Econometrica, Vol. 66, No. 5 (September, 1998), 1017–1098

CHARACTERIZING SELECTION BIAS USING EXPERIMENTAL DATA[1]

BY JAMES HECKMAN, HIDEHIKO ICHIMURA, JEFFREY SMITH, AND PETRA TODD

Semiparametric methods are developed to estimate the bias that arises from using nonexperimental comparison groups to evaluate social programs and to test the identifying assumptions that justify matching, selection models, and the method of difference-in-differences. Using data from an experiment on a prototypical social program and data from nonexperimental comparison groups, we reject the assumptions justifying matching and our extensions of it. The evidence supports the selection bias model and the assumptions that justify a semiparametric version of the method of difference-in-differences. We extend our analysis to consider applications of the methods to ordinary observational data.

KEYWORDS: Selection bias, program evaluation, training programs, semiparametric estimation.

1. INTRODUCTION

A STANDARD METHOD FOR EVALUATING social programs uses the outcomes of nonparticipants to estimate what participants would have experienced had they not participated. The difference between participant and nonparticipant outcomes is the estimated gross impact of a program reported in many evaluations. The outcomes of nonparticipants may differ systematically from what the outcomes of participants would have been without the program, producing selection bias in estimated impacts. A variety of nonexperimental estimators

[1] A previous version of this paper appeared under the title "Nonparametric Characterization of Selection Bias Using Experimental Data: A Study of Adult Males in JTPA. Part I. Definitions, Applications and Empirical Results." An earlier version of it appeared in August, 1994, under the title "Evaluating the Impact of Training on the Earnings and Labor Force Status of Young Women: Better Data Help A Lot." This research was supported by NSF SBR 91-11-455, NSF SBR 93-21-048 and by a grant from the Russell Sage Foundation. This paper was presented as an invited lecture at the Latin American Econometric Society Meeting, Caracas, Venezuela, August 1994. We have benefited from comments received from workshops in September, October, and November 1994 at Yale, Princeton, Chicago, UC-San Diego, USC, Rand-UCLA, UC-Irvine, UC-Riverside, Northwestern, and U.C. London, and workshops in the Winter and Spring of 1995 at UC-Berkeley, Oslo, Washington-St. Louis, Tel Aviv, and Virginia and an NSF-sponsored conference on econometrics held in Madison, Wisconsin in June 1995. We also presented this paper in the Malinvaud Workshop in Paris, March, 1995. We are grateful to three anonymous referees, a co-editor, Derek Bandler, Lars Hansen, Bo Honoré, Lance Lochner, Thierry Magnac, Christopher Taber, Ed Vytlacil and Adonis Yatchew for helpful comments and Derek Bandler, Jingjing Hsee, Lance Lochner, and Annie Zhang for programming assistance.

adjust for this selection bias under different assumptions.[2] Under certain conditions, randomized social experiments eliminate this bias.[3]

Social experiments are costly and the identifying assumptions required to justify them are not always satisfied.[4] However, it is widely held that there is no valid alternative to experimentation as a method for evaluating social programs (see, e.g., Burtless, 1995). In an important paper, LaLonde (1986) combines data from a social experiment with data from nonexperimental comparison groups to evaluate the performance of many commonly-used nonexperimental estimators. For the particular group of parametric estimators that he investigates, and for his particular choices of regressors, he finds that the estimators chosen by econometric model selection criteria produce a range of impact estimates that is unacceptably large.

This paper uses data from a social experiment on a prototypical social program combined with data on comparison groups of persons who chose not to participate in the program evaluated by the experiment. As documented by Heckman, LaLonde, and Smith (1999), many programs in place around the world are very similar to the program we analyze in this paper.

Our analysis is based on the following principles. Neither the experimental control group nor the comparison group we analyze receives treatment, so that differences in measured outcomes between the two groups can be attributed solely to selection bias. Instead of examining the performance of specific parametric estimators based on specific sets of regressors in eliminating selection bias, as LaLonde (1986) and scholars who follow him have done, we use semiparametric econometric methods to estimate the functional form of the selection bias directly using a variety of regressors and data sets. We use the estimated bias functions to test identifying assumptions that have been maintained in the literature, and to suggest estimators that might be effective in eliminating selection bias in future evaluations of similar programs. Our method for characterizing bias is general and can be applied in a variety of settings, including the study of the analytically similar problem of sample attrition.

By characterizing the bias nonparametrically, and by examining the sensitivity of the estimated bias to many alternative sets of conditioning variables, we analyze the suitability of entire classes of estimators, rather than trying out a few parametric members of those classes with a limited set of conditioning variables. Evidence that a particular estimator with a particular set of regressors "works" in a particular data set is properly discounted by most serious analysts. There is always the suspicion that the success of an estimator in a particular instance is the consequence of a diligent specification search. We avoid that

[2] These estimators and the identifying assumptions that justify them are summarized in Heckman and Robb (1985, 1986), Heckman (1990a), Heckman and Smith (1996), and Heckman, Smith, and LaLonde (1999).

[3] See Heckman (1992), Heckman and Smith (1993; 1995a), and Heckman, Smith, and LaLonde (1999) for statements of those assumptions.

[4] See Torp, et al. (1993), Heckman, Khoo, Roselius, and Smith (1996) and Heckman, Hohmann, Khoo, and Smith (1997).

difficulty in this paper by presenting the identifying assumptions that justify broad classes of estimators in a nonparametric setting, by testing the identifying assumptions using both nonparametric and semiparametric methods, by using two separate comparison groups drawn from different data sources, and by using a rich variety of conditioning variables.

In particular, we test the nonparametric identifying assumptions that justify three widely-used types of estimators for eliminating selection bias. The first type of estimator is the class of "index-sufficient" models introduced in Heckman (1980), which assumes that mean selection bias depends only on P, the probability of being selected into the program. The original parametric econometric models of selection bias are special cases of the index-sufficient model. We develop and apply a new test of index sufficiency and find support for this characterization of bias. However, the functional form of the index-sufficient selection bias that we estimate is different from that assumed in traditional econometric selection models. Regions of support where the selection bias for nonparticipants is negligible are required in order to use the index-sufficient selection estimator to construct the counterfactuals required to evaluate programs.[5,6] Such regions are not found in our data. To produce them requires a comprehensive sampling plan for collecting the data on comparison group members.

The second type of estimator whose identifying assumptions we test is the method of matching. It pairs participants and nonparticipants with common P values to estimate program impacts.[7] In general, matching is not guaranteed to reduce bias and may increase it (see Heckman and Siegelman (1993) and Heckman, LaLonde, and Smith (1999)). Moreover, matching is open to many of the same criticisms that have been directed against traditional econometric estimators because the method relies on arbitrary assumptions. Even with the rich data at our disposal, the method of matching is not, in general, an effective evaluation method. In our samples, it reduces but does not eliminate the conventional measure of selection bias. Matching eliminates bias *averaged* over certain intervals of P but does *not* eliminate pointwise bias in P. We demonstrate that this feature of the method is shared with the classical econometric selection model based on index sufficiency.

The third type of estimator whose identifying assumptions we test is an extension of the widely-used method of "difference-in-differences." Conditional on P, outcomes of participants before and after they participate in a program are differenced and differenced again with respect to before and after differences for members of the comparison group. The unconditional version of this estimator and its close cousin—the fixed effects estimator—are widely used.

[5] The supports of P are the domains of P with positive density.

[6] See Heckman (1990a) for a discussion of "identification at infinity," whereby parameters of interest can be identified from subgroups of individuals for whom there is no selection bias.

[7] The relationship between matching models based on P and classical selection models based on P was first discussed in Heckman and Robb (1986).

The assumptions required to justify the conditional version of this estimator are weaker than those required to justify matching. They are generally supported by our data. The effectiveness of the conditional difference-in-differences estimator is consistent with our evidence that the index-sufficient model characterizes bias. Since in our data selection bias as a function of P is constant over time for most values of P, it can be differenced out.

A major finding of this study is that the empirical distribution of P for program participants is very different from the distribution of P for members of the comparison group. Not only are the shapes of the empirical distributions different over regions of common support, but the supports differ as well. Conventional measures of selection bias employed by Ashenfelter (1978), LaLonde (1986), and Heckman and Hotz (1989) do not distinguish the bias arising from comparing participants and nonparticipants at the same P values from the bias arising from comparing persons at different P values.

We present a new decomposition of the conventional measure of selection bias that isolates these conceptually distinct sources of bias. We find broad regions of P values over which the difference between the outcomes of participants and nonparticipants conditional on a particular value of P is not defined because the supports of the distributions of participants and nonparticipants do not overlap. Comparing incomparable people contributes substantially to selection bias as conventionally measured. This finding, in conjunction with our evidence that the impact of the program measured in the region of common support differs from the overall impact of the program, reveals an important limitation of all nonexperimental methods for evaluating social programs. Even when these methods solve the selection problem, they can only identify the effect of treatment for participants who have counterparts in the comparison group.

Our discovery of the empirical importance of imposing a common support condition in reducing bias as conventionally measured demonstrates the benefit of the nonparametric approach to econometrics. Rigorous application of nonparametric methods entails careful specification of the domain over which estimators can be identified and consistently estimated.

This paper also shows the value of having good data. We show that access to a geographically-matched comparison group administered the same questionnaire as program participants and access to detailed information on recent labor force status histories and recent earnings are essential in constructing comparison groups that have outcomes close to those of an experimental control group. Data and method both matter in devising effective nonexperimental estimators of program impacts.

In the concluding sections, we discuss how to extend and apply the methods analyzed in this paper to analyze the effect of treatment on the treated in the more common situation where analysts do not have access to experimental data. Two of the three methods require no modification. The semiparametric selection bias estimator requires additional exclusion restrictions when applied to ordinary observational data.

2. THE EVALUATION PROBLEM, THE PARAMETER OF INTEREST IN THIS PAPER AND HOW RANDOMIZATION ESTIMATES IT

Following Fisher (1935), Roy (1951), and Quandt (1972), we assume that each person has two possible outcomes, Y_0 and Y_1, in the untreated and treated states, respectively. Let $D = 1$ signify receipt of treatment and $D = 0$ its absence. General equilibrium effects are ignored so that the outcomes for any person do not depend on the overall level of participation in the program.[8]

The problem of program evaluation arises because we observe only Y_0 or Y_1 for each person, but never both. That is, we observe Y where $Y = DY_1 + (1 - D)Y_0$. Thus we cannot form the gross gain $\Delta = Y_1 - Y_0$ for anyone. In the standard evaluation problem, analysts have access to participant records and to data on a comparison group of nonparticipants. Hence, one can construct the conditional distribution of Y_1 given a vector of conditioning variables X and $D = 1$, and the conditional distribution of Y_0 given X and $D = 0$, and can consistently estimate $\Pr(D = 1 \mid X) = P(X)$.[9]

This paper only considers the evaluation problem for mean impacts.[10] We focus on the parameter that receives the most attention in the evaluation literature: the effect of treatment on the treated, defined as

$$\Delta(X) = E(\Delta \mid X, D = 1) = E(Y_1 \mid X, D = 1) - E(Y_0 \mid X, D = 1), \tag{1}$$

or an averaged version for X in some region K,

$$\bar{\Delta}(K) = \int_K \Delta(X)\, dF(X \mid D = 1) \Big/ \int_K dF(X \mid D = 1). \tag{2}$$

The average impact parameter is the focus of many evaluation studies, especially those based on the method of matching. Other aspects of a program may also be interesting, but parameters (1) and (2) are useful in evaluating the gross benefit of an existing program—the main ingredient required to make a decision to continue it or shut it down.[11,12]

To make the parameter (1) clearly interpretable, we require that the conditional distribution of X satisfy $F(X \mid Y_0, Y_1, D) = F(X \mid Y_0, Y_1)$, i.e. that condi-

[8] Lewis (1963) discusses the failure of this assumption in the context of evaluating the effects of unionism on wages. This assumption is relaxed in an evaluation of skill promotion policies in Heckman, Lochner, and Taber (1997, 1998).

[9] Thus we do not consider the intrinsically more difficult evaluation problems considered by Marschak (1953) and Lancaster (1971), who consider forecasting the effects of policies never previously implemented (Marschak) or estimating the demand for goods never previously consumed (Lancaster).

[10] Heckman (1990b, 1992), Heckman, Smith, and Clements (1997; first draft 1993), and Heckman and Smith (1993, 1995a, 1998) consider the identification and estimation of distributions of impacts.

[11] See Heckman and Robb (1985), Heckman (1992), Moffitt (1992), Heckman (1997), Heckman and Smith (1993, 1995a, 1998), and Heckman, Smith, and Taber (1998) for discussions of alternative parameters of interest.

[12] In a cost-benefit analysis the other required ingredient is the cost. See, e.g., Heckman and Smith (1998).

tional on potential outcomes, realized D does not "cause" or predict X. This avoids the problem of conditioning on variables that are determined by D and hence masking the total effect of D. This condition is not strictly required but it simplifies the interpretation of our estimates. See Heckman, LaLonde, and Smith (1999) for further discussion.

Data on program participants identify $E(Y_1 \mid X, D = 1)$. Missing is the information required to identify $E(Y_0 \mid X, D = 1)$. The method of comparison groups uses data on nonparticipants to estimate it. The method assumes that, conditional on X, the outcomes of nonparticipants approximate what participants would have experienced had they not participated; that is, it assumes $E(Y_0 \mid X, D = 0) \cong E(Y_0 \mid X, D = 1)$. The selection bias, $B(X)$, associated with the program impact $E(\Delta \mid X, D = 1)$ that arises when this assumption fails to hold is

$$B(X) = E(Y_0 \mid X, D = 1) - E(Y_0 \mid X, D = 0). \tag{3}$$

Under certain conditions, the parameter of interest can be identified with data from a social experiment. If experiments do not disrupt the program being evaluated, and if control group members do not have access to close substitutes for the experimental treatment, then experimental data identify $E(Y_0 \mid X, D = 1)$. Thus $E(\Delta \mid X, D = 1)$ can be identified for any set of conditioning variables X within the support of X for $D = 1$ with data from a social experiment.[13] When it is valid, randomization avoids all of the traditional econometric problems of model selection. It avoids the need to specify the functional forms of the estimating equations that relate Y_1 and Y_0 to X, or to specify which variables are included in or excluded from outcome equations or program participation equations. This is an important advantage of randomization compared to other evaluation procedures.

3. CHARACTERIZING SELECTION BIAS

Since social experiments are costly, there is considerable interest in knowing if a nonexperimental strategy can be devised that produces estimates close to what would be produced from an ideal experiment on a prototypical job training program. This paper uses the data from the control group in a social experiment, together with unusually rich comparison group data collected under our supervision, to characterize the selection bias, $B(X)$, for different specifications of X. Knowledge of $B(X)$ is informative about the effectiveness of entire classes of selection bias correction methods. We now briefly describe the three types of estimators considered in this paper.

[13] Randomization is an instrumental variable that identifies parameters (1) and (2) even when all of the X are endogenous variables in the traditional sense of the term. See Heckman (1996) for an elaboration of this point.

3.1. *The Method of Matching*

To our knowledge, the method of matching was first used by Fechner (1860). It has been extensively applied to the evaluation of job training programs in studies conducted in the late 70's and early 80's.[14] The method is based on the identifying assumption that, conditional on some X, Y_0 is independent of D. In the notation of Dawid (1979), it assumes that

$$\text{(A-1)} \qquad Y_0 \perp\!\!\!\perp D \mid X, \qquad X \in \chi_c,$$

for some set χ_c, where "$\perp\!\!\!\perp$" denotes independence and variables to the right of "|" are the conditioning variables.[15] This assumption produces a comparison group that resembles the control group of an experiment in one key respect: conditional on X, the distribution of Y_0 given $D = 1$ is the same as the distribution of Y_0 given $D = 0$. In particular, when the means exist,

$$(4) \qquad E(Y_0 \mid X, D = 1) = E(Y_0 \mid X, D = 0),$$

so that *pointwise* in X, bias $B(X) = 0$.

Many matching estimators have been proposed that exploit (A-1) or its implication (4). Traditional matching methods pair nonparticipants with participants that are "close" in terms of X using different metrics.[16] For each observation i in the participant sample, a weighted average of comparison sample observations is formed to estimate the effect of treatment on i:

$$(5) \qquad Y_{1i} - \sum_{j \in \{D=0\}} W_{N_0 N_1}(i, j) Y_{0j}$$

where $\{D = 0\}$ is the set of indices for the nonparticipants and $\{D = 1\}$ is the set of indices for participants, N_0 is the number of observations in the comparison group, $\{D = 0\}$, N_1 is the number of observations in the treatment group, $\{D = 1\}$, and $\sum_{j \in \{D=0\}} W_{N_0 N_1}(i, j) = 1$ for all i.[17]

Matching estimators differ in the weights attached to members of the comparison group. Define a neighborhood $C(X_i)$ for each participant i. The persons matched to i are in A_i where $A_i = \{j \in \{D = 0\} \mid X_j \in C(X_i)\}$. Different matching methods use different neighborhoods. Nearest neighbor matching sets $C(X_i) = \{X_i \mid X_i = \min_i \| X_i - X_j \|, j \in \{D = 0\}\}$ where $\| \ \|$ is a norm, $W_{N_0 N_1}(i, j) =$

[14] See the detailed references to the historical literature in Heckman, LaLonde, and Smith (1999).

[15] A stronger version of (A-1) is usually stated: $(Y_0, Y_1) \perp\!\!\!\perp D \mid X$. Given our focus on parameters (1) and (2), this stronger version is not needed. The omitted assumption of conditional independence of Y_1 and D given X would be useful if we sought to evaluate the impact of lack of treatment on the untreated, $E(Y_0 - Y_1 \mid X, D = 0)$, using the outcomes of participants to proxy what nonparticipants would have earned had they participated. Note that we can estimate the impact of the program on a randomly selected person as a combination of the impact of treatment on the treated and treatment on the untreated.

[16] See Heckman, Ichimura, and Todd (1997, 1998; first drafts 1993) for a detailed discussion of alternative matching methods.

[17] The weights are allowed to depend on N_0 and N_1 to allow for use of an optimal bandwidth. See Heckman, Ichimura, and Todd (1996; first draft 1994).

1, $j \in A_i$, and $W_{N_0N_1}(i,j) = 0$ otherwise.[18] Nearest neighbors may be very far apart. For that reason a criterion must be imposed to ensure that the match is close in some sense. Caliper matching defines $C(X_i) = \{X_j \mid \|X_i - X_j\| < \varepsilon\}$ where ε is arbitrarily prespecified (see Cochrane and Rubin (1973)). If there is no such X_j, the observation i is not matched to any observations. If more than one person is in A_i, the nearest neighbor in terms of norm $\| \ \|$ is used to pick the match.

Kernel matching defines

$$W_{N_0N_1}(i,j) = \frac{G_{ij}}{\sum_{k \in \{D=0\}} G_{ik}},$$

where $G_{ik} = G((X_i - X_k)/a_{N_0})$ is a kernel that downweights distant observations from X_i and a_{N_0} is a sequence of smoothing parameters with the property that $\lim_{N_0 \to \infty} a_{N_0} = 0$. Nonzero values of this weight implicitly define $C(X_i)$ for this version of matching. In Section 5 of this paper, we extend kernel matching to permit regression adjustment of outcome equations. To estimate impacts over a set K as in (2), form a weighted sum of (5) over K:

$$\hat{M}(K) = \sum_{i \in \{D=1\}} \omega_{N_0N_1}(i) \left[Y_{1i} - \sum_{j \in \{D=0\}} W_{N_0N_1}(i,j) Y_{0j} \right] \quad \text{for} \quad X_i \in K, \tag{6}$$

where $\omega_{N_0N_1}(i)$ is a weight accounting for scale and possibly heteroskedasticity as well as the choice of support K.

Regression estimators have also been proposed that exploit (A-1), or its implication (4), in a linear regression setting. The econometric procedure of Barnow, Cain, and Goldberger (1980) assumes that Y_0 is linearly related to observables X and an unobservable U_0, so that $E(Y_0 \mid X, D = 0) = X\beta + E(U_0 \mid X, D = 0)$, and that $E(U_0 \mid X, D = 0) = E(U_0 \mid X)$ is linear in X. Under these assumptions, controlling for X via linear regression allows one to identify $E(Y_0 \mid X, D = 1)$ from the data on nonparticipants $E(Y_0 \mid X, D = 0)$. These functional form assumptions do not exploit the richness of assumption (4), which can be used to produce a nonparametric estimator of treatment effects using conditioning instead of projection or linear regression methods. Moreover, in practice, users of the method of Barnow, Cain, and Goldberger (1980) do not impose a common support condition in generating the estimates obtained from the method. The distribution of X may be very different in the $\{D = 0\}$ and $\{D = 1\}$ samples, so that comparability is only achieved by imposing linearity and extrapolating over different regions.

Recently, attention has focused on matching techniques that compare persons based on their probability of participation. Define the probability of participation or "propensity score" as $P(X) = \Pr(D = 1 \mid X)$. A theorem of Rosenbaum

[18] There are two versions of this method that differ depending on whether or not each comparison group observation may be matched to more than one participant observation. For expositional simplicity, we ignore ties.

and Rubin (1983) demonstrates that if (A-1) is satisfied, then

$$\text{(A-2)} \qquad Y_0 \perp\!\!\!\perp D \mid P(X) \quad \text{for} \quad X \in \chi_c,$$

provided $0 < P(X) < 1$ for $X \in \chi_c$, so that there is a positive probability that the events $D = 0$ and $D = 1$ occur for all elements in χ_c. Conditioning on $P(X)$ rather than on X produces conditional independence. An implication of (A-2), and not (A-2) itself, is all that is required to construct the desired counterfactual conditional mean. That implication is

$$\text{(7)} \qquad E(Y_0 \mid P(X), D = 1) - E(Y_0 \mid P(X), D = 0) = B(P(X)) = 0,$$

where (7) could be assumed directly in place of (A-2) or (A-1). Conditioning on $P(X)$ sets $B(P(X)) = 0$ and reduces the dimension of the matching problem down to matching on the scalar $P(X)$. Below we test condition (7) as a statistical hypothesis and reject it in our data.

Rosenbaum and Rubin (1983) assume that $P(X)$ is known rather than estimated. They do not present a distribution theory for the pointwise estimators of (1) or (2). Heckman, Ichimura, and Todd (1997, 1998; first drafts 1993) present the asymptotic distribution theory for the kernel matching estimator for the cases where P is known and where it is estimated.[19]

Comparison groups produced assuming (A-1) is valid differ from the control groups produced by a random experiment in an important way. Randomization equates the distributions of characteristics in the treatment and control groups. Without randomization, the distributions of characteristics in the treatment and comparison groups are not necessarily equated even if (A-1) is satisfied. The supports of the distributions of X may be different in the two groups and the shapes of the distributions may be different over regions of common support. Because counterparts to participants cannot always be found in the comparison group, estimators based on (A-1) or equation (7) do not necessarily identify treatment impacts for all values of X among program participants, unless the impacts do not depend on X.

A major advantage of the method of randomized trials over the method of matching in evaluating programs is that randomization works for any choice of X. In the method of matching, there is the same uncertainty about which X to use as there is in the specification of conventional econometric models. Even if one set of X values satisfies condition (A-1), an augmented or reduced version of this set may not. Heckman, Ichimura, and Todd (1997; first draft 1993) discuss tests that can be used to determine the appropriate choice of X variables. We discuss this problem in Section 4.3 below. Since nonparametric methods can be used to perform matching, the method does not, in principle, require that arbitrary functional forms be imposed to estimate program impacts.

[19] Heckman, Ichimura, and Todd (1998; first draft 1993) also answer the question, "If $P(X)$ were known would we match on it or on X?" Using the variance of the average impacts (2) as the choice criterion, the answer is "it depends."

