results of Table 5; prediction errors are generally larger when

using single locations rather than the combined set of alternative locations. Predicting results for controls in Arkansas with no prior work experience is the exception to this pattern; combining other locations does not seem to help for predicting the outcomes of these women.

8.5. Choosing pre-training variables

In Tables 9–12 we investigate sensitivity to the choice of control variables for predicting each of the four outcome variables. Once we separate the sample into those with and without recent employment experience the results are remarkably insensitive to the inclusion of additional variables, contrasting with some of the other literature (e.g. Heckman et al., 1998; Dehejia and Wahba, 1999). The first four rows of each table report results adjusting for no covariates, personal characteristics only, personal characteristics and prior earnings, and finally personal characteristics and

Table 9
Adjusted differences between target location and all other locations in first year post-training earnings

	SD/others		AR/others		Balt./others		VA/others	
	Dif	*t*-stat	Dif	*t*-stat	Dif	*t*-stat	Dif	*t*-stat
Differences for controls								
Panel A: Some previous employment								
Adjusted for								
No covariates	−0.136	[−0.6]	0.319	[1.0]	0.213	[1.2]	−0.175	[−0.8]
Personal only	−0.227	[−1.1]	0.115	[0.4]	0.282	[1.6]	−0.145	[−0.7]
Pers., earn.	0.182	[0.6]	−0.171	[−0.3]	0.511	[1.6]	0.009	[0.0]
Pers., earn., agg.	0.181	[0.6]	−0.741	[−1.4]	−0.108	[−0.3]	0.129	[0.4]
Panel B: No previous employment								
No covariates	0.209	[2.1]	−0.595	[−11.3]	−0.204	[−3.3]	0.199	[2.5]
Personal only	0.178	[1.8]	−0.622	[−11.8]	−0.122	[−1.9]	0.201	[2.5]
Pers., earn.	0.178	[1.8]	−0.622	[−11.8]	−0.122	[−1.9]	0.201	[2.5]
Pers., earn., agg.	0.284	[2.8]	−0.678	[−12.9]	−0.523	[−8.3]	0.478	[5.8]
Differences for trainees								
Panel A: Some previous employment								
Adjusted for								
No covariates	0.141	[0.8]	−0.723	[−1.7]	0.011	[0.1]	−0.034	[−0.2]
Personal only	0.123	[0.7]	−0.698	[−1.6]	0.009	[0.1]	−0.015	[−0.1]
Pers., earn.	0.473	[1.8]	−0.855	[−1.4]	0.396	[1.4]	0.078	[0.3]
Pers., earn., agg.	0.460	[1.7]	0.527	[0.9]	−0.102	[−0.4]	0.907	[3.6]
Panel B: No previous employment								
No covariates	0.368	[4.0]	−0.455	[−5.7]	−0.236	[−3.5]	−0.039	[−0.7]
Personal only	0.292	[3.1]	−0.421	[−5.3]	−0.165	[−2.5]	−0.015	[−0.3]
Pers., earn.	0.292	[3.1]	−0.421	[−5.3]	−0.165	[−2.5]	−0.015	[−0.3]
Pers., earn., agg.	0.326	[3.5]	−0.333	[−4.2]	−0.492	[−7.3]	0.289	[4.9]

Table 10
Adjusted differences between target location and all other locations in second year post-training earnings