3.2. *Index Sufficient Methods and the Classical Econometric Selection Model*

The traditional econometric approach to the selection problem adopts a more tightly-specified model relating outcomes to regressors X. This is in the spirit of much econometric work that builds models to estimate a variety of counterfactual states, rather than just the single counterfactual required to estimate the mean impact of treatment on the treated, the parameter of interest in most applications of the methods of matching or random assignment.[20] In the simplest econometric approach, two functions are postulated: $Y_1 = g_1(X, U_1)$ and $Y_0 = g_0(X, U_0)$, where U_0 and U_1 are unobservables. A selection equation is specified to determine which outcome is observed. Separability between X and (U_0, U_1) is assumed, so that

$$(8) \qquad Y_1 = g_1(X) + U_1 \quad \text{and} \quad Y_0 = g_0(X) + U_0,$$

where $E(U_1) = E(U_0) = 0$. This assumption *defines* functions called structural functions that do not depend on unobserved variables. In this notation, the parameter of interest defined in (1) becomes

$$(9) \qquad E(\Delta \mid X, D = 1) = g_1(X) - g_0(X) + E(U_1 - U_0 \mid X, D = 1).$$

Parameter (9) is an unconventional object for an econometric investigation. It combines the $g_1(X)$ and $g_0(X)$ functions that are the usual objects of econometric interest with the conditional mean of the difference in unobservables $E(U_1 - U_0 \mid X, D = 1)$.

Much applied econometric activity is devoted to eliminating the mean effect of unobservables on estimates of functions like g_0 and g_1. However, the mean difference in unobservables is an essential component of the definition of the parameter of interest in evaluating social programs.[21] In the traditional separable framework, the selection bias that arises from using a nonexperimental comparison group is

$$(10) \qquad B(X) = E(U_0 \mid X, D = 1) - E(U_0 \mid X, D = 0).$$

In the standard evaluation problem, the goal is to set $B(X) = 0$, *not* to eliminate dependence between (U_0, U_1) and X. The X can fail to be exogenous and parameters (1) and (2) can still be identified.

The conventional economic approach partitions the observed variables X into two not necessarily disjoint sets (R, Z) corresponding to those in the outcome equations and those in the participation equation, and postulates exclusion

[20] This emphasis on econometric models as devices to generate a variety of counterfactuals can be traced back to Haavelmo (1944) or Marschak (1953).

[21] If $U_1 = U_0$, as is assumed in the dummy endogenous variable model, then $E(U_1 - U_0 \mid X, D = 1) = 0$. If $U_1 - U_0$ is not forecastable with respect to X and $D = 1$ at the time the decision to participate in the program is made, then $E(U_1 - U_0 \mid X, D = 1) = 0$. See Heckman (1992, 1996, 1997) and Heckman and Smith (1993, 1996, 1998). The model $Y = Y_1 D + Y_0(1 - D) = g_0(X) + [g_1(X) - g_0(X) + U_1 - U_0]D + U_0$ is a model with a random coefficient on D.

restrictions. Thus it is assumed that certain variables appear in Z but not in R. The conventional approach further restricts the model so that the bias $B(X)$ only depends on Z through a scalar index. Note that exclusion restrictions are neither required nor used to justify matching as an estimator of (1) or (2).[22]

The latent index variable model with index I motivates the characterization of bias as a function of a scalar index. Define $I = H(Z) - \nu$ where $H(Z)$ is the mean difference in utilities or discounted earnings between the participation and nonparticipation states and ν is assumed to be independent of Z.[23] Then $D = 1$ if $I > 0$ and $D = 0$ otherwise, so that $\Pr(D = 1 \mid Z) = F_\nu(H(Z))$. The conventional econometric selection model further assumes that the dependence between D and (U_0, U_1) that gives rise to bias (10) arises only through ν and that R and Z are independent of (U_0, U_1). This implies that

$$E(U_0 \mid Z, R, D = 1) = E(U_0 \mid \nu < H(Z)),$$
$$E(U_0 \mid Z, R, D = 0) = E(U_0 \mid \nu \geq H(Z)),$$
$$E(U_1 \mid Z, R, D = 1) = E(U_1 \mid \nu < H(Z)), \quad \text{and}$$
$$E(U_1 \mid Z, R, D = 0) = E(U_1 \mid \nu \geq H(Z)).$$

Therefore, both $B(Z)$ and the mean gain of the unobservables, $E(U_1 - U_0 \mid Z, R, D = 1)$, depend on Z only through the index $H(Z)$. When F_ν is assumed to be strictly monotonic almost everywhere, we may write $H(Z) = F_\nu^{-1}(\Pr(D = 1 \mid Z))$ and the bias and mean gain terms depend on Z solely through P. The bias is

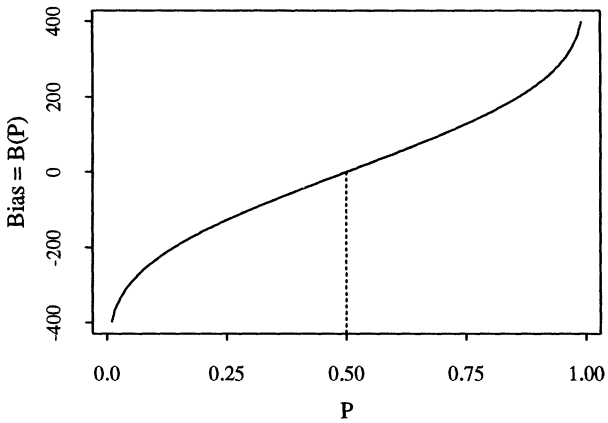
$$B(P(Z)) = E(U_0 \mid P(Z), D = 1) - E(U_0 \mid P(Z), D = 0). \tag{11}$$

This is the "index sufficient" representation where $P(Z)$, or equivalently $H(Z)$, is the index.[24] Conventional econometric models (see, e.g., Amemiya (1985)) assume that the latent variables ν and U_0 are symmetrically distributed around zero, so that $B(P(Z))$ is symmetric around $P = \frac{1}{2}$. Figure 1 presents an example of a normal selection model. If P itself is symmetrically distributed around $P = \frac{1}{2}$, the *average bias* over symmetric intervals around that value is zero even though the pointwise bias is nonzero. Thus, the classical selection model sometimes justifies matching as a consistent estimator of parameter (2) over intervals of P where the bias cancels out. To test the index sufficient model, we use our pooled sample of controls and comparison group members to determine if the estimated bias is solely a function of $P(Z)$ for different sets of variables Z, or if a more general conditioning set (R, Z) is required to characterize the bias.

[22] Heckman, Ichimura, and Todd (1998; first draft 1993) extend the theory of matching to consider separable models and models with exclusion restrictions and discuss the efficiency gains from using such restrictions. Exclusion restrictions are natural in the context of panel data models where the variables in the outcome equation are measured in periods after the decision to participate in the program is made.

[23] Absolute continuity of ν is often assumed although technically it is not required.

[24] This argument is due to Heckman (1980). If there are multiple decision rules for admission into the program, then a multiple index model is required. See Heckman and Robb (1985).



Note: This is the index model introduced in Section 3.2 where ν and U_0 are assumed to be normal and $\sigma_{\nu} = 1$, $\sigma_{U_0} = 375$, and $\rho = \text{cov}(U_0, \nu)/\sigma_{U_0} = 0.16$.

FIGURE 1.—Prototypical selection model, normal example; $B(P(X)) = E(U_0 \mid P(X), D = 1) - E(U_0 \mid P(X), D = 0)$.

Index sufficiency is only a necessary condition for applying the classical index sufficient selection model in a nonparametric or semiparametric setting. As noted by Heckman (1990a), it is also necessary to know a point or interval of P where $E(U_0 \mid P(X), D = 0) = 0$. Unless this condition is satisfied, it is not possible to use the index-sufficient selection model to construct the required counterfactual.[25] Thus in order to implement this method, it is necessary (a) that such a point or interval exist and (b) that it be possible to discover it.

The traditional selection-correction method parameterizes the bias function $B(P(Z))$ and eliminates bias by estimating $B(P(Z))$ along with the other

[25] To see why this condition is necessary, suppose that $Y_0 = \beta_0 + U_0$ and that index sufficiency holds. Then $E(Y_0 \mid X, D = 0) = \beta_0 + E(U_0 \mid P(X), D = 0)$. To construct $E(Y_0 \mid X, D = 1)$, the classical selection bias model requires that $E(U_0) = 0$ and that β_0 be identified along with $E(U_0 \mid P(X), D = 0)$. Then using the fact that

$$E(U_0) = E(U_0 \mid P(X), D = 1)P(X) + E(U_0 \mid P(X), D = 0)(1 - P(X)) = 0,$$

it follows that

$$E(U_0 \mid P(X), D = 1) = -\frac{1 - P(X)}{P(X)} E(U_0 \mid P(X), D = 0).$$

To use this result to construct $E(Y_0 \mid X, D = 0)$ nonparametrically, it is necessary to know β_0. If this is known, then $E(Y_0 \mid X, D = 0) - \beta_0 = E(U_0 \mid P(X), D = 0)$, and it is possible to construct

$$E(Y_0 \mid X, D = 1) = -\frac{1 - P(X)}{P(X)} E(Y_0 \mid X, D = 0) + \frac{\beta_0}{P(X)}.$$

Heckman (1990a) shows that β_0 is identified only if there is a set of values X such that $E(U_0 \mid P(X), D = 0) = 0$. If there is no such set, then one cannot separate a constant associated with $E(U_0 \mid P(X), D = 0)$ from β_0.

parameters of the model.[26] Heckman and Robb (1985, 1986) term the dependence between U_0 and D operating through the ν "selection on unobservables" while the dependence between U_0 and D operating through dependence between Z and U_0 is termed "selection on observables." In their framework, the method of matching assumes selection on observables, because conditioning on Z controls the dependence between D and U_0, producing a counterpart to (4) for the residuals: $E(U_0 \mid Z, D = 1) = E(U_0 \mid Z, D = 0)$. When selection is on unobservables, it is impossible to condition on ν and eliminate the selection bias. Thus the choice of an appropriate econometric model critically depends on the properties of the data on which it is applied.

3.3. *Difference-in-Differences*

The classical before-after estimator compares the outcomes of participants after they participate in the program with their outcomes before they participate. With the difference-in-differences estimator, common time and age trends are eliminated by subtracting the before-after change in nonparticipant outcomes from the before-after change for participant outcomes. This method can be generalized to include regressors.[27] The simplest application of the method does not condition on X and forms simple averages over the treatment and comparison groups.

In this paper, we introduce conditional semiparametric and nonparametric versions of the difference-in-differences estimator to a panel or to repeated cross sections of persons. Differencing is done conditional on X. The critical identifying assumption in our proposed method is that conditional on X, the biases are the same on average in different time periods before and after the period of participation in the program so that differencing the differences between participants and nonparticipants eliminates the bias.

To see how this estimator works, let t be a post-program period and t' a preprogram period. The method identifies parameters (1) and (2) conditional on X under the assumption

$$B_t(X) - B_{t'}(X) = 0, \qquad \text{for some } t, t', \tag{12}$$

where B_t denotes the bias in time t, defined in (10). This method extends the method of matching because it does not require that the bias vanish for any X, just that it be the same for some t and t' conditional on X. Notice further that (12) is implied by the conventional econometric selection estimator if $E(U_{0t} \mid P(X), D = 1) - E(U_{0t'} \mid P(X), D = 1)$ is the same for some choice of t and t'. In application, (12) is often assumed to hold for all t and t' or for t and t' defined

[26] Heckman and Robb (1985), Heckman (1990a), and Cosslett (1991) discuss this strategy in a semiparametric model.

[27] Heckman and Robb (1985, p. 218) discuss the difference-in-differences estimator and demonstrate that it can be implemented using repeated cross-section data. They also present economic models that justify its use. See also Heckman, LaLonde, and Smith (1999).

symmetrically around $t = 0$, the date of participation in the program (i.e., $t = -t'$).

We now compare $B(X)$ to the more conventional measure of bias used in the literature.

4. RE-EXAMINING THE CONVENTIONAL MEASURE OF SELECTION BIAS

The selection bias measure $B(X)$ is rigorously defined only over the set of X values common to the $D = 1$ and $D = 0$ populations. Define $S_{1X} = \{X \mid f(X \mid D = 1) > 0\}$ to be the support of X for $D = 1$, where $f(X \mid D = 1)$ is the conditional density of X given $D = 1$. Let $S_{0X} = \{X \mid f(X \mid D = 0) > 0\}$ be the support of X for $D = 0$ and let $S_X = S_{0X} \cap S_{1X}$ denote the region of overlap. Using the X distribution of participants, we define the mean selection bias $\overline{B}_{S_X}$ as

$$\overline{B}_{S_X} = \frac{\int_{S_X} B(X)\, dF(X \mid D = 1)}{\int_{S_X} dF(X \mid D = 1)}.$$

A comparable definition of $\overline{B}_{S_P}$ replaces X with $P(X)$ in the definition of $\overline{B}_{S_X}$. The conventional measure of selection bias $B = E(Y_0 \mid D = 1) - E(Y_0 \mid D = 0)$ used by LaLonde (1986) and others does not condition on X.

The conventional measure of bias B can be decomposed into a portion corresponding to a properly-weighted average of $B(X)$ and two other components.[28] First note that

$$B = \int_{S_{1X}} E(Y_0 \mid X, D = 1)\, dF(X \mid D = 1) - \int_{S_{0X}} E(Y_0 \mid X, D = 0)\, dF(X \mid D = 0). \tag{13}$$

Decompose B into three terms:

$$B = B_1 + B_2 + B_3, \tag{14}$$

where

$$B_1 = \int_{S_{1X} \setminus S_X} E(Y_0 \mid X, D = 1)\, dF(X \mid D = 1) - \int_{S_{0X} \setminus S_X} E(Y_0 \mid X, D = 0)\, dF(X \mid D = 0),$$

$$B_2 = \int_{S_X} E(Y_0 \mid X, D = 0)[dF(X \mid D = 1) - dF(X \mid D = 0)], \quad \text{and}$$

$$B_3 = P_X \overline{B}_{S_X},$$

[28] One can place the conventional method in a regression framework. Run a least squares regression of Y_0 on D, with $Y_0 = \pi_0 + \pi_1 D + \tau$, and $E(\tau) = 0$. Then plim $\hat{\pi}_1 = B$ as long as a law of large numbers is valid for the (Y_0, D) data sequence.

where $P_X = \int_{S_X} dF(X \mid D = 1)$ is the proportion of the density of X given $D = 1$ in the overlap set S_X, $S_{1X} \setminus S_X$ is the support of X given $D = 1$ that is not in the overlap set S_X, and $S_{0X} \setminus S_X$ is the support of X given $D = 0$ that is not in the overlap set S_X.

Term B_1 in (14) arises when $S_{0X} \setminus S_X$ or $S_{1X} \setminus S_X$ is nonempty. In this case we fail to find counterparts to $E(Y_0 \mid X, D = 1)$ in the set $S_{0X} \setminus S_X$ and counterparts to $E(Y_0 \mid X, D = 0)$ in the set $S_{1X} \setminus S_X$. Term B_2 arises from the differential weighting of $E(Y_0 \mid X, D = 0)$ by the two densities for X given $D = 1$ and $D = 0$ within the overlap set. Term B_3 arises from differences in outcomes that remain even after controlling for observable differences. Selection bias, rigorously defined as $\overline{B}_{S_X}$, may be of a different magnitude and even a different sign than the conventional measure of bias B.

Matching methods that impose the condition of pointwise common support eliminate two of the three sources of bias in (14). Matching only over the common support necessarily eliminates the bias arising from regions of nonoverlapping support given by term B_1 in (14). The bias due to different density weighting is eliminated because matching on participant P values effectively reweights the nonparticipant data. Thus $P_X \overline{B}_{S_X}$ is the only component of (14) that is not eliminated by matching.[29] $\overline{B}_{S_X}$ is the bias associated with a matching estimator.

4.1. *Examining the Validity of Matching on P*

We examine the validity of matching on $P(X)$ by estimating the three components of the bias B. If matching is valid, the third component of the decomposition should be negligible for each value of $P(X)$. Form the orthogonal decomposition of the conditional mean given X into two components: $E(Y_0 \mid X, D = 1) = E(Y_0 \mid P(X), D = 1) + V$ where $V = E(Y_0 \mid X, D = 1) - E(Y_0 \mid P(X), D = 1)$ and $E(V \mid P(X), D = 1) = 0$. Heckman, Ichimura, and Todd (1997, 1998; first drafts 1993) show that constructing the mean conditional on $P(X)$ permits consistent, but possibly inefficient, estimation of the terms in decomposition (14). The conditional means are integrated against the empirical counterparts of the conditional distributions for $P(X)$, $F(P(X) \mid D = 1)$, and $F(P(X) \mid D = 0)$, i.e., the means are self-weighting.

Before presenting our estimates of the components of (14), we describe the data used to generate them and the variables Z that best predict participation in the program.

4.2. *Our Data*

The data used in this study come from four training centers participating in a randomized evaluation of the Job Training Partnership Act (JTPA).[30] Along

[29] Since B_1 and B_2 may be of any sign, the matching estimator may have a bias component bigger than B.

[30] See Orr, et al. (1995) for a description of the National JTPA Study.

with data on the experimental treatment and control groups, information was collected on a nonexperimental comparison group of persons located in the same four labor markets who were eligible for the program but chose not to participate in it at the time random assignment was conducted. These persons are termed ENPs—for eligible nonparticipants.[31]

Random assignment took place at the point where individuals had applied to and been accepted into JTPA (i.e., admitted by a JTPA administrator). Under ideal conditions, randomization at this point identifies parameters (1) and (2). Members of the control group were excluded from receiving JTPA services for 18 months after random assignment. The controls completed the same survey instrument as the ENP comparison group members.[32] This instrument included detailed retrospective questions on labor force participation, job spells, earnings, marital status, and other characteristics. In this paper, we analyze a sample of adult males age 22 to 54. Table I defines the variables used in this study. Appendix B describes the data more fully and gives summary statistics for our sample.

4.3. *Determining the Probability of Program Participation P*

The participation probability $P(X)$ plays a central role in our analysis. In this paper, participation means that a person applies and is accepted into the program. Heckman and Smith (1995b) find that for all groups, including adult males, recent (past six months) labor force status transitions, not the pre-program earnings dip emphasized by Ashenfelter (1978), are the key predictors of participation. The relative participation rates presented in the fifth column of Table II demonstrate this point. Persons recently entering unemployment are the most likely to seek to participate in the program. Participation in job training is a form of job search for many unemployed workers. Earnings at the time of the participation decision are an important secondary predictor of participation.

Table III presents the estimated coefficients of the logit model $P(X)$. Variables are included in the model on the basis of two criteria: (a) minimization of classification error when $\hat{P}(X) > P_c$ is used to predict $D = 1$ and $\hat{P}(X) \leq P_c$ is used to predict $D = 0$, where $P_c = E(D)$; and (b) statistical significance of the included regressors. For adult males, the two criteria produce the same model. See Appendix C for a more extensive discussion of the variable selection criteria used in this paper.

Figure 2 presents the distributions of the estimated $P(X)$ in the $\{D = 0\}$ and $\{D = 1\}$ groups. We obtain similar distributions for $P(X)$ using alternative sets of regressors.[33] This figure indicates the potential importance of defining bias on

[31] See Smith (1994) and Appendix B for descriptions of the ENP sample.

[32] Treatment group members did not complete the long baseline survey instrument administered to the controls and ENPs, and so cannot be used in the estimation of the participation model.

[33] These results are available on request from the authors.

TABLE I

DEFINITION OF VARIABLES

Variable Name	Description
Training Center: Corpus Christi, Fort Wayne, Jersey City, Providence.	Indicator variables for the geographic location of the individual.
Race and Ethnicity: black, white, Hispanic.	Indicator variables for the race/ethnicity of the individual. Individuals who reported Asian or "other" were included in the Hispanic category in R but not in Z.
Age: age 22–29, age 30–39, age 40–49, age 50–54.	Indicator variables for the age of the individual calculated using the average age in years of the individual within the quarter of the observation.
Education: less than 10th grade, 10–11th grade, 12th grade, 1–3 years of college, 4 or more years of college.	Indicator variables for the educational attainment of the individual at the time of random assignment or eligibility determination. Missing values are imputed.[a]
Marital Status: currently married, last married 1–12 months before RA/EL, last married > 12 months before RA/EL, single, never married at RA/EL.	Indicator variables for marital status at the time of random assignment or eligibility determination (RA/EL). Missing values are imputed.[a]
Children less than 6 years of age	Indicator variable for the presence of young children in the household at the time of the baseline interview. Missing values are imputed.[a]
Calendar Quarter: quarter 1, quarter 2, quarter 3, quarter 4.	Indicator variables for the calendar quarter for the observations. Quarter 1 refers to January, February, and March etc. If an observation overlaps two quarters, then the variable takes on fractional values.
Calender Year: year 1987, year 1988, year 1989, year 1990.	Indicator variables for the calendar year of the observation. If the observation overlaps two years, then the year indicators take on fractional values.
Local Unemployment Rate (Sources: U.S. Department of Labor's publication "Labor Force, Employment, and Unemployment Estimates for States, Labor Market Areas, Counties, and Selected Cities" for the years 1986–1991 provide the unemployment rates. Population weights are obtained from annual total population data available in the U.S., Department of Commerce's Regional Economic Information System (REIS)).	This variable gives the monthly unemployment rate. The data are published at the county and metropolitan area levels. We calculate the unemployment rate as a population-weighted average of the unemployment rates of the counties and metropolitan areas served by each of the four training centers in the JTPA data.

TABLE I—*Continued*

Variable Name	Description
Labor Force Status Transition: employed → employed, unemployed → employed, OLF → employed, employed → unemployed, unemployed → unemployed, OLF → unemployed, employed → OLF, unemployed → OLF, OLF → OLF.	The two most recent labor force statuses during the period composed of the month of random assignment or eligibility determination and the six preceding months define a set of nine labor force status patterns. In each case, the second status is that in the month of random assignment or eligibility determination and the first status (if different) is the most recent preceding status. Repeated patterns such as "employed → employed" indicate persons in the same labor force status for all seven months. Missing values are imputed.[a]
Number of Persons in the Household	Continuous variable indicating the number of persons in the individual's household as of the baseline interview. Missing values are imputed.[a]
Earnings in the Month of Random Assignment or Eligibility Determination	Self-reported monthly earnings in the month of random assignment or eligibility determination from the baseline survey. Persons for whom the survey covers only a part of the month have their responses scaled up to a full month.
Ever had Vocational Training	Indicator variable for whether the respondent ever had vocational or technical training as of the baseline interview date, excluding courses taken while in high school. Missing values are imputed.[a]
Currently Receiving Vocational Training	Indicator variable for current receipt of vocational or technical training as of the baseline interview. Excludes courses taken in high school. Missing values are imputed.[a]
Number of Job Spells in the 18 Months Prior to Random Assignment or Eligibility Determination: zero, one, two, more than two.	Categories for the number of full or partial job (not employment) spells experienced during the 18 months prior to random assignment or eligibility determination. Missing values are imputed.[a]
Work Experience	Continuous variable indicating months of work experience prior to random assignment or eligibility determination. It is calculated using the Mincer method, (age-education-6)*12, for the period prior to our data, adding in actual experience in months for the five years prior to RA/EL.

[a]An appendix available upon request from the authors describes the imputation procedure for these variables.

TABLE II

ESTIMATED BIAS BY LABOR FORCE STATUS TRANSITION CELLS AND THE PROBABILITY OF PARTICIPATION AND ITS LOGIT BY LABOR FORCE STATUS, CRUDE RATES, AND RATES IMPLIED BY LOGIT

Quarterly Earnings Expressed in Monthly Dollars, Experimental Control and Eligible Nonparticipant (ENP) Samples, Adult Males, 508 Controls and 388 ENPs

Cell	Percentage of Controls in Cell	Percentage of ENPS in Cell	Estimated Bias in Cell[a]	Difference in Population Program Participation Rates[b]	Average Derivative from Logit[b]	Coefficient From Logit[b]	Difference in Logits of Program Participation Rates[b]
Employed → Employed	21.16	73.41	−421 (78)	*	*	*	*
Unemployed → Employed	10.79	4.16	−474 (228)	0.038	0.033	1.518 (0.416)	1.703
OLF → Employed	4.77	1.11	−128 (327)	0.061	0.012	0.787 (0.761)	2.139
Employed → Unemployed	27.39	6.65	−148 (188)	0.158	0.079	2.465 (0.456)	3.113
Unemployed → Unemployed	17.43	4.16	420 (69)	0.105	0.093	2.668 (0.583)	2.671
OLF → Unemployed	5.81	0.55	230 (207)	0.109	0.142	3.272 (0.597)	2.715
Employed → OLF	5.60	1.94	398 (209)	0.119	0.085	2.552 (0.608)	2.801
Unemployed → OLF	1.45	1.39	381 (92)	0.072	0.069	2.302 (0.875)	2.298
OLF → OLF	5.60	6.65	297 (157)	0.017	−0.002	−0.154 (0.609)	1.090

[a]This column gives the mean difference in monthly earnings of the experimental controls and eligible nonparticipants conditional on labor force status transition patterns in the six months prior to random assignment (see Table I for the definition of the labor force transition categories). The mean is calculated over the 18 months after the date of random assignment/eligibility determination.

[b]These columns give differences in the population participation rates and in the logits of the population participation rate, relative to the Employed → Employed cell.

a common support of $P(X)$. For the sample of controls, the histogram of $P(X)$ values has support over the entire $[0, 1]$ interval. Surprisingly, however, the mode of the distribution of $P(X)$ for controls is near zero. Many controls have a low estimated probability of participation. In the sample of ENPs, the support of $P(X)$ is concentrated in the interval $[0, 0.225]$. Thus, the bias measure $\overline{B}_{S_P}$, which is the bias defined conditional on $P(X)$ rather than X, is defined only over a fairly limited interval. As a result of this restriction on the support, any nonexperimental evaluation can nonparametrically estimate program impacts defined only over this interval. As we demonstrate below, the difference between the distributions of the estimated values of P has important implications for understanding the sources of selection bias as conventionally measured. Before presenting this decomposition, we first develop some econometric tools that are used to generate many of the empirical results reported in this paper.

TABLE III

COEFFICIENT ESTIMATES AND p VALUES FROM WEIGHTED PARTICIPATION LOGIT[a]
BEST PREDICTOR MODEL FOR THE PROBABILITY OF PARTICIPATION[b]
Experimental Control and Eligible Nonparticipant (ENP) Samples
Dependent Variable: 1 for Experimental Control, 0 for Eligible Nonparticipant
Adult Males, 508 Controls and 388 ENPs

Variables	Coeff	Std Error	p Value[c]
Intercept	−5.07	0.83	0.0000
Fort Wayne, IN	2.45	0.41	0.0000
Jersey City, NJ	0.66	0.43	0.1273
Providence, RI	2.19	0.44	0.0000
Black	0.49	0.33	0.1333
Hispanic	0.43	0.40	0.2837
Other race/ethnicity	0.61	0.55	0.2653
Age 30 to 39	−0.50	0.30	0.0926
Age 40 to 49	−0.60	0.38	0.1115
Age 50 to 54	−0.29	0.62	0.6361
Fewer than 10 years schooling	−0.83	0.40	0.0397
10–11 years schooling	0.66	0.34	0.0510
13–15 years schooling	0.90	0.35	0.0096
16 or more years schooling	−1.38	0.54	0.0101
Last married 1–12 months prior to RA/EL[d]	0.42	0.80	0.5995
Last married > 12 months prior to RA/EL	−0.03	0.61	0.9648
Single, never married at RA/EL	0.71	0.36	0.0498
Child age less than 6 present in household	−0.16	0.38	0.6761
Unemployed → Employed	1.52	0.42	0.0003
OLF → Employed	0.79	0.76	0.3016
Employed → Unemployed	2.46	0.46	0.0000
Unemployed → Unemployed	2.67	0.58	0.0000
OLF → Unemployed	3.27	0.60	0.0000
Employed → OLF	2.55	0.61	0.0000
Unemployed → OLF	2.30	0.87	0.0085
OLF → OLF	−0.15	0.61	0.8002
One job in 18 months prior to RA/EL	0.41	0.39	0.2894
Two jobs in 18 months prior to RA/EL	0.57	0.50	0.2600
More than two jobs in 18 months prior to RA/EL	1.87	0.52	0.0003
Enrolled in vocational training at RA/EL	1.94	0.62	0.0019
Ever had vocational training?	−0.28	0.32	0.3815
Total number of household members	−0.25	0.10	0.0134
Earnings in the month of RA/EL	−0.00	0.00	0.0000

[a]Weights are used in the estimation procedure to account for choice-based sampled data. It is assumed that in a random sample Controls represent 3% and ENPs 97% of the eligible population.

[b]The omitted training center is Corpus Christi, TX; the omitted race is white; the omitted age group is 22–29; the omitted schooling category is twelve years; the omitted marital status is currently married at RA/EL; the omitted labor force transition pattern is Employed → Employed; the omitted number of job spells in the 18 months prior to RA/EL is zero.

[c]Reported p-values are for two-tailed tests of the null hypotheses that the true coefficient equals zero.

[d]RA/EL indicates the month of random assignment (RA) for the experimental controls and the month of eligibility (EL) for Eligible Nonparticipants (ENPs).

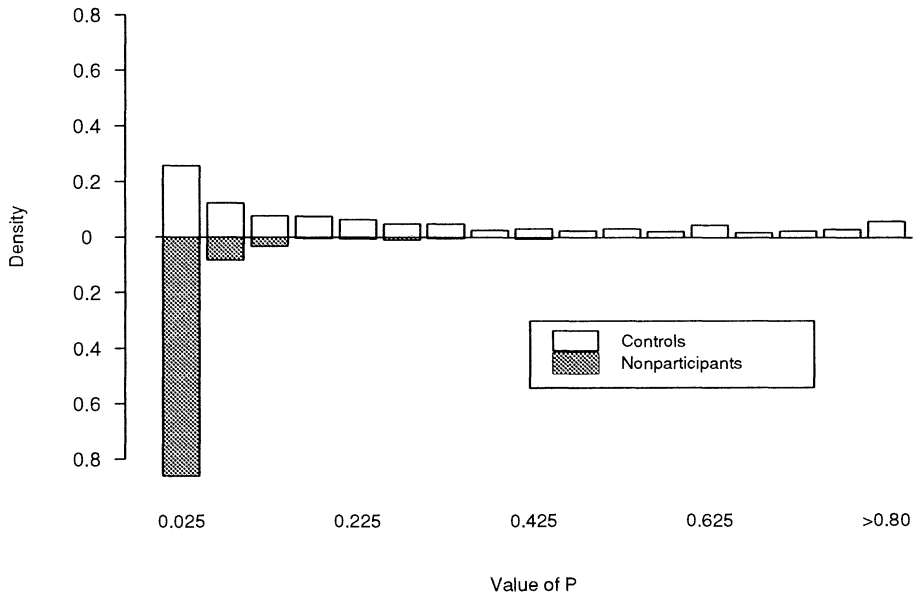


FIGURE 2.—Density of estimated probability of program participation for adult male controls and eligible nonparticipants.

5. NONPARAMETRIC TOOLS FOR ESTIMATING SELECTION BIAS $B(X)$ AND OTHER OBJECTS OF INTEREST

In an econometric sample selection model, the usual goal is to consistently estimate β in $Y_0 = X\beta + U_0$, where $E(Y_0 \mid X, D = 1) = X\beta + E(U_0 \mid X, D = 1)$ and $E(Y_0 \mid X, D = 0) = X\beta + E(U_0 \mid X, D = 0)$. In this paper, the goal is to estimate the bias $B(X) = E(U_0 \mid X, D = 1) - E(U_0 \mid X, D = 0)$ that arises from using a comparison group to identify the parameter $E(\Delta \mid X, D = 1)$.[34] A characterization of $B(X)$ suggests which nonexperimental strategies, if any, are likely to be effective in eliminating it. Our emphasis is thus very different from the standard approach that treats bias terms as nuisance functions to be eliminated.[35]

In the case where the X variables are all discrete, estimation of the bias is straightforward. Only cell means are required. The regression equation used to estimate the bias on comparison and control samples is

$$Y_0 = X\beta + E(U_0 \mid X, D = 0) + B(X)D + \varepsilon, \tag{15}$$

[34] In a context where the treatment impact and not the bias is being estimated, the methods we use can be applied directly by substituting data on Y_1 for participants for the data on Y_0 for controls. To apply the semiparametric index sufficient selection model (but not the other methods we consider) requires an exclusion restriction—some variable in Z not in R. We expand on this point below in Section 11.

[35] See, e.g., Heckman (1979), Cosslett (1991), or Ahn and Powell (1993).

where $B(X) = E(U_0 \mid X, D = 1) - E(U_0 \mid X, D = 0)$ and $E(\varepsilon \mid X, D) = 0$. $B(X)$ can be estimated from a least squares regression of Y_0 on a constant and D interacted with dummy variables for each X cell. The interactions between D and X identify $B(X)$ at the discrete coordinates of X even though β is not identified unless $E(U_0 \mid X, D = 0) = 0$, an assumption not required to identify $B(X)$. If conditioning on X eliminates bias, as is assumed in the method of matching or in the analysis of Barnow, Cain, and Goldberger (1980), then $B(X) = 0$ for each value of X.

A simple application of this method is presented in Table IV. We compute the mean bias within cells defined by a subset of the variables included in the logit for P. This subset and the cells themselves were chosen by cross-validation to minimize the sample misclassification rate using the "hit or miss" method described in Section 4.3 and using the Classification and Regression Tree (CART) method that partitions the data into the best-predicting groups.[36] Within cells, the bias $B(X)$ is large, just as it is in the fourth column of Table II. Averaging over cells using the cell weights for the $D = 1$ population, the estimated bias is much smaller. Thus, although the biases tend to cancel across cells, the method of matching per se is not justified by this partition of the data, nor is the method advocated by Barnow, Cain, and Goldberger (1980).

When $E(U_0 \mid X, D = 0)$ and $E(U_0 \mid X, D = 1)$ are specified more generally as nonparametric functions of *continuous* variables, equation (15) is termed the partial linear regression model.[37] In this paper we focus on nonparametric estimation of the $B(X)$, rather than on estimating the parametric portion of the model, and use the local linear regression methods described in detail in Appendix A.

Our data have a panel structure with individuals observed in periods $t = 1, \ldots, T$. Individuals are subscripted by "i." Define the bias functions as $K_{1t}(P_i) = E(U_{0it} \mid P_i, D_i = 1)$ and $K_{0t}(P_i) = E(U_{0it} \mid P_i, D_i = 0)$, and let $\varepsilon_{it} = U_{0it} - D_i K_{1t}(P_i) - (1 - D_i) K_{0t}(P_i)$ where $E(U_{0it}) = 0$. To conserve on notation we suppress the subscript "0" on Y_0 in the rest of this section and in Appendix A. Define $Y_i = (Y_{i1}, \ldots, Y_{iT})'$, $X_i = (X_{i1}, \ldots, X_{iT})'$, $K_j(P_i) = (K_{j1}(P_i), \ldots, K_{jT}(P_i))'$, $j = 0, 1$, and $\varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{iT})'$. Precise assumptions about ε are stated in Appendix A. In this notation, the seemingly-unrelated partial linear regression model used in this paper is

$$Y_i = X_i \beta + D_i K_1(P_i) + (1 - D_i) K_0(P_i) + \varepsilon_i. \tag{16}$$[38]

[36] The method of picking the best predictors is formalized as the CART method developed in Breiman, Friedman, Olshen, and Stone (1984). We use the CART algorithm in $S+$. See Chambers and Hastie (1993). The method described in Section 4.3 was applied using a parametric logit model. CART is a nonparametric method that searches for the best-predicting partitions of the data and explicitly considers interactions in constructing the model. In fitting the parametric logit model, we do not include interactions terms.

[37] See, e.g., Robinson (1988) and Hastie and Tibshirani (1990). For discrete X, the method used to estimate (15) is fully nonparametric.

[38] In Appendix A, we relax the restriction that β is constant across time periods. Robinson (1988) first proposed the partially linear model in the seemingly-unrelated regression framework.

TABLE IV

ESTIMATED BIAS FROM CLASSIFICATION TREE MODEL USING CROSS VALIDATION
Average Earnings in the 18 Months Following Random Assignment Expressed in Monthly Dollars
Experimental Control and Eligible Nonparticipant (ENP) Samples, Adult Males, 508 Controls and 388 ENPs

	Cell Characteristics								
Cell[a]	Labor Force Status	Earnings at RA/EL	Mean Earnings in 6 Months Prior to RA/EL	Site	#Job Spells	Education	Number of Controls	Number of ENPS	Estimated Bias
(1)	consistently unemployed or had labor force transition in 18 months prior to RA/EL	> $1050	...	Fort Wayne, IN and Jersey City, NJ	...	...	13	8	−608 (296)
(2)	same as (1)	> $1050	...	Providence, RI and Corpus Christi, TX	...	...	0	11	NA[b] (NA)
(3)	same as (1)	< $1050	...	...	...	...	343	54	158 (107)
(4)	consistently employed, consistently out of the labor force or missing	...	< $1008	Fort Wayne, IN and Jersey City, NJ and Providence, RI	0 or 1	> 3 years college or missing	1	14	43 (NA)
(5)	same as (4)	< $233	< $1008	same as (4)	0 or 1	< 10, 10–11, 12 or some college	41	19	222 (217)
(6)	same as (4)	> $233	< $1008	same as (4)	all levels	< 10, 10–11, 12 or some college	11	7	−90 (347)
(7)	same as (4)	...	< $1008	same as (4)	0 or 1	< 10, 10–11, or some college	6	20	−15 (178)
(8)	same as (4)	...	< $1008	same as (4)	2 or more	...	57	18	127 (154)
(9)	same as (4)	...	> $1008	same as (4)	...	...	44	135	−602 (114)
(10)	same as (4)	...	...	Corpus Christi, TX	...	...	20	138	−430 (157)
Average Cell Bias[c]									49 (74)

[a] Variables included in the CART analysis that were not selected as cells were age, average earnings for 12 months prior to random assignment or eligibility determination, number of household members, race, and current and past vocational training.

[b] A bias or standard error value of "NA" indicates that the cell contains only individuals of one type, either all controls or all eligible nonparticipants, so that the bias could not be calculated. If there is exactly one observation in a cell, then a bias can be calculated but a variance cannot (which is the case with cell (4)). A value of "..." indicates that the variable was not included as a cell conditioning variable so that all values of that variable are included in the cell.

[c] The average bias is obtained by a weighted mean of the cell bias values, using the control distribution across cells as the weights. "NA" cells are omitted in taking means.

Participants ($D = 1$) are oversampled in our data relative to their population proportions. We reweight the data to random sample proportions, and use the parametric logit model to estimate P_i.[39] In the general case, K_{0t} and K_{1t} are functions of more than just P_i. In the classical selection model and the extension of the matching method developed in Heckman, Ichimura, and Todd (1997, 1998; first drafts 1993), however, these functions depend only on P_i. The extension of the estimation method to vector-valued arguments for the K functions is straightforward.

We estimate the bias functions using the "double residual regression" method. Form expectations of equation (16) conditional on P_i and D_i to obtain

$$E(Y_i \mid P_i, D_i) = E(X_i \mid P_i, D_i)\beta + D_i K_1(P_i) + (1 - D_i)K_0(P_i).$$

Remove the portion of X_i and Y_i that depends on P_i and D_i (i.e., the conditional means) to form an adjusted version of (16):

$$Y_i - E(Y_i \mid P_i, D_i) = [X_i - E(X_i \mid P_i, D_i)]'\beta + \varepsilon_i. \tag{17}$$

Run "adjusted" least squares on this equation to estimate β.[40] The conditional expectations $E(Y_i \mid P_i, D_i)$ and $E(X_i \mid P_i, D_i)$ are consistently estimated using this method under conditions stated precisely in Appendix A.[41]

To estimate the components of the bias term $B(P(X))$, we use a local linear regression estimator applied to the X-adjusted residuals, $c_i = Y_i - X_i\hat{\beta}$, where $\hat{\beta}$ is estimated using the first stage procedure just described. The pointwise estimator of $K_d(P_0)$ in the neighborhood of P_0 is denoted $\hat{K}_d(P_0)$, where $\hat{K}_d(P_0)$ and $\hat{\gamma}_d(P_0)$ are defined as

$$\underset{K_d,\,\gamma_d}{\arg\min} \sum_{i \in \{D = d\}} \left[c_i - K_d(P_0) - \gamma_d(P_0)\left(\hat{P}_i - P_0\right)\right]^2 G\left(\frac{\hat{P}_i - P_0}{a_N}\right), \tag{18}$$

$$d \in \{0, 1\},$$

[39] The weights are given in the footnote to Table III. As is common in many evaluations (see the discussion and methods of solution in Heckman and Robb (1985)), persons in the $\{D = 1\}$ group are oversampled compared to persons in the $\{D = 0\}$ group. This gives rise to the problem of choice-based sampling. The problems raised by choice-based sampling are a special case of the problem of weighted distributions first analyzed by Rao (1965; 1986) and the solution is the same as his: weight the sampled distributions back to population proportions using population weights. Amemiya (1985) discusses applications of Rao's method in econometrics. Todd (1995) discusses estimation of the model in the text using nonparametric estimators for P. Her evidence suggests that estimation of P assuming a logit functional form is innocuous in our sample. Heckman, Ichimura, and Todd (1996; first draft 1994) show that the correction for choice-based sampling is strictly not required to estimate the bias functions. We reweight the data in order to derive estimates of the selection bias functions that are functions of P.

[40] The "adjusted" least squares trims out observations for which $f(P \mid D = 1)$ is too small. Such "trimming" is required to obtain uniform convergence of the estimator. See Appendix A for details and for the conditions required to secure consistency and asymptotic normality of β. Yatchew (1997) presents a simpler alternative estimator that avoids this first step procedure for estimating β.

[41] See Malinvaud (1970) for references on the origins of the double residual regression method. Robinson (1988) extends it to semiparametric models.

where P_0 is a given point in the support of $\hat{P}_i$ for $\{D = d\}$, G is a kernel with properties fully characterized in Appendix A, $\{a_N\}$ is a sequence of smoothing parameters, and $\hat{P}_i$ is the ith individual's estimated value of P. If $\gamma_d(P_0)$ is set to zero for all P_0, (18) becomes the standard kernel regression estimator. Introducing $\gamma_d(P_0)$ removes the linear bias term in the neighborhood of P_0, gives an estimator that is robust to the distribution of the regressors, and produces better boundary behavior than is produced using standard kernel regression. We account for the estimation of the parameters of P in deriving standard errors and test statistics. See Appendix A for further discussion.

The local linear regression method can be used to construct matches and to extend matching to regression-adjust for X. As demonstrated in Heckman, Ichimura, and Todd (1997; first draft 1993), local linear matching on P defines the $W_{N_0, N_1}(i, j)$ in (5) to be

$$W_{N_0 N_1}(i, j) = \frac{A - B}{C - D}, \tag{19}$$

where

$$A = G_{ij} \sum_{k \in \{D=0\}} G_{ik}(P_k - P_i)^2,$$

$$B = G_{ij}(P_j - P_i)\left[\sum_{k \in \{D=0\}} G_{ik}(P_k - P_i)\right],$$

$$C = \sum_{j \in \{D=0\}} G_{ij} \sum_{k \in \{D=0\}} G_{ik}(P_k - P_i)^2,$$

$$D = \sum_{k \in \{D=0\}} G_{ik}(P_k - P_i)^2, \quad \text{and}$$

$$G_{ik} = G\left(\frac{P_k - P_i}{a_N}\right).$$

This weight can be used to construct consistent pointwise estimators of (1) or averaged estimators of (2). Consistency and asymptotic normality of these estimators is established under conditions specified in Heckman, Ichimura, and Todd (1998; first draft 1993). Regression-adjusted local linear matching removes $X\beta$ from Y_0. Applied to participant and comparison group data, formula (5) or (6) is used with weights (19) and with $(Y_i - X_i \hat{\beta})$ in place of Y_i. The estimates $\hat{\beta}$ are obtained from the first stage estimator of equation (17).

6. ESTIMATING THE COMPONENTS OF OUR DECOMPOSITION OF B

We obtain nonparametric estimates of each of the components in (14) by decomposing our estimate of the bias $\hat{B}$ into the sample analogs of the three terms in (14) as follows:

$$\hat{B} = \hat{E}(Y_0 \mid D = 1) - \hat{E}(Y_0 \mid D = 0) = \hat{B}_1 + \hat{B}_2 + \hat{B}_3 \tag{20}$$

where

$$\hat{B}_1 = \frac{1}{N_1} \sum_{\substack{i \in \{D=1\} \\ P_i \in S_{1P} \setminus S_P}} Y_0(P_i) - \frac{1}{N_0} \sum_{\substack{i \in \{D=0\} \\ P_i \in S_{0P} \setminus S_P}} Y_0(P_i),$$

$$\hat{B}_2 = \frac{1}{N_1} \sum_{\substack{i \in \{D=1\} \\ P_i \in S_P}} \hat{E}(Y_{0i} \mid P_i, D=0) - \frac{1}{N_0} \sum_{\substack{i \in \{D=0\} \\ P_i \in S_P}} Y_0(P_i),$$

$$\hat{B}_3 = \frac{1}{N_1} \sum_{\substack{i \in \{D=1\} \\ P_i \in S_P}} \left[Y_0(P_i) - \hat{E}(Y_{0i} \mid P_i, D_i=0) \right],$$

where N_1 denotes the size of the $D=1$ sample, N_0 denotes the size of the $D=0$ sample, "^" indicates an estimate, $P_i = P(X_i)$ for person i, $Y_0(P_i)$ is the value of Y_{0i} for person i with probability P_i, where $S_P, S_{1P} \setminus S_P, S_{0P} \setminus S_P$ are analogous to $S_X, S_{1X} \setminus S_X$, and $S_{0X} \setminus S_X$ in (14) and where the counterfactual outcome in the no-treatment state for a $D=1$ observation with probability P_i, $E(Y_{0i} \mid P_i, D_i = 0)$, is estimated by a local linear regression of Y_{0i} on P_i using data on persons for whom $D=0$. Each term in the summations on the right-hand side of (20) is self-weighted by averaging over the empirical distribution of the P in either the $D=1$ or $D=0$ sample. Under random sampling, each term is consistently estimated and $\sqrt{N}$ times each term centered around its expected value is asymptotically normal.[42]

Following the analysis of the JTPA experiment reported in Bloom, et al. (1993), we use quarterly earnings and total earnings in the 18 months after random assignment as our outcome measures. Table V presents consistent and asymptotically normal estimates of the three components of decomposition (14) using the earnings data from the JTPA experiment and estimated using the formulas presented below equation (20). The control group sample gives information on Y_0 for those with $D=1$ and the sample of eligible nonparticipants gives Y_0 for those with $D=0$. The first column in Table V indicates the quarter (three month period) for which the estimates are constructed. These quarters are defined relative to the month of random assignment or eligibility determination. Each row corresponds to one quarter, with the bottom row reporting averages over the first six quarters (18 months) after random assignment. Column (1) reports the estimated mean selection bias $\hat{B}$. The next three columns report estimates of the components of the decomposition in (14). The top number in each cell is the estimate, the number in parentheses is the bootstrap standard error, and the number in square brackets is the percentage of $\hat{B}$ for the row that is attributable to the given component. The first component, $\hat{B}_1$, is presented in column (2) of the table. The component arising from misweighting of the data, $\hat{B}_2$, is given in column (3), and the component due to selection bias rigorously defined, $\hat{B}_3$, appears in column (4). Column (5)

[42] The asymptotic normality of each component is justified by Theorem A.1 of Appendix A.