	SD/others		AR/others		Balt./others		VA/others	
	Dif	*t*-stat	Dif	*t*-stat	Dif	*t*-stat	Dif	*t*-stat
Differences for controls								
Panel A: Some previous employment								
Adjusted for								
No covariates	−0.615	[−2.4]	0.476	[1.2]	0.723	[3.2]	−0.237	[−0.9]
Personal only	−0.641	[−2.5]	0.233	[0.6]	0.762	[3.4]	−0.174	[−0.7]
Pers., earn.	−0.420	[−1.0]	−1.245	[−1.6]	1.150	[2.8]	0.449	[1.1]
Pers., earn., agg.	−0.417	[−1.0]	0.529	[0.7]	0.385	[0.9]	0.294	[0.7]
Panel B: No previous employment								
No covariates	0.183	[1.5]	−0.856	[−8.0]	−0.153	[−1.6]	0.286	[2.5]
Personal only	0.138	[1.1]	−0.875	[−8.2]	−0.098	[−1.0]	0.308	[2.7]
Pers., earn.	0.138	[1.1]	−0.875	[−8.2]	−0.098	[−1.0]	0.308	[2.7]
Pers., earn., agg.	0.292	[2.3]	−0.956	[−9.0]	−0.670	[−6.8]	0.594	[5.2]
Differences for trainees								
Panel A: Some previous employment								
Adjusted for								
No covariates	−0.050	[−0.2]	−0.833	[−2.0]	0.235	[1.1]	−0.030	[−0.2]
Personal only	−0.062	[−0.3]	−0.837	[−2.0]	0.251	[1.2]	−0.003	[−0.0]
Pers., earn.	0.487	[1.4]	−1.701	[−2.3]	0.386	[1.0]	0.044	[0.1]
Pers., earn., agg.	0.612	[1.8]	−0.132	[−0.2]	−0.450	[−1.2]	0.656	[2.0]
Panel B: No previous employment								
No covariates	0.323	[2.3]	−0.807	[−6.3]	−0.131	[−1.2]	0.011	[0.1]
Personal only	0.190	[1.3]	−0.840	[−6.6]	0.041	[0.4]	0.074	[0.8]
Pers., earn.	0.190	[1.3]	−0.840	[−6.6]	0.041	[0.4]	0.074	[0.8]
Pers., earn., agg.	0.268	[1.9]	−0.554	[−4.3]	−0.469	[−4.2]	0.619	[6.8]

prior earnings with adjustment for aggregate variables. Adjustments are presented separately for each outcome variable, using the combined sample to predict each individual location. Table 9 reports the results for first year earnings. The inclusion of personal characteristics (race, education and children) matters remarkably little. For example, the prediction error in San Diego for controls with no prior employment is reduced from $209 to $178, although the prediction error for San Diego controls with some prior work experience is actually slightly larger with the inclusion of personal characteristics (−$227 versus $136). The overall pattern across locations and types of participants (based on prior work experience) is that the use of personal characteristics does not matter much for prediction.

Similarly, the inclusion of additional earnings or employment controls does not seem to matter much for prediction. Comparing the second and third rows of Table 9 (results for the first year earnings outcome), the prediction error for San Diego women with some previous employment experience is −$227 versus $182 once

Table 11
Adjusted differences between target location and all other locations in first year post-training employment

	SD/others		AR/others		Balt./others		VA/others	
	Dif	*t*-stat	Dif	*t*-stat	Dif	*t*-stat	Dif	*t*-stat
Differences for controls								
Panel A: Some previous employment								
Adjusted for								
No covariates	−0.007	[−0.4]	0.078	[2.0]	0.034	[−1.7]	0.032	[1.6]
Personal only	−0.019	[−1.0]	0.061	[1.6]	0.022	[−1.1]	0.043	[2.1]
Pers., earn.	−0.039	[−1.1]	0.182	[2.4]	−0.021	[−0.6]	0.037	[1.0]
Pers., earn., agg.	−0.040	[−1.1]	0.394	[5.1]	−0.015	[−0.4]	0.024	[0.6]
Panel B: No previous employment								
No covariates	−0.020	[−1.4]	−0.165	[−10.7]	−0.005	[−0.4]	0.097	[6.0]
Personal only	−0.030	[−2.2]	−0.170	[−11.0]	0.009	[0.6]	0.098	[6.0]
Pers., earn.	−0.030	[−2.2]	−0.170	[−11.0]	0.009	[0.6]	0.098	[6.0]
Pers., earn., agg.	−0.016	[−1.1]	−0.178	[−11.6]	−0.096	[−6.4]	0.163	[10.2]
Differences for trainees								
Panel A: Some previous employment								
Adjusted for								
No covariates	0.031	[1.8]	−0.111	[−2.4]	−0.041	[−2.2]	0.021	[1.3]
Personal only	0.022	[1.3]	−0.105	[−2.3]	−0.032	[−1.7]	0.020	[1.3]
Pers., earn.	0.020	[0.7]	−0.173	[−2.1]	0.030	[0.8]	0.017	[0.6]
Pers., earn., agg.	0.033	[1.1]	−0.325	[−4.0]	0.003	[0.1]	0.190	[5.8]
Panel B: No previous employment								
No covariates	0.056	[3.8]	−0.169	[−7.9]	−0.043	[−2.8]	0.023	[1.7]
Personal only	0.043	[2.9]	−0.170	[−8.0]	−0.034	[−2.2]	0.039	[2.2]
Pers., earn.	0.043	[2.9]	−0.170	[−8.0]	−0.034	[−2.2]	0.029	[2.2]
Pers., earn., agg.	0.048	[3.2]	−0.153	[−7.2]	−0.131	[−8.5]	0.124	[9.3]

prior earnings are included. There is of course no change in the results in Panel B between the rows, as these results only pertain to women with no earnings histories. It appears that once one compares individuals with some, or individuals with no recent labor market experience, the fact that they are eligible for employment training programs is sufficient to make them adequate comparisons for similar populations in other geographical areas without many additional controls.