TABLE V

DECOMPOSITION OF MEAN SELECTION BIAS FOR THE BEST PREDICTOR MODEL FOR THE PROBABILITY OF PROGRAM PARTICIPATION[a]

Experimental Control and Elig. Nonparticipant (ENP) Samples, Adult Males, 508 Controls and 388 ENPs

Quarter	(1) Mean Difference[b] ($\hat{B}$)	(2) Nonoverlap Support[c] ($\hat{B}_1$)	(3) Density Weighting ($\hat{B}_2$)	(4) Selection Bias ($\hat{B}_3$)	(5) Average Bias ($\hat{\bar{B}}_{S_P}$)	(6) Experimental Treatment Impact	(7) Average Bias ($\hat{B}_{S_P}$) as of % of Treatment Impact[d]
Qtr1	−420	190[−45%]	−627[149%]	17[−4%]	29	5	566%
	(38)	(31)	(32)	(34)	(63)	(30)	
Qtr2	−352	209[−59%]	−581[165%]	19[−6%]	32	37	88%
	(47)	(41)	(45)	(35)	(65)	(33)	
Qtr3	−343	221[−65%]	−576[168%]	12[−3%]	20	57	35%
	(55)	(39)	(50)	(43)	(79)	(34)	
Qtr4	−294	234[−80%]	−568[194%]	41[−14%]	68	60	114%
	(57)	(40)	(46)	(42)	(79)	(34)	
Qtr5	−311	232[−75%]	−576[185%]	33[−10%]	54	44	121%
	(57)	(40)	(51)	(41)	(77)	(35)	
Qtr6	−334	223[−67%]	−573[172%]	16[−5%]	27	61	44%
	(63)	(45)	(51)	(44)	(81)	(34)	
Average of 1 to 6	−342	218[−64%]	−584[170%]	23[−7%]	38	44	87%
	(47)	(38)	(41)	(33)	(63)	(14)	

[a]The best predictor model for the probability of program participation includes training center indicators, race, age, education, marital status, children aged less than 6, labor force status transitions, job spells, current and past vocational training, total number of household members, and earnings in the month of random assignment or eligibility determination. (See Table III for model estimates and Appendix C for prediction rate comparisons.)

[b]The percentage of the mean difference attributable to each component appears in square brackets in the appropriate column. Bootstrapped standard errors based on 50 replications with 100% sampling appear in parentheses.

[c]A 2% trimming rule was used in determining the overlapping support region, and a 0.06 fixed bandwidth was used for the nonparametric estimates. (See Appendix A for details.) Proportion of controls in the overlap region $S_P = 0.60$, proportion of ENPs in $S_P = 0.96$.

[d]The final column gives the ratio of the absolute value of $\hat{\bar{B}}_{S_P}$ to the absolute value of the experimental impact estimate, times 100. The experimental impact estimate is based on the full treatment and control sample.

presents $\hat{\bar{B}}_{S_P}$ ($\hat{\bar{B}}_{S_X}$ evaluated with $X = P$), the selection bias for those in the overlap set S_P. Column (6) presents the experimental impact estimate calculated using the full control and treatment group samples while column (7) expresses $\hat{\bar{B}}_{S_P}$ as a fraction of the experimental program impact estimate. All of the values in the table are reported as monthly dollars. Thus the first row and first column of Table V reports a mean earnings difference of −$420 per month over the three months of the first quarter after random assignment. The percentages of controls and ENPs in the common support region for P_i are reported in the table notes.

A remarkable feature of the estimates in Table V is that for the overall 18 month earnings measure, terms $\hat{B}_1$ and $\hat{B}_2$ are substantially larger than the selection bias component $\hat{B}_3$. The selection bias is a small fraction (only 7%) of the conventional measure of selection bias and is not statistically significantly different from zero.[43] These results on the bias for the overall impact of the

[43] For adult women and for youth the estimated selection bias is proportionately higher, although the conventional measure $\hat{B}$ is lower than for adult males. For adult women and youth the bias measures $\hat{B}$ and $\hat{B}_3$ are of the same order of magnitude. These results are reported in Heckman, Ichimura, and Todd (1997; first draft 1993).

program appear to provide a strong endorsement for matching on P as a method of program evaluation. However, the bias $\hat{\bar{B}}_{S_P}$ that is not eliminated by matching is still large relative to the estimated treatment effects, as is shown in the last two columns.

The decompositions for quarterly earnings tell a somewhat different story. There is more evidence of selection bias in quarters 4 and 5, although even in these quarters the selection bias is still dwarfed by the other components of bias in (20). Expressed as a fraction of the experimental impact estimate, the quarter-by-quarter biases are substantial.

The evidence for the empirical importance of selection bias that cannot be removed by matching is even stronger when we examine the bias at particular deciles of the P_i distribution (conditional on $D = 1$) in the overlap set. Table VII, discussed below, shows that the estimates of bias at the deciles of the control distribution of P are large, negative, and statistically significant at the lowest decile, and large and positive at the upper decile. The apparent success of matching on P in eliminating some of the conventionally-measured selection bias in the overall estimate of program impact masks substantial bias over subintervals of P. The bias that remains after matching is a large fraction of the experimentally-estimated program impact. Our evidence of substantial pointwise bias that averages out to small bias over certain intervals is reminiscent of what can occur in the classical selection bias model, as noted in the discussion surrounding Figure 1. Moreover, it is inconsistent with the identifying assumption used to justify matching. This empirical regularity occurs in the other models estimated below and is a central empirical finding of this paper.

7. TESTING THE CONDITIONS THAT JUSTIFY MATCHING, OUR EXTENSION OF MATCHING, THE INDEX SUFFICIENCY HYPOTHESIS, AND THE CONDITIONAL DIFFERENCE-IN-DIFFERENCES METHOD

We now refine our characterization of the bias function by testing several important hypotheses. The first hypothesis is the fundamental identifying assumption (7) required to identify parameter (1) using matching. Rejection of this hypothesis for a broad array of probabilities of participation P, selected on the basis of various criteria, leads us to test the validity of regression-adjusted matching. In that method, we postulate econometric separability and exclusion restrictions and write $X = (R, Z)$, $Y_0 = R'\beta + U_0$, and $E(U_0 \mid X, D) = E(U_0 \mid Z, D)$. In place of (7), we postulate conditional mean independence for the disturbances that parallels the conditions specified in (A-1) and consider $U_0 \perp\!\!\!\perp D \mid Z$ or the disturbance parallel of (A-2), $U_0 \perp\!\!\!\perp D \mid P(Z)$ or its implication

$$E(U_0 \mid P(Z), D = 1) = E(U_0 \mid P(Z), D = 0) = E(U_0 \mid P(Z)). \tag{21}$$

Separability is a familiar econometric restriction. Exclusion restrictions are motivated by the temporal structure of the program we analyze. Outcomes are affected by variables R, like local labor market variables and time effects, that

are experienced after participation decisions are made due to uncertainty about future labor market shocks.

Our evidence on hypothesis (21) is mixed. Using conventional asymptotic standard errors, we reject (21). Using standard errors that adjust for estimation of β, which are justified in an extensive Monte Carlo analysis reported in Heckman, Ichimura, and Todd (1996; first draft 1994), we do not reject the hypothesis. However, the estimated pointwise bias expressed as a function of P is large. We are reluctant to declare a sizable estimated effect to be zero based on these tests and we conclude that even after adjusting for R, matching is not vindicated in our sample. However, the regression-adjusted method improves on simple matching on P in producing somewhat lower average bias over certain intervals.

We test the index sufficiency hypothesis (11), and do not reject it, although the power of our test is not high in the empirically-relevant range of alternatives. Therefore, a key necessary condition justifying the classical econometric selection bias model is consistent with our data. Large pointwise bias and small average bias over certain intervals are consistent with the econometric selection model. Finally, we test the identifying assumptions of the conditional difference-in-differences estimator and find that they are satisfied in our data for all but low values of P in time periods near the date of random assignment or eligibility determination.

7.1. *Testing the Validity of Matching on P*

We construct our test of the hypothesis (7) from estimates of $\hat{m}_1(P) = \hat{E}(Y_0 \mid P, D = 1)$ and $\hat{m}_0(P) = \hat{E}(Y_0 \mid P, D = 0)$ obtained from the separate local linear regressions of Y_0 on P for observations with $D = 1$ and of Y_0 on P for observations with $D = 0$. The asymptotic normality of the two terms $(N_d a_N)^{1/2}(\hat{m}_d(P) - m_d(P)) \sim N(\Psi_d, V_d)$, $d = 0, 1$ is discussed in Section A.5 of Appendix A, where Ψ_d and V_d are also defined. (See Theorem A.3.) We pick the smoothing parameters to satisfy $a_{N_1} = a_{N_0} = a_N$. The statistic used to test hypothesis (7) is

$$(\hat{m}_1(P) - \hat{m}_0(P))' \left(\hat{V}_1/(a_N N_1) + \hat{V}_0/(a_N N_0) \right)^{-1} (\hat{m}_1(P) - \hat{m}_0(P)) \sim \chi^2(1),$$

where $\hat{V}_d$ is a consistent estimator of V_d for $d \in \{0, 1\}$ and N_1 and N_0 are the sample sizes for $D = 1$ and $D = 0$, respectively. For testing hypothesis (21), the test statistics are analogous except that Y_0 is replaced by $\hat{U}_0$. The test statistics and estimators of the variances for this case are presented in Appendix A, Section A.6. The Monte Carlo evidence reported in Heckman, Ichimura, and Todd (1996; first draft 1994) suggests that adjustment for the estimation of β is required to produce correct standard errors for samples of size 500–1,000 with the variation in the regressors found in the samples used in our analysis.

Tables VIA and VIB present the "p values" (rejection rates under the null) for these hypotheses for various values of the probability of program participa-

TABLE VIA

TESTS OF CONDITIONAL MEAN INDEPENDENCE OF EARNINGS AND RESIDUALS BASED ON ASYMPTOTIC STANDARD ERRORS WITHOUT ADJUSTMENT FOR ESTIMATION OF β[a]

Experimental Controls and Elig. Nonparticipant (ENP) Samples
Adult Males, 508 Controls and 388 ENPs

p-Values from Tests of Conditional Mean Independence of Earnings[b] H_0: $E(Y_0 \mid P, D = 1) = E(Y_0 \mid P, D = 0)$		
Value of P	Joint Test for Quarters $t = 1$ to $t = 6$[c]	Joint Test for Quarters $t = -6$ to $t = -1$
0.0025	0.0000	0.0242
0.005	0.0002	0.0803
0.01	0.0042	0.2416
0.02	0.1224	0.1919
0.03	0.4363	0.1238
0.04	0.7659	0.1585
0.05	0.9423	0.3271
0.10	0.1678	0.8464
Joint	0.0000	0.0159

p-Values from Tests of Conditional Mean Independence of Residuals[b] H_0: $E(U_0 \mid P, D = 1) = E(U_0 \mid P, D = 0)$		
Value of P	Joint Test for Quarters $t = 1$ to $t = 6$	Joint Test for Quarters $t = -6$ to $t = -1$
0.0025	0.0002	0.0293
0.005	0.0004	0.0815
0.01	0.0040	0.2586
0.02	0.2056	0.4078
0.03	0.6563	0.5680
0.04	0.9060	0.7177
0.05	0.9885	0.8064
0.10	0.2591	0.7456
Joint	0.0001	0.2515

[a] Densities were estimated using a biweight kernel and using the fixed bandwidth proposed in Silverman (1986) (defined in Appendix A, Section A.2). Conditional means were estimated by local linear regression using a fixed bandwidth of 0.06 and a biweight kernel. (See Appendix A, Section A.1 for a description of local linear regression and Section A.6 for a description of the test procedure.)

[b] Final row presents the p-value from a joint test.

[c] The number of observations within one bandwidth of $P = 0.0025$ in quarter 1 are 140 controls and 328 ENPs. For other P points, the numbers of observations are the following: 143 controls, 331 ENPs ($P = 0.005$), 150 controls and 336 ENPs ($P = 0.01$), 158 controls and 345 ENPs ($P = 0.02$), 170 controls and 350 ENPs ($P = 0.03$), 184 controls and 353 ENPs ($P = 0.04$), 198 controls and 355 ENPs ($P = 0.05$), and 120 controls and 52 ENPs ($P = 0.1$). The number of observations in other quarters are similar, but vary slightly because of the unbalanced panel data.

TABLE VIB

TESTS OF CONDITIONAL MEAN INDEPENDENCE OF RESIDUALS BASED ON ASYMPTOTIC STANDARD ERRORS WITH ADJUSTMENT FOR ESTIMATION OF β[a]

Experimental Controls and Elig. Nonparticipant (ENP) Samples
Adult Males, 508 Controls and 388 ENPs

p-Values from Tests of Conditional Mean Independence of Residuals[b] H_0: $E(U_0 \mid P, D = 1) = E(U_0 \mid P, D = 0)$		
Value of P	Joint Test for Quarters $t = 1$ to $t = 6$	Joint Test for Quarters $t = -6$ to $t = -1$
0.0025	0.0955	0.6421
0.005	0.1129	0.7346
0.01	0.2170	0.7773
0.02	0.5925	0.6738
0.03	0.8289	0.7808
0.04	0.9563	0.8726
0.05	0.9939	0.9042
0.10	0.5790	0.9253
Joint	1.0000	1.0000

[a] Densities were estimated using a biweight kernel and using the fixed bandwidth proposed in Silverman (1986) (defined in Appendix A, Section A.2). Conditional means were estimated by local linear regression using a fixed bandwidth of 0.06 and a biweight kernel. (See Appendix A, Section A.1 for a description of local linear regression and Section A.6 for a description of the test procedure.)

[b] Final row presents the p-value from a joint test.

tion P located at least one bandwidth apart, so that the test statistics are statistically independent. The top portion of Table VIA reports tests of hypothesis (7). The relevant period over which the test should be performed is the post-random assignment period ($t = 1, \ldots, 6$) since it is post-entry time periods on which the program would be evaluated. For the sake of completeness, however, we also record the test results for the pre-random assignment period ($t = -1, \ldots, -6$).[44] The bottom portion of the table reports tests of hypothesis (21). Hypothesis (7), which justifies matching on P, is decisively rejected. In addition, hypothesis (21) is rejected, so regression-adjusted matching is also inconsistent with our data. When second order-adjusted standard errors are used that account for the estimation of β, as in Table VIB, the evidence is less clear cut. However, the pointwise bias is large (see Figure 3 for bias from the best-predictor P) and it seems inappropriate to ignore this bias and accept the null of no selection bias when an asymptotically-equivalent test of the same hypothesis rejects it. Table VII reports the pointwise bias estimates at deciles of the distribution of P for controls. The bias is large, negative, and statistically significant at low values of P and large and positive at high values of P, which is inconsistent with the null hypothesis that matching is a valid estimator.

[44] The same inferences are found when we test over all 12 periods although such a test is not especially interesting for judging the performance of matching as an evaluation estimator on post-random-assignment data.

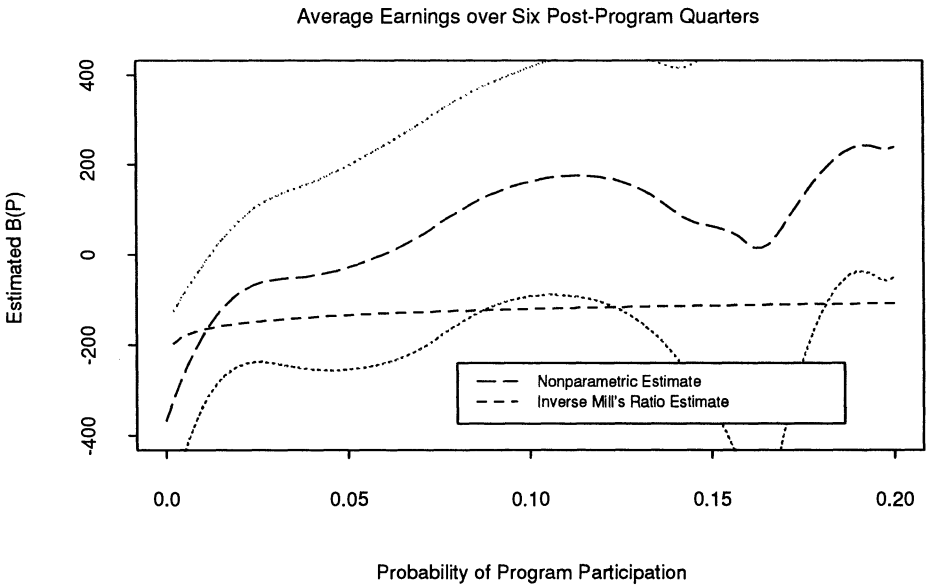


FIGURE 3.—Local linear regression estimates of pointwise bias ($B(P(X))$), adult males, best predictor P model for the probability of program participation; bandwidth = 0.06, trimming = 2%.

7.2. *Testing Index Sufficiency*

Our data are consistent with the hypothesis of index sufficiency. Appendix A, Section A.6.2, presents the test statistic for this hypothesis. We test $E(U_{0t} \mid P, Z, D = 1) - E(U_{0t} \mid P, Z, D = 0) = K_{1t}(P, Z) - K_{0t}(P, Z)$ for different discrete regressors, Z, shown in the subtable headings of Table VIII, using the best-predicting P score selected on the basis of tests discussed in Appendix C.[45] For most cells, and tests over all cells using conventional significance levels, we do not reject the hypothesis in the relevant post-random assignment period ($t = 1, \ldots, 6$) or for that matter in the pre-random assignment period ($t = -1, \ldots, -6$). P values are chosen at least one bandwidth apart so that the test statistics are statistically independent. A Monte Carlo analysis of the test statistic presented in Appendix D reveals that the test is consistent but quite conservative. It rejects at a far higher rate (25%) than the normal size (5%). On the other hand, the power of the test is not especially high (roughly 20%) for a large range of alternatives away from the null. A similar pattern of acceptance

[45] Since the terms $E(U_{0t} \mid P, Z, D = 1)$ and $E(U_{0t} \mid P, Z, D = 0)$ are identified only up to unknown constants, we do not test the hypotheses $K_{0t}(P, Z) = K_{0t}(P)$ and $K_{1t}(P, Z) = K_{1t}(P)$. Our test of index sufficiency is different from that of Fan and Li (1996) because we test the hypothesis that *differences* are index-sufficient, not levels. Our test is also different from that of Aït-Sahalia, Bickel, and Stoker (1994) because we test for index sufficiency of a subfunction and not an entire function and we use local linear regression methods which greatly simplify the derivation of the sampling distribution of test statistics. See the discussion in Appendix A.

TABLE VII

ESTIMATED SELECTION BIAS AT DECILES OF THE CONTROL P DISTRIBUTION FOR THE BEST PREDICTOR P MODEL[a]
Quarterly Earnings Stated in Monthly Dollars[b]
Experimental Control and Elig. Nonparticipant (ENP) Samples, Adult Males, 508 Controls and 388 ENPs

	Decile of the Control Empirical Distribution of P[c] (Decile boundaries shown in brackets)								
Quarter	1 [0.0002, 0.0023)	2 [0.0023, 0.0087)	3 [0.0087, 0.0152)	4 [0.0152, 0.0269)	5 [0.0269, 0.0410)	6 [0.0410, 0.0822)	7 [0.0822, 0.0983)	8 [0.0983, 0.1337)	9 [0.1337, 0.2534]
Qtr1	−338	−229	−139	−83	−20	84	66	−2	517
	(121)	(92)	(83)	(86)	(101)	(122)	(131)	(175)	(320)
Qtr2	−260	−194	−144	−95	−23	130	157	228	492
	(139)	(109)	(94)	(86)	(97)	(131)	(127)	(165)	(348)
Qtr3	−295	−195	−118	−59	6	176	202	275	442
	(140)	(111)	(96)	(86)	(95)	(134)	(127)	(193)	(378)
Qtr4	−193	−103	−50	−21	38	193	152	54	530
	(133)	(107)	(95)	(90)	(102)	(133)	(132)	(183)	(376)
Qtr5	−246	−146	−84	−45	22	257	246	−163	519
	(139)	(112)	(102)	(97)	(119)	(169)	(176)	(240)	(398)
Qtr6	−359	−262	−173	−76	−3	169	191	97	428
	(117)	(94)	(88)	(102)	(130)	(175)	(191)	(205)	(342)
Average of	−282	−188	−118	−63	3	168	169	81	488
1 to 6	(116)	(91)	(81)	(79)	(98)	(130)	(117)	(147)	(281)

[a] The best predictor model is given in Table III.

[b] Bootstrap standard errors are shown in parentheses. They are based on 50 replications with 100% sampling. For the nonparametric estimates, a fixed bandwidth of 0.06 and a biweight kernel function were used. (See Appendix A, Sections A.1 and A.5.2 for additional details concerning the estimation procedure.)

[c] The deciles are based on the distribution of control probabilities of participation in the region of overlapping support, S_P. A 2% trimming rule was used in determining the overlapping support region, and a 0.06 fixed bandwidth was used for the nonparametric estimates. (See Appendix A for details.) Proportion of controls in the overlap region $S_P = 0.60$, proportion of ENPs in $S_P = 0.96$. There are too few eligible nonparticipant observations to estimate the bias reliability in the 10th decile.

TABLE VIII

p-VALUES FROM TESTS OF INDEX SUFFICIENCY[a]
Experimental Controls and Elig. Nonparticipant (ENP) Samples
Best Predictor Model for the Probability of Program Participation
Adult Males, 508 Controls and 388 ENPs

Tests by Race and Ethnicity[b]		
Value of P	Joint Test for Quarters $t = 1$ to $t = 6$	Joint Test for Quarters $t = -6$ to $t = -1$
0.002	0.4229	0.4376
0.039	0.5213	0.7867
0.076	0.2268	0.5307
0.113	0.6526	0.2827
0.150	0.0175	0.2440
Joint[c]	0.0421	0.3983
Tests by Training Center[b]		
Value of P	Joint Test for Quarters $t = 1$ to $t = 6$	Joint Test for Quarters $t = -6$ to $t = -1$
0.008	0.4942	0.8503
0.026	0.3404	0.1392
0.044	0.0952	0.0230
0.062	0.0925	0.0626
0.080	0.4062	0.2633
Joint[c]	0.4667	0.5959
Tests by Years of Schooling Categories[b]		
Value of P	Joint Test for Quarters $t = 1$ to $t = 6$	Joint Test for Quarters $t = -6$ to $t = -1$
0.003	0.4717	0.4646
0.022	0.5736	0.2842
0.042	0.2576	0.0967
0.061	0.0792	0.0188
0.080	0.0967	0.0964
Joint[c]	0.1718	0.1686

[a]Densities were estimated using a biweight kernel and using the fixed bandwidth proposed in Silverman (1986) (defined in Appendix A, Section A.2). Conditional means were estimated by local linear regression using a fixed bandwidth of 0.06 and a biweight kernel. (See Appendix A, Section A.1 for a description of local linear regression and Section A.6 for a description of the test procedure.) Standard errors used in the test are asymptotic and are not adjusted for higher order terms (as described in Appendix A, Section A.6). When adjustment is made for estimation of β, the estimated standard errors are substantially larger.

[b]The tests by race and ethnicity include "White" and "Black" groups. The tests by training center include "Fort Wayne," "Jersey City," and "Providence." The tests by years of schooling category include "Fewer than 10 years of schooling," "10–11 years of schooling," "12 years of schooling," and "More than 12 years of schooling."

[c]Joint tests shown include only a subset of the P points that are at least one bandwidth apart.

of the null of index sufficiency is found for all specifications of P shown in Table VIII, except when P scores are used which exclude both earnings and recent labor force transition information.

Our acceptance of index sufficiency is necessarily qualified because the power of our test is not especially high. The test partitions the data by demographic group, by training center at which the experiment was conducted, and by education group. This partitioning sometimes produces very small cells and it greatly restricts the range of P over which the test can be performed. When certain cells are deleted, the range of P values over which the test can be performed is greatly expanded. For this reason, the tests reported in Table VIII omit the "Hispanic" race/ethnicity and the "Corpus Christi" training site cells. Unlike the case of our test of the conditional independence assumptions that justify the conventional matching estimator, where the rejections are firm, here we can only make the guarded statement that the data are consistent with the null hypothesis of index sufficiency and that further tests with larger samples would be highly desirable.[46] The pointwise differences in the bias are sometimes substantial (see Figure D-2 displayed in Appendix D), but so are the standard errors.

Moreover, as noted in Section 3.2, in order to use the index-sufficient model to construct the desired counterfactual (1) it is necessary to be able to determine a set of X values where $E(U_0 \mid P(X), D = 0) = 0$. The restricted support of $P(X)$ evident in Figure 2 precludes this identification strategy unless parametric restrictions are invoked. The restriction on the support of P in our sample also eliminates the possibility of a more general statement about the shape of $B(P)$ over the full support of P for program participants. Future evaluations should select comparison groups to enlarge S_P to the full support of program participants in order to allow valid inferences about the entire sample of participants.

7.3. *Testing the Identifying Assumption Justifying the Conditional Difference-in-Differences Method*

Maintaining index sufficiency to characterize bias $B(X)$ simplifies the testing of identifying assumption (12). In light of our evidence on index sufficiency we can reformulate it in the following way:

$$B_t(P(X)) - B_{t'}(P(X)) = 0, \qquad \text{for some } t, t' \tag{22}$$

where t is a post-program period and t' is a pre-program period.

Figure 4 plots the pointwise bias estimates over all t. The $B_t(P)$ are not constant over time, or even equal for time periods $t = -t'$ at low values of P for

[46] In general, a multiple index model would characterize participation in the program, reflecting the preferences of the individuals and those of the bureaucrats who accept people into the program. Heckman, Smith, and Taber (1996) report the absence of cream-skimming behavior at one of the JTPA training centers analyzed in this paper. (The required data are not available at the other centers.) In a larger sample, or with different decision rules used by program officials, the single index model might be rejected in favor of a multiple index model. Local linear regression methods can easily be modified to estimate models with multiple indices using higher dimension kernels.

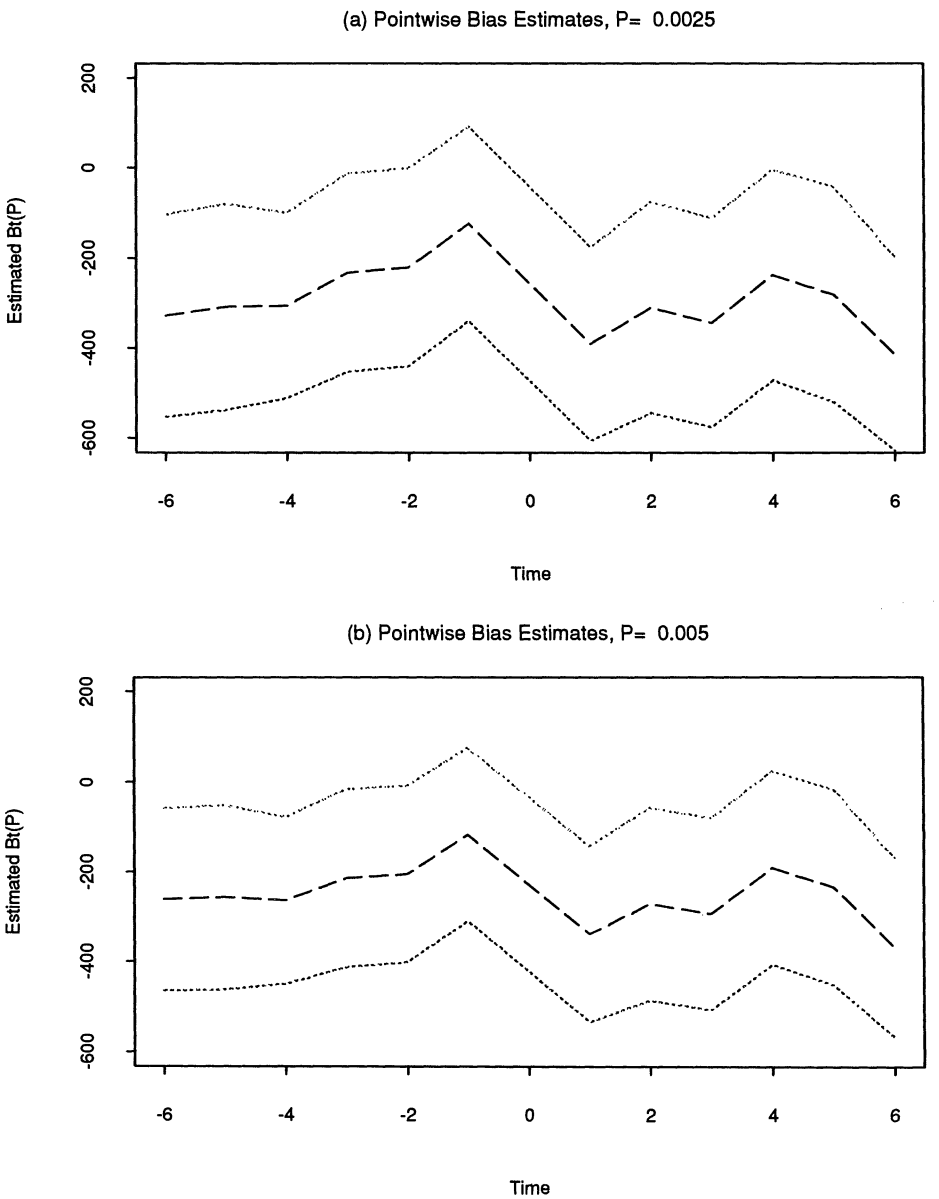


FIGURE 4.—Local linear regression estimates of pointwise bias $B_t(P(X))$ over time, adult males, best predictor P model for the probability of program participation; bandwidth = 0.06, trimming = 2%.

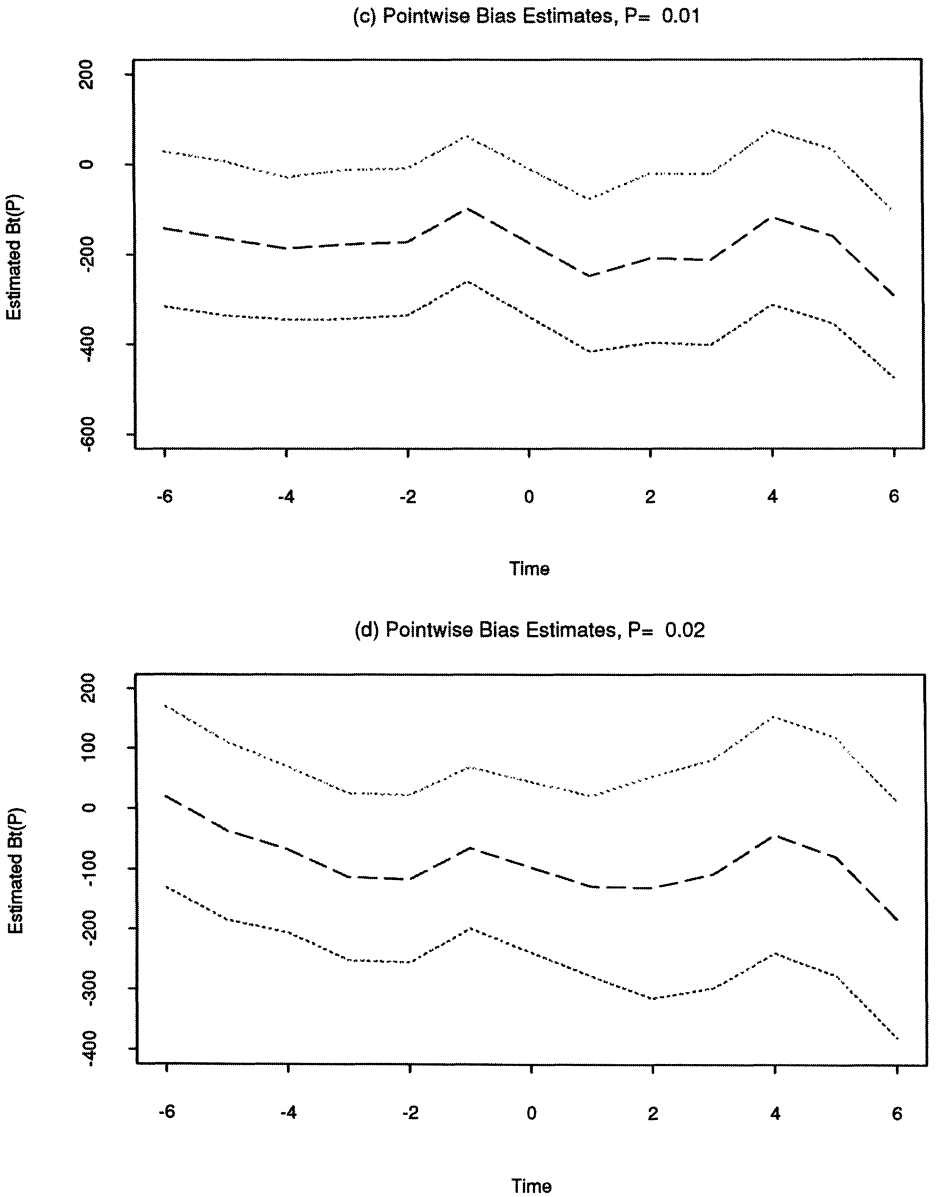
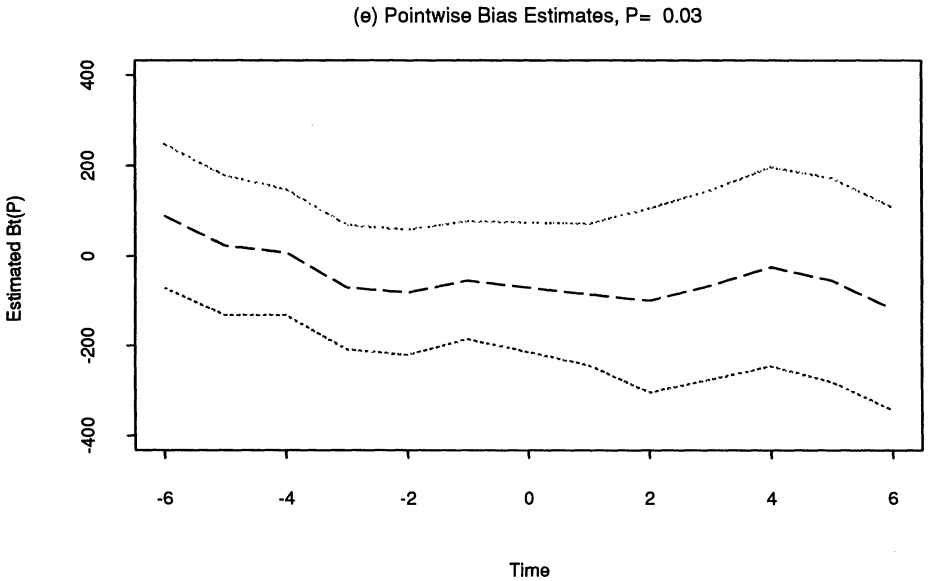


FIGURE 4.—*Continued*



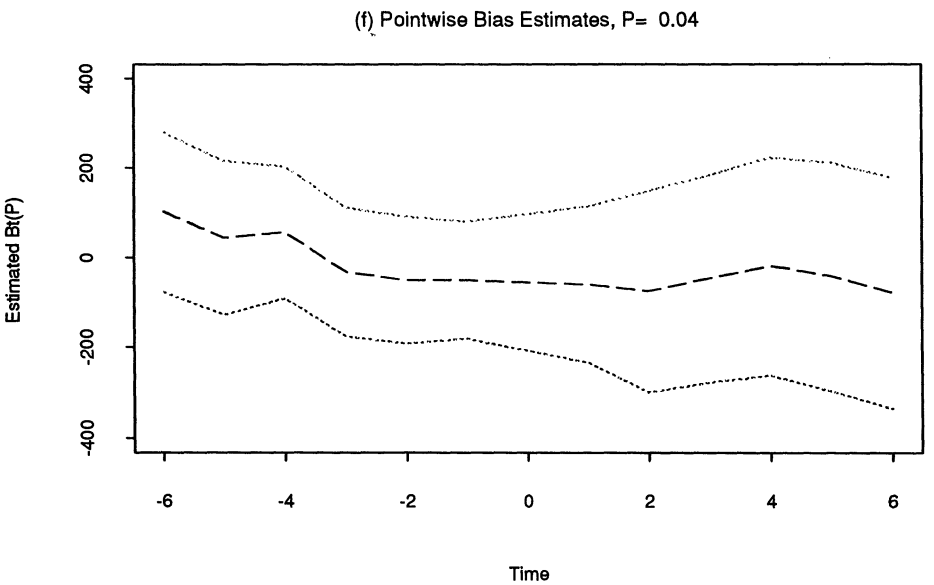


FIGURE 4.—*Continued*

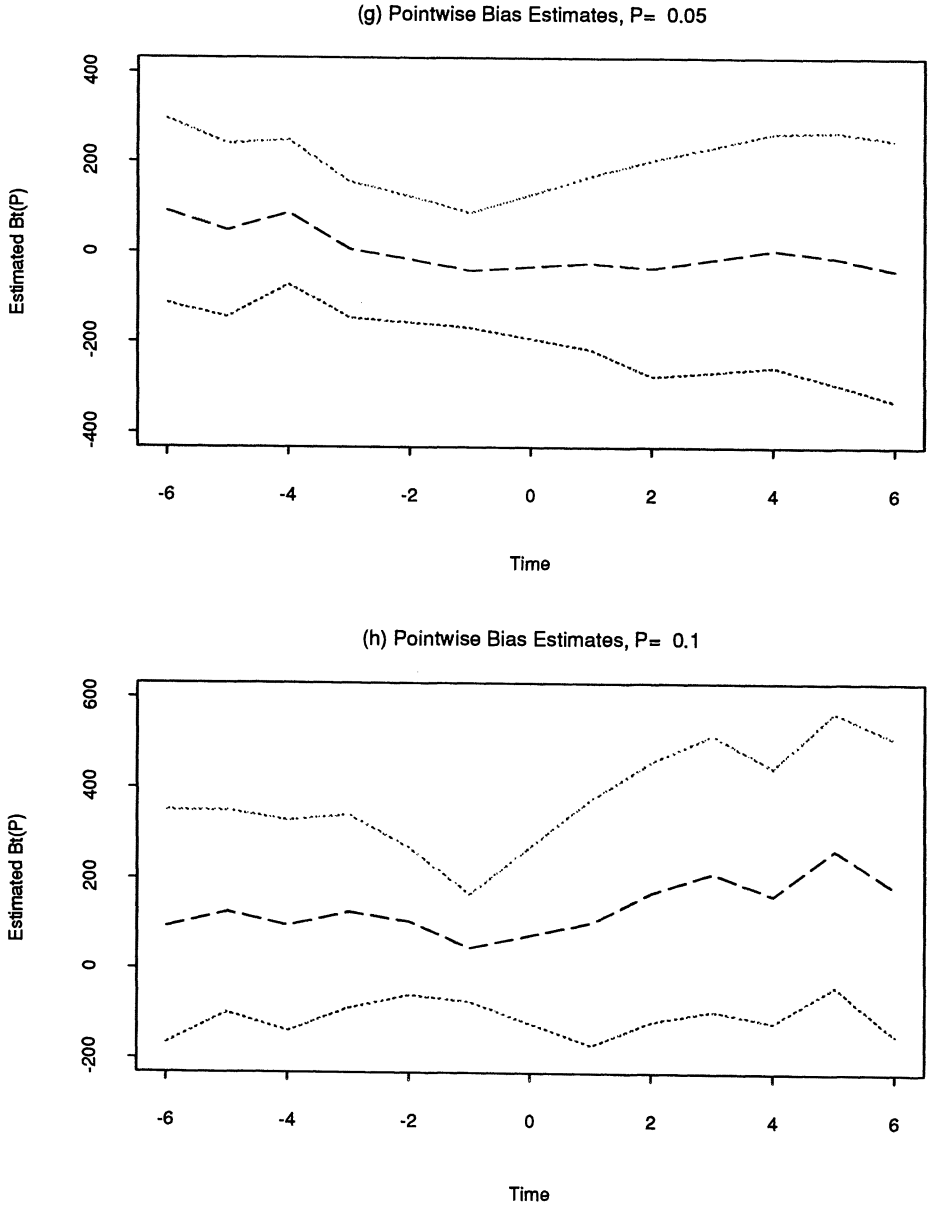


FIGURE 4.—*Continued*

time periods near the time of the participation decision. In general, however, the identifying assumption justifying the conditional difference-in-differences estimator is consistent with our data. The fourth column of Table IX presents p values for tests of hypothesis (22) for symmetric differences around $t = 0$.[47] Only

[47] The inference using the unadjusted standard errors is the same as that reported in Table IX.

TABLE IX

p-VALUES FROM TESTS FOR FIXED EFFECT AND DIFFERENCE-IN-DIFFERENCES SPECIFICATIONS FOR THE BIAS FUNCTION
All Tests are Symmetric Around Time $t = 0$[a,b]
Adult Males, 508 Controls and 388 ENPs

Value of P (1)	Fixed Effect Test for Controls (2)	Fixed Effect Test for ENPs (3)	Difference-in-Differences Test (4)
	Null tested jointly over $t \in \{1,2,3,4,5,6\}$[c]		
0.0025	0.0922	0.8042	0.1221
0.0050	0.0967	0.8578	0.1291
0.0100	0.1148	0.8609	0.1688
0.0200	0.2158	0.3093	0.4279
0.0300	0.2327	0.0948	0.7579
0.0400	0.0807	0.0454	0.9353
0.0500	0.0047	0.0466	0.9707
0.1000	0.0057	0.1785	0.9914
Overall	0.1303	0.0019	0.8087
	Null tested jointly over $t \in \{1,2,3\}$[c]		
0.0025	0.3160	0.9417	0.4083
0.0050	0.3251	0.9832	0.4907
0.0100	0.3614	0.9949	0.7176
0.0200	0.4392	0.6269	0.9899
0.0300	0.3158	0.1877	0.9999
0.0400	0.1159	0.0785	1.0000
0.0500	0.0121	0.0680	0.9999
0.1000	0.0068	0.4363	0.9999
Overall	0.0255	0.0001	0.7008
	Null tested jointly over $t \in \{4,5,6\}$[c]		
0.0025	0.9456	0.9968	0.9386
0.0050	0.8779	0.9832	0.8424
0.0100	0.7217	0.9115	0.6266
0.0200	0.6256	0.8132	0.6123
0.0300	0.7353	0.8901	0.8768
0.0400	0.8467	0.9206	0.9772
0.0500	0.8522	0.9263	0.9906
0.1000	0.4140	0.9931	0.9966
Overall	0.9498	0.3141	0.9376

[a] Densities were estimated using a biweight kernel and using the fixed bandwidth proposed in Silverman (1986) (defined in Appendix A, Section A.2). Conditional means were estimated by local linear regression using a fixed bandwidth of 0.06 and a biweight kernel. (See Appendix A, Section A.1 for a description of local linear regression and Section A.6 for a description of the test procedure.) Standard errors used in the test are asymptotic and are not adjusted for higher order terms (as described in Appendix A, Section A.6). When adjustment is made for estimation of β, the estimated standard errors are substantially larger.

[b] Null hypothesis for fixed effects test for controls is H_0: $K_{1t}(P) - K_{1,-t}(P) = 0$; null hypothesis for fixed effect test for ENPs is H_0: $K_{0t}(P) - K_{0,-t}(P) = 0$; null hypothesis for difference-in-differences test is H_0: $[K_{1t}(P) - K_{1,-t}(P)] - [K_{0t}(P) - K_{0,-t}(P)] = 0$; where $(-t)$ is a pre-program period t periods before random assignment or eligibility determination.

[c] Values of P in the overall test are at least one bandwidth apart.

for the lowest values of P in the joint test over all six pairs of quarters is the null close to being rejected at conventional levels. Outside the interval $t \in [-3, 3]$, hypothesis (22) is never close to being rejected for any values of P. Table X presents the bias by P decile in a format comparable to that of Table VII. For most deciles, the bias is substantially lower than for the matching estimator. Pointwise, the estimated bias using the difference-in-differences matching estimator, which is a differenced version of the regression-adjusted matching estimator, is lower than that for the cross-sectional matching estimator or the regression-adjusted matching estimator.[48]

Column (2) of Table IX reports p values for the test of the identifying assumption of the fixed effect model $(K_{1t}(P) = K_{1,(-t)}(P))$. In a stationary environment, the fixed effect method applied to controls $(D = 1)$ is sufficient to identify the parameter of interest.[49] This hypothesis is decisively rejected overall for the ENPs and in most cases for the controls, but the data are consistent with the hypothesis of fixed effects in the interval outside $t \in [-3, 3]$. The results in column (3) of Table IX show that the same conclusions apply to the hypothesis $K_{0t}(P) = K_{0,(-t)}(P)$.

8. ESTIMATED SELECTION BIAS UNDER ALTERNATIVE ESTIMATORS AND SENSITIVITY OF ESTIMATES TO ALTERNATIVE SPECIFICATIONS OF THE OUTCOME AND PARTICIPATION EQUATIONS

This section presents estimates of selection bias associated with the alternative estimators described above and explores the sensitivity of the estimated average selection bias, $\overline{B}_{S_P}$, to variations in the variables included in the outcome equations (R) and in the participation equation (Z). We also compare the selection bias, rigorously defined, that is obtained from the method of Barnow, Cain, and Goldberger (1980) with the bias from the local linear regression estimator.

Table XI presents estimates of selection bias associated with different matching estimators, where matching is performed using the best-predictor model for P. The first column of Table XI gives the benchmark difference in raw mean earnings between the control and ENP groups. Column (2) is the bias for a local-linear P matching estimator without regression-adjustment, which imposes a common support condition and uses nonparametric local linear regression methods in constructing matches. The average bias estimate of \$47 improves substantially over a simple mean-difference estimator. Column (3) gives the estimated bias for the regression-adjusted version of the same estimator. The fourth and fifth columns present the bias estimates for the difference-in-dif-

[48] The bias by decile for the regression-adjusted matching method is only slightly smaller (less than 10%) for each decile. For the sake of brevity we do not display these results.

[49] See Heckman and Robb (1985).

TABLE X

DIFFERENCE-IN-DIFFERENCES AT DECILES OF THE CONTROL P DISTRIBUTION FOR THE BEST PREDICTOR MODEL FOR P[a]
Differences Are Symmetric Around Time $t = 0$,[b] Quarterly Earnings Stated in Monthly Dollars
Experimental Control and Elig. Nonparticipant (ENP) Samples, Adult Males, 508 Controls and 388 ENPs

	Decile of the Control Empirical Distribution of P[c] (Decile boundaries shown in brackets)								
Quarter	1 [0.0002, 0.0023)	2 [0.0023, 0.0087)	3 [0.0087, 0.0152)	4 [0.0152, 0.0269)	5 [0.0269, 0.0410)	6 [0.0410, 0.0822)	7 [0.0822, 0.0983)	8 [0.0983, 0.1337)	9 [0.1337, 0.2534]
Qtr1	−269	−148	−63	−14	25	68	38	−112	858
	(160)	(125)	(105)	(91)	(92)	(106)	(140)	(215)	(614)
Qtr2	−100	−42	−19	−23	−9	39	96	200	724
	(164)	(127)	(108)	(100)	(114)	(147)	(161)	(177)	(720)
Qtr3	−119	−38	2	−2	2	65	126	413	627
	(156)	(131)	(115)	(114)	(125)	(147)	(183)	(284)	(802)
Qtr4	54	63	26	−45	−34	99	113	141	−102
	(165)	(133)	(115)	(113)	(128)	(160)	(194)	(361)	(681)
Qtr5	3	−2	−40	−68	−11	216	143	−350	38
	(186)	(139)	(114)	(117)	(157)	(209)	(196)	(267)	(432)
Qtr6	−92	−139	−181	−149	−61	177	126	−168	−44
	(172)	(129)	(114)	(138)	(182)	(228)	(217)	(284)	(401)
Average of	−87	−51	−46	−50	−14	111	107	20	350
1 to 6	(147)	(114)	(96)	(94)	(113)	(140)	(135)	(165)	(408)

[a]The best predictor model was used to estimate the probability of participation. It is given in Table III.
[b]For the nonparametric estimates, a fixed bandwidth of 0.06 and a biweight kernel function were used. (See Appendix A, Sections A.1 and A.4.2 for additional details concerning the estimation procedure.) Bootstrap standard errors are shown in parentheses. They are based on 50 replications with 100% sampling.
[c]The deciles are based on the distribution of control probabilities of participation in the region of overlapping support. (See footnote to Table VII.) There are too few eligible nonparticipant observations to estimate the bias reliably in the 10th decile.