The second half of Table 9 confirms the same intuition for trainees; adjusting for different sets of covariates does not matter much for our ability to predict outcomes. For example, consider program trainees with no prior work experience in San Diego. Prediction error using no covariates is $368, compared to a prediction error of $292 after adjusting for personal characteristics. Fully adjusting for all personal characteristics and aggregate variables gives a prediction error of $326. Tables 10–12 report the same general patterns for the other three outcome variables (second-year earnings and first- and second-year employment outcomes).

Table 12
Adjusted differences between target location and all other locations in second year post-training employment

	SD/others		AR/others		Balt./others		VA/others	
	Dif	*t*-stat	Dif	*t*-stat	Dif	*t*-stat	Dif	*t*-stat
Differences for controls								
Panel A: Some previous employment								
Adjusted for								
No covariates	−0.080	[−4.0]	0.050	[1.2]	0.060	[3.1]	0.014	[0.7]
Personal only	−0.081	[−4.0]	0.036	[0.9]	0.067	[3.5]	0.021	[1.0]
Pers., earn.	−0.113	[−3.0]	0.044	[0.5]	0.069	[1.8]	0.044	[1.1]
Pers., earn., agg.	−0.120	[−3.1]	0.246	[2.9]	0.011	[0.3]	−0.014	[−0.3]
Panel B: No previous employment								
No covariates	−0.014	[−1.0]	−0.182	[−10.5]	−0.001	[0.1]	0.090	[5.4]
Personal only	−0.021	[−1.4]	−0.182	[−10.6]	−0.000	[0.0]	0.092	[5.5]
Pers., earn.	−0.021	[−1.4]	−0.182	[−10.6]	−0.000	[0.0]	0.092	[5.5]
Pers., earn., agg.	0.006	[0.4]	−0.207	[−12.1]	−0.097	[−6.1]	0.162	[9.6]
Differences for trainees								
Panel A: Some previous employment								
Adjusted for								
No covariates	−0.040	[−2.2]	−0.020	[−0.4]	−0.013	[−0.7]	0.047	[2.9]
Personal only	−0.046	[−2.5]	−0.023	[−0.5]	−0.009	[−0.5]	0.053	[3.3]
Pers., earn.	−0.046	[−1.4]	−0.095	[−1.1]	0.019	[0.6]	0.056	[1.9]
Pers., earn., agg.	−0.040	[−1.2]	0.010	[0.1]	−0.026	[−0.8]	0.124	[4.2]
Panel B: No previous employment								
No covariates	0.008	[0.5]	−0.189	[−8.7]	−0.025	[−1.6]	0.056	[4.2]
Personal only	−0.005	[−0.3]	−0.196	[−9.0]	−0.016	[−1.0]	0.061	[4.6]
Pers., earn.	−0.005	[−0.3]	−0.196	[−9.0]	−0.016	[−1.0]	0.061	[4.6]
Pers., earn., agg.	0.010	[0.7]	−0.176	[−8.0]	−0.106	[−6.6]	0.169	[12.2]

9. Conclusion

Using data from experimental training programs from four very different locations (San Diego, Baltimore, Arkansas and Virginia), we attempt to predict the effect of the training program in one location given outcomes in other locations. We find that we are able to predict the average outcomes for those with some prior employment fairly accurately, thus eliminating selection bias. In contrast to these results for those with prior employment records, however, we cannot predict outcomes for individuals with no prior employment in the last four quarters accurately. Once we separate out those two groups additional covariates matter little. There is some gain from combining the data from the three control groups. Heterogeneity in the programs seems to contribute little to difficulty in predicting outcomes.

Acknowledgements

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. Imbens' research was partially funded by the National Science Foundation grants SBR 9818644 and SES 0136789 to the National Bureau of Economic Research, and Mortimer's research 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.

References

Abadie, A., Imbens, G., 2004. Large sample properties of matching estimators for average treatment effects. Mimeo, Department of Economics, UC Berkeley.

Angrist, J., 1998. Estimating the labor market impact of voluntary military service using social security data on military applicants. Econometrica 66 (2), 249–288.