TABLE XI

COMPARISON OF ESTIMATED SELECTION BIAS UNDER ALTERNATIVE ESTIMATORS OF PROGRAM IMPACTS FOR THE BEST PREDICTOR MODEL FOR P[a, b]
Quarterly Earnings Stated in Monthly Dollars
Experimental Control ($D = 1$) and Elig. Nonparticipant (ENP) ($D = 0$) Samples
Adult Males, 508 Controls and 388 ENPs

Quarter	(1) Difference in Means	(2) Local Linear P Score Matching	(3) Regression-Adjusted Local Linear Matching	(4) Difference-in-Differences Local Linear P Score Matching	(5) Difference-in Differences Regression-Adjusted Local Linear Matching
Qtr1	−418 (38)	33 (59)	39 (60)	97 (62)	104 (63)
Qtr2	−349 (47)	37 (61)	39 (64)	77 (89)	77 (92)
Qtr3	−337 (55)	29 (78)	21 (80)	90 (114)	74 (114)
Qtr4	−286 (57)	80 (77)	65 (82)	112 (90)	98 (91)
Qtr5	−305 (57)	64 (77)	50 (83)	19 (95)	−5 (99)
Qtr6	−328 (63)	37 (82)	17 (90)	4 (105)	−35 (111)
Average of 1 to 6	−337 (47)	47 (60)	39 (64)	67 (71)	52 (74)
As a % of impact	775%	107%	88%	153%	120%

[a] The best predictor model was used for the probability of participation. It is given in Table III.
[b] For the nonparametric estimates, a fixed bandwidth of 0.06 and a biweight kernel function were used. (See Appendix A, Sections A.1, A.4, and A.5 for additional details concerning the estimation procedure.) Bootstrap standard errors are shown in parentheses. They are based on 50 replications with 100% sampling.

ferences and regression-adjusted difference-in-differences estimators, respectively. The estimated bias is slightly higher.[50]

In Table XII, we explore the sensitivity of the bias estimates to alternative sets of variables included in the outcome equation. That is, we use the best-predictor P model defined in Appendix C throughout the sensitivity analysis but vary R. Table XII reveals that there is relatively little sensitivity in the estimates of selection bias across specifications of the outcome equations. For example, comparing the baseline specification with Model I, which includes no regressors except for an intercept, shows little effect of inclusion of the baseline regressors on the estimated overall bias. Addition of training center indicators, race/ethnicity, age, and calendar quarter and year dummies (Model II) to the stripped-down Model I decreases the estimated overall selection bias roughly by a factor of two. Augmenting the regressors of Model II to include measures of previous training, work experience, the local unemployment rate, and a dummy variable for whether or not a child is present (Model III) increases estimated overall selection bias only by a small amount compared to Model II. Adding schooling, age, and marital status to the Model III specification to produce

[50] Heckman, Ichimura, and Todd (1997; first draft 1993) apply the conditional difference-in-differences estimator to data from three other demographic groups and find that it generally yields bias estimates similar to those obtained using cross-sectional matching estimators.

TABLE XII

SELECTION BIAS UNDER DIFFERENT OUTCOME EQUATION MODELS: ESTIMATED BIAS FROM REGRESSION-ADJUSTED LOCAL LINEAR MATCHING ESTIMATOR[a, b]
Quarterly Earnings Stated in Monthly Dollars
Experimental Control ($D = 1$) and Elig. Nonparticipants (ENP) ($D = 0$) Samples
Adult Males, 508 Controls and 388 ENPs

Quarter	Baseline[c]	Model I[c]	Model II[c]	Model III[c]	Model IV[c]	Model V[c]	Method of Barnow Cain and Goldberger[d]
Qtr1	39 (60)	33 (59)	32 (68)	38 (68)	33 (65)	63 (62)	150 (34)
Qtr2	39 (64)	37 (61)	33 (64)	38 (68)	32 (63)	63 (65)	126 (28)
Qtr3	21 (80)	29 (78)	5 (76)	14 (78)	4 (76)	40 (83)	82 (16)
Qtr4	65 (82)	80 (77)	40 (78)	54 (78)	41 (79)	82 (82)	125 (27)
Qtr5	50 (83)	64 (77)	32 (75)	44 (78)	28 (77)	66 (80)	142 (45)
Qtr6	17 (90)	37 (82)	−5 (81)	9 (85)	−7 (84)	38 (87)	108 (53)
Average of 1 to 6	39 (64)	47 (60)	23 (61)	33 (63)	22 (62)	59 (65)	134 (51)
As a % of impact	88%	107%	52%	76%	50%	135%	304%

[a]The regression-adjusted average bias is defined in Appendix A, Section A.5.3. In the estimation of the model, densities were estimated using a biweight kernel and the fixed bandwidth proposed in Silverman (1986) (defined in Appendix A, Section A.2). The bias function was estimated by local linear regression using a fixed bandwidth of 0.06 and a biweight kernel. The overlapping support region was determined using a 2% trimming rule and a biweight kernel function. (See Appendix A, Section A.1 for a description of local linear regression, and Section A.4.1 for the method used to determine the overlapping support region.)

[b]Bootstrap standard errors are shown in parentheses. They are based on 50 replications with 100% sampling.

[c]Baseline outcome model includes dummy variables for specific training center, race or ethnicity, schooling, age, and previous training. It also includes work experience, local unemployment rate, an indicator for marital status, an indicator for presence of a child age less than six, and quarter and year indicators.

Outcome Model I includes no R variables and is equivalent to local linear P score matching.

Outcome Model II augments I with training center indicators, race or ethnicity, age, and quarter and year indicators.

Outcome Model III augments II with previous training, work experience, local unemployment rate, and presence of a child age less than six.

Outcome Model IV augments II with local unemployment rate, presence of a child age less than six, schooling, and a marital status indicator.

Outcome Model V augments the baseline model with labor force transition indicators.

[d]The Barnow, Cain, and Goldberger (1980) method is based on equation (15) in the text with the $X = (R, Z)$ the same as in the baseline model, and $B(X) = B(Z)$ and $E(U_0 \mid X, D = 0) = E(U_0 \mid R, D = 0) = E(U_0 \mid R)$ assumed linear. We impose a common support restriction in defining the sample used for estimation where observations are used with $P(Z) \in \hat{S}_P$ for the baseline model. The bias is computed using the distribution of $P \mid D = 1$.

Model IV barely changes the estimated selection bias. Adding the labor force transition variables (Model V) that prove useful in estimating the probability of participation substantially increases the estimated selection bias. These variables are not included in the baseline model and are typically not used as regressors in earnings equations.

The final column of Table XII presents the selection bias that arises from using the method of Barnow, Cain, and Goldberger (1980). This is a weighted linear regression version of our method of regression-adjusted matching. Using the same outcome variables (R) and selection variables (Z) that appear in the baseline model, we estimate linear regression (15) where $B(X) = B(Z)$ is postulated to be a linear function of Z and $E(U_0 \mid X, D = 0) = E(U_0 \mid R, D = 0) = E(U_0 \mid R)$ under their hypothesis, is postulated to be linear in R. We impose the condition of common support to secure estimates from the method by using the observations with $P(Z) \in \hat{S}_P$, and we impose common weighting in estimat-

ing the regression across ENP ($D = 0$) and control samples by weighting the ENP observations by the ratio of the estimated control and ENP densities $\hat{f}(P \mid D = 1)/\hat{f}(P \mid D = 0)$.[51] The estimated selection biases are large when compared with those obtained from the baseline semiparametric model. Our semiparametric alternative to linear regression methods offers substantial benefits in reducing selection bias.[52]

Table XIII presents a sensitivity analysis of the effect of changes in Z on the estimated selection bias for both the regression-adjusted local linear matching estimator and the difference-in-differences version of the estimator. The baseline regressors R from the previous table are maintained through all of the specifications examined here. The second column of the table presents the baseline selection bias for the regression-adjusted model. "Coarse P I" is a model that only includes demographics, schooling, and training center dummies in Z. If there is no access to information on earnings or labor force histories to include in Z, the estimated bias for the local linear estimator is substantial. For the difference-in-differences estimator, the quarterly bias estimates are also substantial but they average out to a low value of $32 per month. Access to information on earnings from the year preceding random assignment or eligibility determination greatly improves but does not eliminate the estimated selection bias for the local linear regression estimator, as shown by the estimates for the "Coarse P II" model. The estimates for the "Coarse P III" model demonstrate that adding local labor force transition variables to the "Coarse P I" model greatly reduces the estimated selection bias. The importance of recent labor force transitions in predicting P and eliminating selection bias is a major empirical finding of this paper. This information was not used in earlier evaluations of U.S. job training programs because it was not available.

9. SENSITIVITY OF THE ESTIMATED BIAS TO ALTERNATIVE DEFINITIONS OF ELIGIBILITY, MISMATCH OF GEOGRAPHY, AND ALTERNATIVE FORMATS OF SURVEY QUESTIONS

National comparison group samples are commonly used to evaluate local programs. These samples do not place comparison group members and participants in the same labor markets. Moreover, the variables and interview formats

[51] Following the analysis of White (1980), such weighting reduces misspecification error for $E(U_0 \mid X, P(Z), D = 1) - E(U_0 \mid X, P(Z), D = 0)$ when the bias function is assumed to be linear and is in fact not linear. The densities are estimated by kernel methods using the kernel defined in Appendix A. Imposing the common support condition ensures that the denominator is nonzero. In results not reported for the sake of brevity, we use an alternative way to impose the common weighting condition. A regression is first estimated without rewriting to obtain an estimate of $B(X_i)$ for each person, and then the common weighting by $f(P \mid D = 1)$ is used in averaging individual $\hat{B}(X)$ estimates. Introducing weighting in the first stage regression makes a substantial difference in the resulting estimates of bias. The estimated bias is about four times larger if the regression is unweighted and the weighting is performed in the second stage.

[52] Below, in Table XIX, we report estimated bias for a more standard version of the Barnow, Cain, and Goldberger (1980) estimator that does not impose common support or common weighting.

TABLE XIII

COMPARISON OF ESTIMATED SELECTION BIAS $\overline{B}_{S_P}(adj)$ UNDER DIFFERENT MODELS FOR P ESTIMATED BIAS FROM REGRESSION-ADJUSTED LOCAL LINEAR MATCHING ESTIMATOR AND DIFFERENCE-IN-DIFFERENCES ESTIMATOR[a, b]

Quarterly Earnings Stated in Monthly Dollars Experimental Control ($D = 1$) and Elig. Nonparticipant (ENP) ($D = 0$) Samples, Adult Males, 508 Controls and 388 ENPs

Quarter	Best Predictor P[c]	Difference-in-differences Best Predictor P	Coarse PI[c]	Difference-in-differences Coarse PI	Coarse PII[c]	Difference-in-differences Coarse PII	Coarse PIII[c]	Difference-in-differences Coarse PIII
Qtr1	39 (60)	104 (63)	−390 (50)	167 (67)	−228 (67)	31 (57)	−84 (77)	67 (68)
Qtr2	39 (64)	77 (92)	−312 (58)	143 (82)	−193 (61)	−80 (62)	−39 (88)	103 (107)
Qtr3	21 (80)	74 (114)	−286 (62)	62 (95)	−153 (57)	−158 (71)	−36 (96)	105 (134)
Qtr4	65 (82)	98 (91)	−231 (64)	33 (93)	−104 (66)	−150 (82)	−9 (92)	47 (109)
Qtr5	50 (83)	−5 (99)	−244 (72)	−73 (104)	−146 (70)	−254 (86)	20 (96)	−29 (122)
Qtr6	17 (90)	−35 (111)	−286 (84)	−143 (106)	−172 (79)	−255 (96)	−3 (111)	−36 (129)
Average of 1 to 6	39 (64)	52 (74)	−291 (54)	32 (78)	−166 (56)	−144 (61)	−25 (83)	43 (95)
As a % of impact	88%	120%	670%	73%	382%	332%	58%	98%

[a]The regression-adjusted average bias is defined in Appendix A, Section A.5.3. In the estimation of the model, densities were estimated using a biweight kernel and the fixed bandwidth proposed in Silverman (1986) (defined in Appendix A, Section A.2). The bias function was estimated by local linear regression using a fixed bandwidth of 0.06 and a biweight kernel. (See Appendix A, Section A.1 for a description of local linear regression.)

[b]Bootstrapped standard errors are shown in parentheses. They are based on 50 replications with 100% sampling.

[c]Best predictor model for P is the same as shown in Table III.

Coarse P model I includes indicator variables for training site, race or ethnicity, age, schooling, marital status, and presence of a child age less than six.

Coarse P model II augments I with earnings from the year prior to random assignment or eligibility determination.

Coarse P model III augments I with the labor force status transition patterns used in the best predictor model for program participation.

sometimes differ across surveys creating further sources of discrepancy between participant and comparison groups unrelated to selection bias, rigorously defined. LaLonde (1986) uses comparison groups situated in different markets from his participants, all of which were administered different questionnaires than those given to participants. Part of the bias that he reports arises from market and survey mismatch. This section investigates these sources of bias and also explores the impact on the estimated bias of imposing different eligibility criteria in creating nonexperimental comparison group samples.

We use SIPP (Survey of Income and Program Participation) data to investigate these issues. These data are sufficiently rich that it is possible to determine whether surveyed persons are eligible for JTPA. However, because of sample size and confidentiality restrictions, it is not possible to make close geographical matches between controls and nonparticipants. In addition, the SIPP survey asks questions about earnings and labor force participation in a different format than does the survey used to produce our data.[53]

Table XIVA presents estimates of the bias (B), the average bias after local linear matching on P, $\overline{B}_{S_P}$, and the regression-adjusted bias, $\overline{B}_{S_P}(adj)$, from three alternative comparison samples. The first sample ("full sample") uses all SIPPs. The second sample uses SIPPs screened for eligibility for JTPA using the rough guidelines employed by Ashenfelter and Card (1985) in their evaluation of the closely-related CETA program. The third sample used only JTPA-eligible persons.[54] The raw bias B greatly diminishes as more refined eligibility criteria are imposed to create comparison samples. For the first two samples, matching and regression-adjusted matching eliminate a substantial portion of the raw bias but the bias that remains is still large relative to the program impact. Imposing eligibility actually increases the measured bias obtained from either method of matching for the SIPP sample of persons, constructed using either the Ashenfelter-Card criterion or exact eligibility for JTPA. Table XIVB presents analogous estimates for the difference-in-differences estimators but the benefits of imposing eligibility criteria on the sample are small. Using samples of eligible individuals as comparison group members may be intuitively appealing but is not guaranteed to reduce selection bias compared to the estimates obtained from other samples. The estimator performs comparably for the full sample and the Ashenfelter and Card (1985) eligible sample, but the bias increases for the sample imposing the more refined eligibility criterion.

Our estimates demonstrate the importance of basic data quality in producing valid program evaluations. The bias from use of SIPP data is generally substantially greater than the bias that arises from using the ENP data (compare the biases in Table XIVA and XIVB with the biases in Table XI).

Unlike the SIPP sample, the ENP sample was drawn from the same geographic locations as program participants and was administered the same survey questionnaire. To isolate the effect of geographic mismatch in producing

[53] Our data are collected in the format of the NLSY. For elaboration of these issues, see Smith (1995).

[54] See Devine and Heckman (1996) for an analysis of eligibility for the JTPA program.

TABLE XIVA

EFFECTS OF JTPA ELIGIBILITY ON ESTIMATED SELECTION BIAS USING SIPP SAMPLES[a, b]
Quarterly Earnings Expressed in Monthly Dollars SIPP Full Sample, Ashenfelter and Card Sample and Eligible Sample[c]
Adult Males

	Full Sample			Sample Constructed Using Criteria Used By Ashenfelter and Card (1985)			Eligible Sample		
Quarter	Difference in Means B	Local Linear Matching[d] $\overline{B}_{S_P}$	Regression-Adjusted Local Linear Matching[e] $\overline{B}_{S_P}(adj)$	Difference in Means B	Local Linear Matching $\overline{B}_{S_P}$	Regression-Adjusted Local Linear Matching $\overline{B}_{S_P}(adj)$	Difference in Means B	Local Linear Matching $\overline{B}_{S_P}$	Regression Adjusted Local Linear Matching $\overline{B}_{S_P}(adj)$
Qtr1	−1537 (37)	234 (83)	88 (97)	−1111 (36)	302 (67)	289 (72)	−43 (68)	317 (54)	249 (77)
Qtr2	−1587 (38)	188 (81)	45 (94)	−1150 (36)	248 (68)	233 (74)	−119 (62)	201 (66)	123 (79)
Qtr3	−1613 (37)	115 (82)	−25 (98)	−1180 (36)	174 (71)	159 (79)	−167 (55)	157 (65)	76 (81)
Qtr4	−1648 (39)	69 (84)	−73 (103)	−1218 (36)	117 (73)	99 (83)	−237 (63)	102 (74)	13 (93)
Average of 1 to 4	−1596 (37)	151 (80)	9 (96)	−1165 (35)	210 (68)	195 (75)	−142 (56)	194 (60)	115 (78)
As a % of impact	3669%	348%	20%	2677%	484%	448%	325%	446%	265%

[a] Only four quarters of data are available due to the short length of the SIPP Panel. Full sample includes all SIPPs with non-missing data in month 12 of the 24 month 1988 SIPP Full Panel. Ashenfelter and Card sample consists of SIPP sample members with annual earnings below $20,000 (1975 dollars), household income below $30,000 (1975 dollars) and who were in the labor force in month 12 of the 24 month 1988 SIPP Full Panel. Eligible sample consists of SIPP sample members eligible for JTPA in month 12 of the 24 month 1988 Full Panel (see Appendix B for details).

[b] Bootstrap standard errors are shown in parentheses. They are based on 50 replications with 100% sampling.

[c] Full sample proportion of controls in common support is 0.19, proportion of SIPP is 0.97. Ashenfelter and Card sample proportion of controls in common support is 0.22, proportion of SIPP is 0.97. Eligible sample proportion of controls in common support is 0.41, proportion of SIPP is 0.96.

[d] The probability of program participation logit includes age, race, education, marital status, child age less than six, labor force transitions and earnings in the preceding year. The probability of program participation is estimated separately for each sample. Variables included in the outcome equation are race, age, education, marital status, children less than six, and indicators for season and year.

[e] 2% trimming is used to estimate the region of overlapping support. A fixed bandwidth of 0.06 is used for the nonparametric estimates (see Appendix A for more details).

TABLE XIVB

EFFECTS OF JTPA ELIGIBILITY ON ESTIMATED SELECTION BIAS USING SIPP SAMPLES[a, b]
Quarterly Earnings Expressed in Monthly Dollars, SIPP Full Sample, Ashenfelter and Card Sample and Eligible Sample[c]
Adult Males

	Full Sample			Sample Constructed Using Criteria Used By Ashenfelter and Card (1985)			Eligible Sample		
Quarter	Difference in Means	Difference-in Differences for Local Linear Matching[d]	Difference-in-Differences for Regression-Adjusted Local Linear Matching[e]	Difference in Means	Difference-in-Diffeences for Local Linear Matching	Difference-in-Differences for Regression-Adjusted Local Linear Matching	Difference in Means	Difference-in-Differences for Local Linear Matching	Difference-in Differences for Regression-Adjusted Local Linear Matching
Qtr1	−1537 (37)	−35 (25)	−33 (30)	−1111 (36)	−27 (26)	−24 (32)	−43 (68)	−80 (34)	−97 (38)
Qtr2	−1587 (38)	−143 (26)	−140 (36)	−1150 (36)	−138 (23)	−137 (31)	−119 (62)	−199 (45)	−230 (51)
Qtr3	−1613 (37)	−264 (26)	−240 (41)	−1180 (36)	−256 (31)	−246 (41)	−167 (55)	−234 (42)	−277 (52)
Qtr4	−1648 (39)	−314 (32)	−305 (53)	−1218 (36)	−319 (35)	−326 (42)	−237 (63)	−283 (53)	−338 (72)
Average of 1 to 4	−1596 (37)	−189 (20)	−180 (31)	−1165 (35)	−185 (20)	−183 (28)	−142 (56)	−199 (33)	−236 (45)
As a % of impact	3669%	434%	413%	2677%	426%	421%	325%	457%	542%

[a]Only four quarters of data are available due to the short length of the SIPP Panel. Full sample includes all SIPPs with nonmissing data in month 12 of the 24 month 1988 SIPP Full Panel. Ashenfelter and Card sample consists of SIPP sample members with annual earnings below $20,000 (1975 dollars), household income below $30,000 (1975 dollars), and who were in the labor force in month 12 of the 24 month 1988 SIPP Full Panel. Eligible sample consists of SIPP sample members eligible for JTPA in month 12 of the 24 month 1988 SIPP Full Panel (see Appendix B for details).

[b]Bootstrap standard errors are shown in parentheses. They are based on 50 replications with 100% sampling.

[c]Full sample proportion of controls in common support is 0.19, proportion of SIPP is 0.97. Ashenfelter and Card sample proportion of controls in common support is 0.22, proportion of SIPP is 0.97. Eligible sample proportion of controls in common support is 0.41, proportion of SIPP is 0.96.

[d]The probability of program participation logit includes age, race, education, marital status, child age less than six, labor force transitions, and earnings in the preceding year. The probability of program participation is estimated separately for each sample. Variables included in the outcome equation are race, age, education, marital status, children less than six, and indicators for season and year.

[e]2% trimming is used to estimate the region of overlapping support. A fixed bandwidth of 0.06 is used for the nonparametric estimates (see Appendix A for more details).

TABLE XV

EFFECT OF GEOGRAPHY ON ESTIMATED BIAS
COMPARING CONTROLS AT TWO SITES TO ELIGIBLE NON-PARTICIPANTS AT TWO SITES
Earnings in the 18 Months After Random Assignment
Quarterly Earnings Expressed in Monthly Dollars
Elig. Nonparticipant (ENP) Sample at Corpus Christi and Fort Wayne
Experimental Control Sample at Jersey City and Providence
Adult Males, 149 Controls and 276 ENPs

Quarter	Difference in Means B	Local Linear Matching[a] $\overline{B}_{S_P}$	Regression-Adjusted Local Linear Matching $\overline{B}_{S_P}(adj)$	Difference-in differences for Local Linear Matching	Difference-in-differences for Regression-Adjusted Local Linear Matching
Qtr1	−534 (53)	−203 (85)	−184 (110)	−143 (111)	−135 (126)
Qtr2	−504 (73)	−166 (107)	−154 (120)	−125 (118)	−72 (130)
Qtr3	−515 (78)	−177 (120)	−147 (127)	−73 (131)	−9 (141)
Qtr4	−485 (78)	−200 (121)	−164 (132)	−87 (141)	19 (151)
Qtr5	−527 (72)	−272 (127)	−211 (132)	−254 (160)	−136 (167)
Qtr6	−524 (75)	−281 (110)	−189 (112)	−257 (162)	−82 (165)
Average of 1 to 6	−515 (63)	−216 (95)	−175 (108)	−157 (110)	−69 (123)
As a % of impact	1183%	497%	402%	360%	159%

[a] 2% trimming is used to estimate the overlapping support region. A fixed bandwidth of 0.06 is used for the nonparametric estimates. (See Appendix A for more details on the estimation procedure.) Bootstrap standard errors are shown in parentheses. They are based on 50 replications with 100% sampling.

bias, and to evaluate the effectiveness of econometric methods in reducing the bias, we scramble the ENP-control data and mismatch by geography within these samples. Since all observations are administered the same questionnaire, this enables us to estimate a pure geographic mismatch effect. Table XV reports the result of matching ENPs ($D = 0$) in two training centers to controls ($D = 1$) from two other training centers. For three of the estimators, the bias B in Table XV is two or three times as large as the bias in the geographically-aligned data (compare with the results in Table XI). Matching and regression-adjusted matching reduce, but by no means eliminate, the bias (compare the second and third columns of Table XV with the second and third columns of Table XI). When data are geographically misaligned, the difference-in-differences estimators generally perform better than the cross-sectional estimators. Geographic mismatch is an important source of bias in evaluating training programs.[55,56]

[55] Roselius (1996) builds on our analysis and creates a variety of SIPP samples using alternative definitions of region and city size. She finds substantial bias in all of her SIPP samples that is far in excess of the ENP-control bias reported in the text. Adjusting for labor market variables like the unemployment rate in the state or metropolitan statistical area does not reduce the bias she estimates.

[56] Smith (1995) uses other data sources and considers the consequences of alternative definitions of variables and survey instruments on the estimated bias.

Access to comparison samples of persons who are administered the same questionnaire and located in the same labor market as participants greatly improves the quality of nonexperimental evaluations. Econometric methods generally reduce, but do not eliminate, these sources of bias and are no panacea for the problems created by using bad data to evaluate social programs.

10. THE CONSEQUENCES OF *P*-DEPENDENCE OF THE IMPACTS

If the program impact $E(Y_1 - Y_0 \mid P, D = 1)$ depends on P, then econometric methods applied to nonexperimental comparison groups that have P support in regions different from the support of the participant group estimate a parameter that differs from what is estimated by an ideal experiment. This is true even if there is no selection bias so that $B(P(X)) = 0$ everywhere. This section presents evidence on this additional source of bias.

Using data on eighteen-month outcomes from the treatment and control groups of the JTPA experiment, we use local linear regression methods to determine how $E(Y_1 - Y_0 \mid P, D = 1)$ depends on P. The estimates are graphed in Figure 5. The point estimates suggest a modest dependence in the neighborhood of $P = 0.15$, but the formal statistical test whose results we report in Table XVI does not allow us to reject the null hypothesis of no dependence.[57]

However, measuring the program impact only over the limited support of the overlap set S_P adds an additional $-\$19$ to the bias arising from using a nonexperimental estimator adapted to a common support. The overall impact estimated over S_P is \$38 per month. The overall impact for the program estimated without any restriction on the support is \$57 per month. Thus the restriction to a common support reduces the estimated program impact by 33%. The difference between the two estimates of program impact is statistically significant. (See Table XVII.) A major lesson of this paper for the design of future evaluations is that comparison groups should be selected to have P distributions similar to those of program participants in order to mitigate the support problem.

11. IMPLEMENTATION OF ESTIMATORS WITH ORDINARY NONEXPERIMENTAL SAMPLES

The methodologies that we have devised to estimate the bias in samples that combine experimental and nonexperimental data can also be applied to ordinary nonexperimental samples to estimate a variety of evaluation parameters of interest. For the nonparametric sample selection estimator, the only new ingredient that is required is an exclusion restriction—at least one variable in Z not in R—that satisfies certain conditions specified below.

[57] The test statistic is formally equivalent to the test for index sufficiency of the outcome differences for a model with $R = 1$.

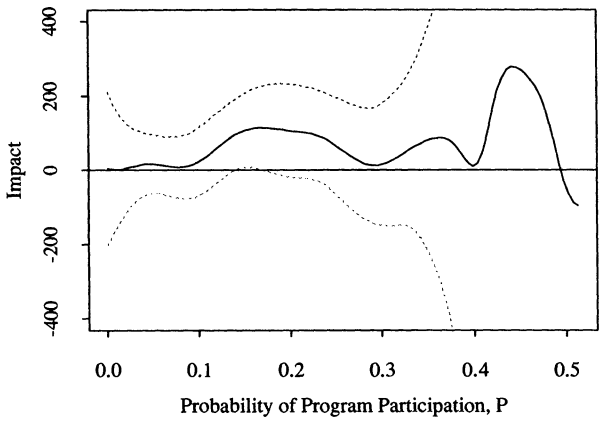


FIGURE 5.—Adult males, experimental treatments and controls, P dependence of treatment impacts, $E(Y_1 - Y_0 \mid P)$, bandwidth = 0.03.

TABLE XVI

p-VALUES AND POINT ESTIMATES FROM TESTS OF P-DEPENDENCE OF TREATMENT IMPACTS

H_0: $E(Y_1 - Y_0 \mid P, D = 1) - E(Y_1 - Y_0 \mid D = 1) = 0$

Experimental Control and Treatment Samples

Average Monthly Earnings Over the First Six Quarters After Random Assignment, Adult Males, 649 Controls and 1478 Treatments

Test Values of P	p-values[a]	Point Estimates[b]
0.00040	0.9607	−29
0.00081	0.9587	−31
0.0020	0.9529	−36
0.0024	0.9511	−38
0.0095	0.8952	−50
0.0159	0.7765	−47
0.0315	0.5760	−28
0.0494	0.7137	−25
0.0691	0.3765	−41
0.0970	0.6485	−38
0.1272	0.4745	38
0.1632	0.2647	74
0.2119	0.4271	57
0.2712	0.8116	−20

[a]A bandwidth equal to 0.06 and a biweight kernel were used for the nonparametric estimates (see Appendix A, Sections A.1 for additional details concerning the estimation procedure). The distribution of the test statistic is chi-squared with one degree of freedom under the null.

[b]Values shown are the difference between the conditional and unconditional estimated means.

TABLE XVII

COMPARISON OF MONTHLY IMPACTS ESTIMATED
OVER THE ENTIRE SUPPORT AND OVER THE RESTRICTED SUPPORT[a]
Quarterly Earnings Expressed in Monthly Dollars
Experimental Control and Treatment Samples
Adult Males, 649 Controls and 1478 Treatments

Quarter	Estimated Impact Using Entire Support[b]	Estimated Impact Using Restricted Support	Difference
Qtr1	4	18	−15
	(30)	(33)	(12)
Qtr2	26	42	−16
	(38)	(36)	(11)
Qtr3	51	69	−18
	(36)	(36)	(13)
Qtr4	57	85	−28
	(42)	(51)	(15)
Qtr5	39	58	−18
	(39)	(36)	(14)
Qtr6	49	71	−23
	(44)	(35)	(15)
Average of 1 to 6	38	57	−20
	(16)	(16)	(5)

[a]Bootstrap standard errors are shown in parentheses. They are based on 50 replications with 100% sampling.
[b]In our data the experimental control group was administered a long baseline survey that gathered five years of retrospective data while the experimental treatment group was not (see "Appendix B). Since information on recent labor force status and on recent earnings is missing for treatments, we are only able to obtain coarse estimates of P for the treated group. In particular, we use the coarse II model described in the notes to Table XIII. The support region in the nonexperimental analysis is determined using the best predictor P model, so it is necessary to estimate which treatment group members would be excluded by the common support restriction in order to obtain impact estimates within the support region that would be estimated by a nonexperimental method. The adjusted treatment impacts were obtained as follows. For controls and treatments, we first divide the coarse P distribution into 20 equal-size bins, then within-bin treatment impacts are estimated. The average unadjusted impact estimate is obtained as the weighted average of the within-bin estimates, with weights given by the proportion of controls within each bin. The adjusted impact estimate is equal to the weighted average of the within-bin estimates, with the weights given by the proportion of controls within each bin after deleting controls whose values of P lie outside the overlap region.

Consider equation system (8) and suppose that index sufficiency characterizes the bias term and that

$$E(Y_1 \mid Z, R, D = 1) = g_1(R) + E(U_1 \mid P(Z), D = 1) \tag{23a}$$

and

$$E(Y_0 \mid Z, R, D = 0) = g_0(R) + E(U_0 \mid P(Z), D = 0). \tag{23b}$$

If there is at least one element in Z not in R that satisfies the conditions

$$\lim_{Z \to Z^{c_1}} E(U_1 \mid P(Z), D = 1) = 0 \tag{24a}$$

and

$$\lim_{Z \to Z^{c_0}} E(U_0 \mid P(Z), D = 0) = 0, \tag{24b}$$

where Z^{c_0} and Z^{c_1} may be values or sets of values, and need not be the same sets of values, we can identify $g_1(R)$ and $g_0(R)$ following the argument in

Heckman (1990a, b). This enables us to construct $E(Y_1 - Y_0 \mid R)$ and $E(Y_1 - Y_0 \mid R, P(Z), D = 1)$. To see how to construct the latter, observe that the left-hand sides of (23a) and (23b) can be constructed from sample data using, e.g., local linear regression methods. If (24b) holds, and $E(U_0 \mid P(Z)) = 0$, we can use the iterated expectation argument and construct

$$E(U_0 \mid P(Z), D = 1) = -E(U_0 \mid P(Z), D = 0)\left[\frac{1 - P(Z)}{P(Z)}\right]$$

$$= -[E(Y_0 \mid Z, R, D = 0) - g_0(R)]\left[\frac{1 - P(Z)}{P(Z)}\right].$$

Thus we can construct

$$E(Y_1 \mid Z, R, D = 1) - [g_0(R) + E(U_0 \mid P(Z), D = 1)]$$
$$= E(Y_1 - Y_0 \mid Z, R, D = 1).$$

Observe that condition (24a) is not required. The empirical evidence on the support of P presented in this paper suggests that producing the sample counterparts to (24b) or (24a) may be difficult in practice.

Cross-sectional matching and difference-in-differences methods considered in this paper can be applied as formulated to nonexperimental data.[58] They do not require the limit sets defined by (24a) and (24b).

12. SUMMARY, SYNTHESIS, AND CONCLUSIONS

This paper develops a framework for combining experimental and nonexperimental data to test the identifying assumptions that justify three widely-used nonexperimental methods of evaluating social programs based on comparison groups: (i) the method of matching; (ii) the classical econometric selection bias model which represents the bias solely as a function of the probability of participation P; and (iii) the method of difference-in-differences.

We decompose the conventional measure of bias into three components corresponding to (a) differences in the supports of the regressors between participants and members of the comparison group; (b) differences in the shapes of the distributions of the regressors in the two groups in the region of common support; and (c) selection bias, rigorously defined at common values of the regressors for both groups. The first two components are eliminated by matching on characteristics that are "close" in the two groups. Only the third component —selection bias—remains.

We apply our methods to unusually rich data from the control group of a random experiment on a prototypical job training program combined with a

[58] As noted by Heckman and Smith (1996), the difference-in-differences estimator identifies the "treatment on the treated" parameter only when no baseline observations have received treatment. For the general case, see their paper.

nonexperimental comparison group of nonparticipants. Our decomposition reveals that selection bias rigorously defined is generally the smallest of the three components of bias as conventionally measured but it is still a substantial fraction of the experimentally-determined impact of the program we study. In our data, both of the forms of matching we examine reduce but do not eliminate the conventional measure of bias. Matching cannot eliminate a nonzero selection bias, rigorously defined, and in fact the method is based on the assumption that it is zero. In related work, Heckman, Ichimura, and Todd (1997; first draft 1993) find that for other demographic groups, matching sometimes increases the estimated bias, at least for some sets of conditioning variables.

Our data are consistent with the index sufficiency assumption that underlies the classical selection bias model. This model cannot be implemented semiparametrically in our data because the support of P is limited. To apply the method semiparametrically in future evaluations, it is necessary to enlarge the support of P for comparison group members so that it matches the full support of participants ($P \in (0, 1)$).

Our data are also consistent with the identifying assumptions required to justify application of a conditional version of the method of difference-in-differences to the evaluation of job training programs for all but low values of P. The conditional difference-in-differences estimator is consistent with the index-sufficient model of selection bias and only requires that the bias be the same before and after the date of the program participation decision, or at least be the same in symmetric intervals around the date of the program participation decision.

The method of matching and the classical selection bias model share one important feature: under the assumptions that justify each method, selection bias $B(X)$ averages out to zero *over certain intervals*. Matching is based on the assumption that selection bias is *zero* for all intervals, however small. Our tests clearly reject this assumption, which also underlies the regression method advocated by Barnow, Cain, and Goldberger (1980). The cross-section bias detected in our analysis is characterized by a *crossing property*. Sizeable negative bias in some cells or intervals is offset by sizeable positive bias in other cells or intervals. A weighted average across cells can reduce the overall bias substantially. This is why some form of matching reduces the bias in our sample, although it does not eliminate it.

As shown in Figure 3, estimated selection bias as a function of P is sizeable, especially in the vicinity of $P = 0$. In that neighborhood, the shape is broadly consistent with the form of the classical selection bias displayed in Figure 1. However, our analysis rejects the application of the normal selection bias model of Heckman (1979). The dashed lines in Figure 3 reveal a large difference between the estimates of selection bias obtained using the nonparametric methods developed in this paper and the classical parametric selection bias model based on the inverse Mills' ratio.

We also demonstrate the substantial benefits of having access to nonexperimental data that (a) place nonparticipants in the same labor markets as pro-

gram participants; (b) administer the same questionnaire to both groups; and (c) include information on recent labor force status histories. Recent labor force status transitions turn out to be more important predictors of program participation than the recent earnings histories emphasized in the analysis of Ashenfelter (1978). Failure to use comparison groups of persons situated in the same labor markets as participants and administered the same questionnaires contributes substantially to the bias as conventionally measured. These sources of bias are empirically more important than selection bias, rigorously defined. Access to recent labor force histories in estimating the probability of program participation considerably improves the performance of nonexperimental methods. These findings enhance our ability to design future nonexperimental evaluations of training programs. Since the JTPA program we consider is typical of a variety of training programs in place around the world, the lessons from our study apply more generally. (See Heckman, LaLonde, and Smith (1999).)

Although further testing with larger samples would be highly desirable, our analysis suggests that semiparametric sample selection bias methods of the sort proposed by Heckman (1980), Cosslett (1991), and Ahn and Powell (1993) are one potentially promising method for evaluating training programs provided that comparable data are collected on nonparticipants and participants located in the same geographic areas and administered the same questionnaire and provided that the support of the distribution of P for nonparticipants is enlarged. Labor force status history variables, local labor market variables and personal characteristics that determine participation (i.e., Z variables) but are excluded from the outcome equations are valid exclusion restrictions for identifying the semiparametric selection model. The temporal structure of the program makes some of the Z and R variables distinct.

Another very promising method that does not require an exclusion restriction is our extension of the method of difference-in-differences. Conditioning on P, the bias function $B_t(P)$ tends to be constant over all time periods t, except possibly for low values of P in time periods near the date of random assignment or eligibility determination. It is for this reason that the index sufficient selection model and our conditional version of the method of difference-in-differences are consistent with each other.

We stress the importance of collecting information on recent labor force status histories and of designing nonparticipant samples so that the distributions of P have the same support for both participants and nonparticipants. It is essential to get the full support to identify parameters (1) and (2) for the entire population of participants.[59] Lack of common support—comparing the incomparable—is a major source of selection bias as it is conventionally measured. Our evidence leads us to a rigorous reformulation of the definition of selection

[59] In practical terms, for training programs such as JTPA, stratified sampling of nonparticipants based on their labor force status or labor force status histories seems a promising strategy. The original ENP data collection plan called for stratification on labor force status, but this plan was abandoned for cost reasons.

bias so that it excludes bias arising from gaps in the common support and from differences in the weights applied to participant and comparison group samples over the region of common support.

Using a common support and a common set of weights applied to participant and comparison group samples goes a long way toward improving the performance of any econometric evaluation estimator. Table XVIIIA clearly demonstrates this point. Column (1) presents the raw bias ($\hat{B}$) quarter-by-quarter and overall using the means for the control and ENP samples. Column (2) shows how the bias is reduced simply by matching to the nearest neighbor using P. (Recall that nearest neighbors can be far apart.) Column (3) shows how the imposition of the common support condition improves the nearest-neighbor matching estimator. Quarter-by-quarter, there is a substantial reduction in bias. However, the overall average is slightly higher in (3). Column (4) presents estimates of the bias that arise from local linear matching (on P) while column (5) presents the estimates that arise from regression-adjusted local linear matching. Both procedures impose common support and common weighting and both improve over the raw mean or crude nearest-neighbor estimators.

TABLE XVIIIA

COMPARISON OF ESTIMATED MEAN BIAS
UNDER ALTERNATIVE ESTIMATORS OF MEAN PROGRAM IMPACTS[a]
Quarterly Earnings Expressed in Monthly Dollars
Adult Males, 508 Experimental Controls and 388 Elig. Nonparticipants (ENPs)

Quarter	Difference in Means (1)[b]	Nearest Neighbor w/o Common Support (2)	Nearest Neighbor w/Common Support (3)	Local Linear Matching (4)	Regression-Adjusted Local Linear Matching (5)
Qtr1	−418 (38)	221 (56)	123 (67)	33 (59)	39 (60)
Qtr2	−349 (47)	−166 (151)	77 (83)	37 (61)	39 (64)
Qtr3	−337 (55)	−58 (206)	53 (96)	29 (78)	21 (80)
Qtr4	−286 (57)	161 (178)	86 (96)	80 (77)	65 (82)
Qtr5	−305 (57)	167 (196)	87 (100)	64 (77)	50 (83)
Qtr6	−328 (63)	45 (191)	34 (113)	37 (82)	17 (90)
Average of 1 to 6	−337 (47)	62 (127)	77 (80)	47 (60)	39 (64)
As a % of impact	775%	142%	176%	107%	88%

[a] Bootstrap standard errors are shown in parentheses. They are based on 50 replications with 100% resampling.

[b] The estimates for each column are defined as follows:

(1) $\hat{B} = \hat{E}(Y_{0t} \mid D = 1) - \hat{E}(Y_{0t} \mid D = 0)$, where $\hat{E}$ denotes the sample mean.

(2) $\hat{B} = \hat{E}_{f(P \mid D=1)}(Y_0 \mid D = 1) - \hat{E}_{f(P \mid D=1)}(Y_0 \mid D = 0)$ where $\hat{E}_{f(P \mid D=1)}(Y_0 \mid D = 1)$ is the sample mean of $\{D = 1\}$ outcomes and $\hat{E}_{f(P \mid D=1)}(Y_0 \mid D = 0, P)$ the sample mean of nearest neighbor matched $\{D = 0\}$ outcomes. The nearest neighbor match for each observation in $\{D = 1\}$ is the observation in $\{D = 0\}$ that is closest in terms of P. Matching is done with replacement. (See Section 3.1 in the text.)

(3) $\hat{B}_{S_P} = \hat{E}_{f(P \mid P \in S_P, D=1)}(Y_0 \mid P \in S_P, D = 1) - \hat{E}_{f(P \mid P \in S_P, D=1)}(Y_0 \mid P \in S_P, D = 0)$. Same estimator as (2) except that matches are only constructed within the region of overlapping support S_P, which is precisely defined in Appendix A.

(4) Estimates are constructed using local linear regression on P, as described in the text. There are no variables in the outcome equation. (See Section 5.0 in the text.)

(5) $\hat{B}_{S_P(adj)} = \hat{E}_{f(P \mid P \in S_P, D=1)}(Y_0 - R\hat{\beta} \mid P \in S_P, D = 1) - \hat{E}_{f(P \mid P \in S_P, D=1)}(Y_0 - R\hat{\beta} \mid P \in S_P, D = 0)$. This is the same estimator as in (4) except matching is performed on the residuals $Y_0 - R\hat{\beta}$ instead of on outcomes Y_0. (See Section 5.0 in the text.) The following regressors R are included in the outcome equation: dummy variables for training center, race, schooling, age, previous training, work experience in months, local unemployment rate, marital status, presence of a child age less than six, and quarter and year effects.

TABLE XVIIIB

COMPARISON OF ESTIMATED MEAN BIAS
UNDER ALTERNATIVE ESTIMATORS OF MEAN PROGRAM IMPACTS[a]
Quarterly Earnings Expressed in Monthly Dollars
Adult Males, 508 Experimental Controls and 388 Elig. Nonparticipants (ENPs)

Quarter	Difference-in-Differences w/o Common Support (1)[b]	Conditional on P Difference-in-Differences w/Common Support (2)	Regression-Adjusted Conditional on P Difference-in-Differences w/Common Support (3)
Qtr1	172 (42)	97 (62)	104 (63)
Qtr2	142 (47)	77 (89)	77 (92)
Qtr3	41 (56)	90 (114)	74 (114)
Qtr4	43 (61)	112 (90)	98 (91)
Qtr5	−54 (63)	19 (95)	−5 (99)
Qtr6	−111 (64)	4 (105)	−35 (111)
Average of 1 to 6	39 (47)	67 (71)	52 (74)
As a % of impact	89%	153%	120%

[a]Bootstrap standard errors are shown in parentheses. They are based on 50 replications with 100% sampling.
[b]The estimates for each column are defined as follows:
(1) $\hat{B}_D = \hat{E}(Y_{0,t} \mid D=1) - \hat{E}(Y_{0,(-t)} \mid D=1) - [\hat{E}(Y_{0,t} \mid D=0) - \hat{E}(Y_{0,(-t)} \mid D=0)]$, where $\hat{E}$ denotes the sample mean.
(2) $\hat{B}_{D,S_P} = \hat{E}_{f(P \mid P \in S_P, D=1)}(Y_{0,t} \mid P \in S_P, D=1) - \hat{E}_{f(P \mid P \in S_P, D=1)}(Y_{0,t} \mid P \in S_P, D=0) - [\hat{E}_{f(P \mid P \in S_P, D=1)}(Y_{0,(-t)} \mid P \in S_P, D=1) - \hat{E}_{f(P \mid P \in S_P, D=1)}(Y_{0,(-t)} \mid P \in S_P, D=0)]$, where $\hat{E}_{f(P \mid P \in S_P, D=1)}(Y_{0,t} - Y_{0,(-t)} \mid P \in S_P, D=1)$ is the sample mean of the $D=1$ outcomes, and $E_{f(P \mid P \in S_P, D=1)}(Y_{0,t} - Y_{0,(-t)} \mid P \in S_P, D=0)$ is the sample mean of the $D=0$ matched outcomes. Matches are constructed by local linear regression on P as described in the text (see Section 5.0 in the text). The model does not include regressors in the outcome model.
(3) Same as (2) except the following regressors are included in the outcome equation: training site, age, education, marital status, children less than 6 indicator, indicator for currently enrolled in training, labor market experience, local unemployment rate, season and year.

Similar patterns appear in Table XVIIIB for the difference-in-differences estimator. Simple differencing symmetrically before and after the date of random assignment or eligibility determination eliminates person-specific components of bias. Compare column (1) of that table with column (1) of Table XVIIIA. Imposing common support and common density in column (2) generally reduces the quarter-by-quarter bias. However, as we found for the nearest neighbor estimator, the overall average bias is slightly higher. Using regressors to adjust for the bias reduces it slightly as shown in column (3). Note in comparing Tables XVIIIA and XVIIIB that the overall bias from our conditional difference-in-differences estimator and from the cross-sectional matching estimator are of the same order of magnitude. Column (3) of Table XVIIIC reveals that even though the inverse Mills' ratio as typically applied is badly biased (see the estimates in the first column), weighting by a common density ($f(P \mid D=1)$) greatly improves the performance of the estimator.[60] Imposing

[60] For column (3), the ENP observations (for which $D=0$) in the regression are weighted by the ratio $\hat{f}(P \mid D=1)/\hat{f}(P \mid D=0)$, where the densities are estimated by standard kernel methods. Imposing the common support condition ensures that the weights are nonzero. The control observations are self-weighting by the $f(P \mid D=1)$ distribution.

TABLE XVIIIC

COMPARISON OF ESTIMATED MEAN BIAS
UNDER ALTERNATIVE ESTIMATORS OF MEAN PROGRAM IMPACTS[a]
Quarterly Earnings Expressed in Monthly Dollars
Adult Males, 508 Experimental Controls and 388 Elig. Nonparticipants (ENPs)

Quarter	Inverse Mills' Ratio w/o Common Support w/o Density Weighting (1)[b]	Inverse Mills' Ratio w/Common Support w/o Density Weighting (2)	Inverse Mills' Ratio w/Common Support w/Density Weighting (3)
Qtr1	−611 (86)	−619 (161)	−147 (176)
Qtr2	−515 (95)	−403 (194)	3 (220)
Qtr3	−498 (96)	−365 (190)	30 (215)
Qtr4	−494 (97)	−421 (191)	−80 (215)
Qtr5	−511 (98)	−441 (190)	−69 (215)
Qtr6	−499 (102)	−323 (196)	48 (222)
Average of 1 to 6	−521 (86)	−553 (161)	−36 (37)
As a % of impact	1198%	985%	83%

[a]Bootstrap standard errors are shown in parentheses. They are based on 50 replications with 100% sampling.