Angrist, J.D., Krueger, A.B., 1999. Empirical strategies in labor economics. In: Ashenfelter, A., Card, D. (Eds.), Handbook of Labor Economics, Vol. 3. Elsevier Science, New York.

Ashenfelter, O., Card, D., 1985. Using the longitudinal structure of earnings to estimate the effect of training programs. Review of Economics and Statistics 67, 648–660.

Card, D., Sullivan, D., 1988. Measuring the effect of subsidized training programs on movements in and out of employment. Econometrica 56(3), 497–530.

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

Dehejia, R., 1997. A decision-theoretic approach to program evaluation. Ph.D. Dissertation, Department of Economics, Harvard University (Chapter 2).

Dehejia, R., Wahba, S., 1998. Causal effects in non-experimental studies: re-evaluating the evaluation of training programs. NBER technical working paper.

Dehejia, R., Wahba, S., 1999. Causal effects in non-experimental studies: re-evaluating the evaluation of training programs. Journal of the American Statistical Association 94, 1053–1062.

Fraker, T., Maynard, R., 1987. The adequacy of comparison group designs for evaluations of employment-related programs. Journal of Human Resources 22 (2), 194–227.

Friedlander, D., Gueron, J., 1992. Are high-cost services more effective than low-cost services? Evidence from experimental evaluations of Welfare-to-Work programs. In: Garfinkel, I., Manski, C. (Eds.), Evaluating Welfare and Training Programs. Harvard University Press, Cambridge, MA, pp. 143–198.

Friedlander, D., Robins, P., 1995. Evaluating program evaluations: new evidence on commonly used nonexperimental methods. American Economic Review 85, 923–937.

Greenberg, D., Wiseman, M., 1992. What did the OBRA demonstrations do? In: Garfinkel, I., Manski, C. (Eds.), Evaluating Welfare and Training Programs. Harvard University Press, Cambridge, MA.

Gueron, J., Pauly, E., 1991. From Welfare to Work. Russell Sage Foundation, New York.

Hahn, J., 1998. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. Econometrica 66 (2), 315–331.

Heckman, J., Hotz, V.J., 1989. Alternative methods for evaluating the impact of training programs. Journal of the American Statistical Association (with discussion) 29 84(408) 862–880.

Heckman, J., Robb, R., 1984. Alternative methods for evaluating the impact of interventions. In: Heckman, Singer (Eds.), Longitudinal Analysis of Labor Market Data. Cambridge University Press, Cambridge.

Heckman, J., Ichimura, H., Smith, J., Todd, P., 1998. Characterizing selection bias using experimental data. Econometrica 66 (5), 1017–1098.

Hirano, K., Imbens, G., Ridder, G., 2003. Efficient estimation of average treatment effects using the estimated propensity score. Econometrica, 71(4), 1161–1189.

Hotz, V.J., 1992. Recent experience in designing evaluations of social programs: the case of the national JTPA study. In: Garfinkel, I., Manski, C. (Eds.), Evaluating Welfare and Training Programs. Harvard University Press, Cambridge, MA, pp. 76–114.

Imbens, G., 2004. Semiparametric estimation of average treatment effects under exogeneity: a review. Review of Economics and Statistics, forthcoming.

Imbens, G., Angrist, J., 1994. Identification and estimation of local average treatment effects. Econometrica 62 (2), 467–475.

Lalonde, R., 1986. Evaluating the econometric evaluations of training programs with experimental data. American Economic Review 76 (4), 604–620.

Lechner, M., 1999. Earnings and employment effects of continuous off-the-job training in East Germany after unification. Journal of Business and Economic Statistics 17 (1), 79–90.

Meyer, B., 1995. Natural and quasi-experiments in economics. Journal of Business and Economic Statistics 13 (2), 151–161.

Rosenbaum, P., 1995. Observational Studies. Springer Series in Statistics. Springer, New york.

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

Rosenbaum, P., Rubin, D., 1984. Reducing bias in observational studies using subclassification on the propensity score. Journal of the American Statistical Association 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., 1978. Bayesian inference for causal effects: the role of randomization. Annals of Statistics 6, 34–58.

Smith, J.A., Todd, P.E., 2001. Reconciling conflicting evidence on the performance of propensity-score matching methods. American Economic Review 91, 112–118.

Smith, J.A., Todd, P.E., 2004. Does matching address LaLonde's critique of nonexperimental estimators. Journal of Econometrics, forthcoming.