[b]The estimates for each column are defined as follows:

(1) $\hat{B}_{\lambda} = \hat{E}(\hat{\lambda}_1(P) \mid D = 1) - \hat{E}(\hat{\lambda}_0(P) \mid D = 0)$, where $\hat{E}$ denotes the sample mean, $\hat{\lambda}_1$ is an estimator of $E(U_0 \mid X, D = 1)$ obtained under the Mills' ratio assumption and $\hat{\lambda}_0$ is an estimator of $E(U_0 \mid X, D = 1)$.

(2) $\hat{B}_{\lambda, S_P} = \hat{E}(\hat{\lambda}_1(P) \mid P \in S_P, D = 1) - \hat{E}(\hat{\lambda}_0(P) \mid P \in S_P, D = 0)$, where $\hat{E}$ denotes the sample mean. $\hat{\lambda}_1$ is an estimator of $E(U_0 \mid X, D = 1)$ and $\hat{\lambda}_0$ is an estimator of $E(U_0 \mid X, D = 0)$ obtained under the Mills' ratio assumption. (Same as (1) except mean is only taken over observations in the overlapping support, S_P.)

(3) $\hat{\bar{B}}_{\lambda, S_P} = \hat{E}(\hat{\lambda}_1(P) \mid P \in S_P, D = 1) - \hat{E}[(\hat{f}(P \mid D = 1)/\hat{f}(P \mid D = 0))\hat{\lambda}_1(P) \mid P \in S_P, D = 1]$. This estimator is same as (2) except with density weighting as described in the text.

common support alone without reweighting does not lead to substantial improvement, as shown in column (2).

It is instructive to contrast the biases defined over a common support and with common weighting with the biases defined in the conventional way (e.g., as in Ashenfelter (1978) or LaLonde (1986)). One conventional measure of bias is the OLS estimate of π in the model

$$Y = g(X) + D\pi + U,$$

applied to controls and comparison group members, where $g(X)$ depends on the specification used. The normal selection bias method introduces the inverse Mills' ratio terms into $g(X)$ in conducting a cross-section analysis. The difference-in-differences method uses Y or regression-adjusted Y differenced symmetrically around the date of random assignment or eligibility determination. Estimates of π reveal the bias in the conventional common coefficient model ($U_0 = U_1$), where the program impact is assumed not to depend on X. This estimate of bias combines the three sources of bias distinguished in this paper plus any bias arising from correlation between U_0 and X.[61] In contrast, estimates of the bias that condition on a common support and impose a common weighting of participant and comparison group data produce an estimate of selection bias as rigorously defined in this paper.

[61] Heckman and Todd (1994) decompose the bias π for the model with $g(X) = X\beta$ and present the contribution for the case where U_0 is correlated with X.

TABLE XIX

COMPARISON OF ESTIMATED MEAN BIAS
UNDER ALTERNATIVE ESTIMATORS OF MEAN PROGRAM IMPACTS[a]
Quarterly Earnings Expressed in Monthly Dollars
Adult Males, 508 Experimental Controls and 388 Elig. Nonparticipants (ENPs)

Quarter	Difference in Means (1)[b]	Nearest Neighbor w/o Common Support (2)	Method of Barnow, Cain and Goldberger w/o Common Support w/o Density Weighting (3)	Difference-in-Differences w/o Common Support (4)	Inverse Mills' Ratio w/o Common Support w/o Density Weighting (5)
Qtr1	−418 (38)	221 (56)	−15 (47)	173 (42)	−611 (86)
Qtr2	−349 (47)	−166 (151)	53 (55)	142 (47)	−515 (95)
Qtr3	−337 (55)	−58 (206)	62 (58)	40 (56)	−498 (96)
Qtr4	−286 (57)	161 (178)	107 (60)	43 (61)	−495 (97)
Qtr5	−305 (57)	167 (196)	94 (62)	−54 (63)	−511 (98)
Qtr6	−328 (63)	45 (191)	54 (62)	−111 (64)	−499 (102)
Average of 1 to 6	−337 (47)	62 (127)	62 (58)	39 (47)	−521 (86)
As a % of impact	775%	143%	143%	90%	1198%

[a]Bootstrap standard errors are shown in parentheses. They are based on 50 replications with 100% sampling.
[b]The estimates for each column are defined as follows:
(1) $\hat{B} = \hat{E}(Y_0 \mid D = 1) - \hat{E}(Y_0 \mid D = 0)$ where $\hat{E}$ denotes the sample mean.
(2) $\hat{B} = \hat{E}_{f(P \mid D=1)}(Y_0 \mid D = 1) - \hat{E}_{f(P \mid D=1)}(Y_0 \mid D = 0)$, where $\hat{E}_{f(P \mid D=1)}(Y_0 \mid D = 1)$ is estimated by the sample mean of $\{D = 1\}$ outcomes and $E_{f(P \mid D=1)}(Y_0 \mid D = 0)$ by the sample mean of $\{D = 0\}$ nearest neighbor matches. (See Section 3.1 in the text.)
(3) Same as described in footnote d in Table XII except without imposing a common support restriction.
(4) $\hat{B}_D = \hat{E}(Y_{0,t} \mid D = 1) - \hat{E}(Y_{0,(-t)} \mid D = 1) - [\hat{E}(Y_{0,t} \mid D = 0) - \hat{E}(Y_{0,(-t)} \mid D = 0)]$, where $\hat{E}$ denotes the sample mean.
(5) $\hat{E}(\hat{\lambda}_1(P) \mid D = 1) - \hat{E}(\hat{\lambda}_0(P) \mid D = 0)$, where $\hat{E}$ denotes the sample mean, $\hat{\lambda}_1$ is the estimator for $E(U_0 \mid R, D = 1)$ under the Mills' ratio assumption, and $\hat{\lambda}_0$ is the estimator of $E(U_0 \mid R, D = 0)$.

The estimates of π for the different methods are presented in Table XIX. Except for the inverse Mills' ratio, the overall biases (π) from the other commonly-used estimators are of the same order of magnitude. All except the inverse Mills' ratio estimator produce biases that are smaller than the raw mean $\hat{B}$. At the same time, all are large relative to the program impact and exhibit substantial variability across quarters. The different sources of bias tend to cancel each other out. This is especially true of the Barnow, Cain, and Goldberger (1980) estimator. (Compare Column (3) of Table XIX with the last column of Table XII).

By decomposing the bias π into its components, we determine whether a small estimated π is due to a fortuitous combination of offsetting biases or whether each component of the bias is small. Sources of bias such as the failure of common support and discrepancies in the weights across participants and comparison group members depend on the sampling plan used to collect the data for the comparison group and so are likely to vary across evaluations. The factors generating self-selection are more likely to be similar across evaluations. The focus in this paper is on the estimation of the stable components of the conventional measure of bias. Knowledge of these components facilitates generalization of the evidence from any one study to other environments, and is more

informative about the sources of bias than the measure B or π traditionally used to summarize bias. Our decomposition demonstrates that in our data, selection bias, rigorously defined, is large relative to experimentally-estimated program impacts but is small relative to the conventional measure of bias.

Our analysis highlights the benefits of randomized trials. While the bias is reduced using nonexperimental methods that impose common support and common weighting, it is not eliminated. Experiments avoid the need to specify precise functional forms of econometric models or to select regressors to appear in outcome or participation equations. Typically, experimental treatment and control groups reside in the same location and are administered the same questionnaires. Experiments solve the problem of common support by balancing the distributions of characteristics between treatments and controls and producing an impact estimate for all P values. However, experiments have their own important limitations (Heckman, LaLonde, and Smith (1999)). If a nonexperimental evaluation method is used, semiparametric selection bias models estimated on data with full support for nonparticipants or conditional difference-in-differences estimators fit outside the period immediately surrounding the period of initial participation in the program appear to be promising methods that deserve much further exploration and testing.

Dept. of Economics, University of Chicago, 1126 E. 59th St., Chicago, IL 60637, U.S.A.,

Dept. of Economics, University of Pittsburgh, 4M35 Forbes Quadrangle, 230 Bonquet St., Pittsburgh, PA 15260, U.S.A.,

Dept. of Economics, University of Western Ontario, Social Science Centre, London, Ontario, N6A 5C2, Canada,

and

Dept. of Economics, University of Pennsylvania, 3718 Locust Walk, Philadelphia, PA 19104, U.S.A.

Manuscript received August, 1994; final revision received August, 1997.

APPENDIX A

A.1. THE LOCAL LINEAR REGRESSION ESTIMATOR

Fan (1992, 1993) develops the distribution theory for the local linear regression estimator of $E(Y \mid P = P_0)$, where Y and P are random variables. The estimator of the expectation at P_0 is defined as $\hat{\gamma}_1$, the solution from the problem

$$\min_{\gamma_1, \gamma_2} \sum_{i \leq N} [Y_i - \gamma_1 - \gamma_2(P_i - P_0)]^2 G((P_0 - P_i)/a_N), \tag{A-1}$$

where $G(\cdot)$ is a kernel function and a_N is a bandwidth parameter. The local linear estimator at each point is obtained by weighted least squares, with greater weight given to points closer to P_0 when G is a symmetric single-peaked function. $\hat{\gamma}_2$ consistently estimates the first derivative of $E(Y \mid P = P_0)$, a result we use below. Higher order derivatives can be consistently estimated under additional smoothness assumptions on $E(Y \mid P = P_0)$ from the coefficients of the higher order terms in $(P_i - P_0)$

in a local polynomial regression. For example, if it exists, the qth-order derivative of the regression function can be estimated as the coefficient on $[(P_i - P_0)^q]/q!$ of the local polynomial regression.

There are several advantages of local linear estimation over standard kernel methods. If $E(Y \mid P = P_0)$ is twice continuously differentiable with respect to P_0, then the bias of the local linear regression estimator is of the same order in the boundary regions of the support of P as it is in the interior regions, whereas the kernel estimator suffers from a lower order bias at boundary points. As shown in Figure 1 in the text, conventional selection bias methods exhibit the greatest bias in the neighborhood of the boundary values $P \in \{0, 1\}$, and as shown in Figure 2 a lot of our data is near $P = 0$, so the better performance of local linear estimators at these values is potentially important for our study. In addition, the first order bias of the local linear estimator does not depend on the distribution of P. This property makes the local linear estimator robust to different distributions of P and produces dramatic simplifications in the distribution theory of our test statistics, compared to what would be obtained from standard kernel methods.

A.2. ESTIMATING THE PARTIALLY LINEAR MODEL

We adopt the following notation. The R variables appear in the outcome equation. The Z variables appear in the probability that $D = 1$, $\Pr(D = 1 \mid Z) = P(Z'\theta)$. In this paper, a logit model is used to estimate P. Estimators are designated by "^", and $\hat{P}_i = P(Z_i'\hat{\theta})$.

A.2.1. *Estimation Method*

The outcome model that we estimate is

$$Y_{it} = R_{it}'\beta + K_{1t}(P_i) + \varepsilon_{it}, \qquad \text{for} \quad i \in \{D = 1\}, \qquad t \in \mathscr{T},$$

$$Y_{it} = R_{it}'\beta + K_{0t}(P_i) + \varepsilon_{it}, \qquad \text{for} \quad i \in \{D = 0\}, \qquad t \in \mathscr{T},$$

where $\{D = 1\}$ is the set of i indices for which $D_i = 1$, $\{D = 0\}$ is the set of i indices for which $D_i = 0$, and $\mathscr{T}$ is the set of time periods used to estimate the model, $\mathscr{T} = \{1, \ldots, T\}$. $N = N_0 + N_1$ and N_0 and N_1 are the number of observations in $\{D = 0\}$ and $\{D = 1\}$, respectively.

We may write these equations as

$$\text{(A-2)} \qquad Y_{it} = R_{it}'\beta + D_i K_{1t}(P_i) + (1 - D_i)K_{0t}(P_i) + \varepsilon_{it}, \qquad t \in \mathscr{T}.$$

In implementing this model, we replace P_i with $\hat{P}_i$. Let $R_i = (R_{i1}, \ldots, R_{iT})'$ denote the matrix of stacked regressors for individual i over all time periods and let $\mathscr{R}_{it} = (R_{i1}, \ldots, R_{it})$ denote the submatrix for individual i through period t. For $t > t'$, we assume, (i) $E\{\varepsilon_{it} \mid \mathscr{R}_{it}, Z_i, D_i\} = 0$, (ii) $E\{\varepsilon_{it}^2 \mid \mathscr{R}_{it}, Z_i, D_i = d\} = \sigma_t^2(\mathscr{R}_{it}, Z_i, D_i = d)$, (iii) $E\{\varepsilon_{it}\varepsilon_{it'} \mid \mathscr{R}_{it}, Z_i, D_i = d\} = \sigma(\mathscr{R}_{it}, Z_i, D_i = d)$. This model is an extension of the partially linear regression model of Wahba (1984) and Robinson (1988).

We first estimate $\hat{P}_i = P(Z_i'\hat{\theta})$ by weighted logistic regression. Using the estimator $\hat{P}_i$, we then estimate β, $K_{1t}(P_i)$, and $K_{0t}(P_i)$. The slope coefficients β are restricted to be the same for observations with $D_i = 0$ and $D_i = 1$ and are assumed constant over time. The nonparametric components K_{1t} and K_{0t} are allowed to vary across groups and over time.

We use the observations for which $D_i = 1$ to nonparametrically estimate $E(Y_{it} \mid P_i, D_i = 1)$ and $E(R_{it} \mid P_i, D_i = 1)$ and observations with $D_i = 0$ to nonparametrically estimate $E(Y_{it} \mid P_i, D_i = 0)$ and $E(R_{it} \mid P_i, D_i = 0)$. Let $\hat{Y}_{itd} = Y_{it} - \hat{E}(R_{it} \mid \hat{P}_i, D_i = d)$ and $\hat{R}_{itd} = R_{it} - \hat{E}(R_{it} \mid \hat{P}_i, D_i = d)$, where $d \in \{0, 1\}$ and we leave the choice of bandwidth a_{N_d} implicit. Throughout this paper $a_{N_0} = a_{N_1} = a_N$. β is estimated by pooling observations across groups over $\mathscr{T}$:

$$\hat{\beta} = \left[\sum_{t \in \mathscr{T}} \left(\sum_{d \in \{0,1\}} \sum_{i \in \{D = d\}} \hat{R}_{itd}\hat{R}_{itd}'\hat{Q}_{di} \right)\right]^{-1} \sum_{t \in \mathscr{T}} \left(\sum_{d \in \{0,1\}} \sum_{i \in \{D = d\}} \hat{R}_{itd}\hat{Y}_{itd}\hat{Q}_{di} \right),$$

where $\hat{Q}_{0i}$ and $\hat{Q}_{1i}$ are indicator functions that exclude a small fraction (2%) of the data with low estimated densities. More precisely, $\hat{Q}_{di} = 1\{\hat{f}(\hat{P}_i \mid D_i = d) > \hat{q}_{2d}\}$, where the estimated density of P_i given $D_i = d$, $\hat{f}(\hat{P}_i \mid D_i = d)$, is obtained using a standard kernel density estimator with the biweight kernel and where $\hat{q}_{2d}$ is the second percentile of the estimates of $\hat{f}(\hat{P}_i \mid D_i = d)$. The expression for the biweight kernel is

$$G(s) = \begin{cases} \frac{15}{16}(s^2 - 1)^2 & \text{for } |s| \leq 1; \\ 0 & \text{otherwise.} \end{cases}$$

Such "trimming" is required to ensure that the nonparametric estimator is uniformly consistent.[62] Estimates of $K_{1t}(P_i)$ and $K_{0t}(P_i)$ are then obtained by local linear regression of $Y_{it} - R'_{it}\hat{\beta}$ on $\hat{P}_i$ performed separately within the $\{D = 1\}$ and $\{D = 0\}$ groups. If β is allowed to vary across different t, the estimator is

$$\hat{\beta}_t = \left[\sum_{d \in \{0,1\}} \sum_{i \in \{D = d\}} \hat{R}_{itd} \hat{R}'_{itd} \hat{Q}_{di} \right]^{-1} \sum_{d \in \{0,1\}} \sum_{i \in \{D = d\}} \hat{R}_{itd} \hat{Y}_{itd} \hat{Q}_{di}.$$

The terms K_{1t} and K_{0t} are then estimated using a local linear regression of $Y_{it} - R'_{it}\hat{\beta}$ on $\hat{P}_i$ for the two groups ($D_i = 0, 1$) for each period.

A.2.2. *Choice of Kernel Function and Smoothing Parameters*

Heckman, Ichimura, and Todd (1996) establish that the choice of $G(\cdot)$ does not affect the asymptotic variance of $\hat{\beta}$ but does affect the variance of the estimator of the nonparametric components K_{0t} and K_{1t}. We use a fixed bandwidth of 0.06 in constructing the estimates of β. The empirical results are not sensitive to perturbations of the bandwidth in the interval [0.04, 0.08].

A.3 ASSUMPTIONS

Heckman, Ichimura, and Todd (1996) establish the asymptotic properties of the estimators and test statistics used in this paper under the following assumptions. Our analysis allows for data to be randomly missing for some quarters. To focus on the main ideas, and to simplify the notation, we abstract from this complication in stating the propositions, but in presenting computational formulae we allow for it.

ASSUMPTION 1: $\{(R_{i,1}, \ldots, R_{i,T}; Y_{i,1}, \ldots, Y_{i,T}; Z_i; D_i)\}_{i \in \{D = d\}}$, $d = \{0, 1\}$ *are independent across individuals i for each d, but data may be correlated across time for each individual.*

ASSUMPTION 2: $P(Z'_i\theta)$ *is twice continuously differentiable with respect to* θ *and both derivatives have finite second moments.*

This condition is satisfied for a logit because the first and second derivatives of the logit CDF are uniformly bounded and because of Assumption 3 which we now present. Let $|\cdot|_2$ denote the Euclidean norm. We make the following assumption:

ASSUMPTION 3: $E\{\sum_{t \in \mathcal{T}}(|R_{it}|_2^{2+\delta} + |Z_i|_2^{2+\delta} + |Y_{it}|_2^3)\} < \infty$ *for some* $\delta > 0$.

[62] The global bandwidth parameter for the density estimates is chosen following the recommendation of Silverman (1986), which in our case gives $a_N = A(\hat{H}/1.34)N^{-1/5}$, where A is a constant that depends on the kernel ($A = 2.7768$) and $\hat{H}$ is the interquartile range of $\hat{P}_i$.

We estimate θ by a weighted logistic likelihood to account for choice-based sampling. (See, e.g., Amemiya (1985).) Let $\hat{\theta}$ denote the estimator of θ. Asymptotic normality for the weighted score vector

$$\psi_i \overset{\text{def}}{=} \psi_i(Z_i, D_i) = E\left\{ N^{-1} \sum_{i \in \{D=0\} \cup \{D=1\}} \frac{\partial^2 \log L_i}{\partial \theta \, \partial \theta'} \right\}^{-1} \frac{\partial \log L_i}{\partial \theta}$$

is assumed where L_i is the contribution of the ith observation to the weighted logistic likelihood, $N = N_0 + N_1$, and N_0 and N_1 are the number of observations in $\{D=0\}$ and $\{D=1\}$, respectively.

ASSUMPTION 4: $\sqrt{N}(\hat{\theta} - \theta) = N^{-1/2} \sum_{i \in \{D=0\} \cup \{D=1\}} \psi(Z_i, D_i) + o_P(1)$ *converges in distribution to a* $\mathcal{N}(0, V_\theta)$ *random vector, where* V_θ *is the asymptotic variance-covariance matrix of* $\hat{\theta}$.

To state the next assumption, we define

$$C_1(G) = \int_{l(P_0)}^{u(P_0)} s^2 G(s)\, ds \int_{l(P_0)}^{u(P_0)} G(s)\, ds - \left[\int_{l(P_0)}^{u(P_0)} sG(s)\, ds \right]^2,$$

$$C_2(G) = \left[\int_{l(P_0)}^{u(P_0)} s^2 G(s)\, ds \right]^2 - \int_{l(P_0)}^{u(P_0)} s^3 G(s)\, ds \int_{l(P_0)}^{u(P_0)} sG(s)\, ds,$$

$$C_3(G) = \int_{l(P_0)}^{u(P_0)} \left[\int_{l(P_0)}^{u(P_0)} s^2 G(s)\, ds - u \int_{l(P_0)}^{u(P_0)} sG(s)\, ds \right]^2 G^2(u)\, du,$$

where the upper and lower limits of integration, $u(P_0)$ and $l(P_0)$, satisfy $u(1) = 0$, $u(P_0) = \infty$ if $P_0 \in [0,1)$ and $l(0) = 0, l(P_0) = -\infty$ if $P_0 \in (0,1]$. Operationally, when the estimated P is within a bandwidth a_N of 0 or 1, C_j changes discontinuously for $j = 1,2,3$.

We impose the following conditions on G and on $C_1(G)$, $C_2(G)$, $C_3(G)$:

ASSUMPTION 5: (a) *The second derivative of* $G(s)$ *is finite*; (b) $C_1(G) \neq 0$; *and* (c) $C_1(G)$, $C_2(G)$, *and* $C_3(G)$ *are finite.*

Observe that $C_1(G)/[\int_{l(P_0)}^{u(P_0)} G(s)\, ds]^2$ corresponds to the variance of a random variable with density $G(s)/\int_{l(P_0)}^{u(P_0)} G(s)\, ds$ if $G(s) \geq 0$. Assumption 5 holds, for example, if $G(\cdot)$ is taken to be a smooth density supported in a finite interval. These restrictions on the kernel function are satisfied by the biweight kernel that we use (defined in A.2.1).

Since θ is estimated, we impose the following two regularity conditions on the behavior of the conditional expectations and the conditional densities of P_i given $D_i = 0$ or $D_i = 1$, $f_{\theta_0}(P \mid D = 0)$ and $f_{\theta_0}(P \mid D = 1)$, in the neighborhood of the true value $\theta = \theta_0$:

ASSUMPTION 6: $E(R_{it} \mid P_i, D_i = d)$ *and* $E(Y_{it} \mid P_i, D_i = d), d \in \{0,1\}$, *are twice continuously differentiable with respect to* θ *in the neighborhood of* $\theta = \theta_0$.

ASSUMPTION 7: *For* $d \in \{0,1\}$, (a) $f_{\theta_0}(P \mid D = d)$ *is bounded and continuous on* $[0,1]$, *and* (b) *for any* $\varepsilon > 0$ *there exists* $\delta > 0$ *such that if* $\|\theta - \theta_0\| < \delta$, *then*

$$\sup_{0 \leq P \leq 1} |f_\theta(P \mid D = d) - f_{\theta_0}(P \mid D = d)| < \varepsilon.$$

It is possible to weaken Assumption 6 and still obtain consistency and asymptotic normality of the estimated β and K functions, but the advantages of the local linear estimator described in Section A.1 materialize only when it is maintained.

To construct a consistent estimator of the asymptotic variances we make the following assumption.

ASSUMPTION 8: *Let β_0 be the true value of β. For $d \in \{0,1\}$, $\text{var}(Y_{it} \mid P_i, D_i = d)$ and $\text{var}(Y_{it} - R_{it}'\beta_0 \mid P_i, D_i = d)$ are continuous functions of P evaluated at $\theta = \theta_0$.*

A.4. DECOMPOSING THE CONVENTIONAL MEASURE OF SELECTION BIAS

A.4.1. *Estimation Methods*

To obtain consistent and asymptotically normal estimates of B_{1t}, B_{2t}, and B_{3t} defined below equation (14) in the text, it is necessary to estimate the overlapping support region, S_P, and $E(Y_{it} \mid P_i, D_i = 0)$. To estimate the region of overlapping support, S_P, we estimate the densities $f_{\theta_0}(\cdot \mid D = d)$ for $d \in \{0,1\}$, using a standard kernel density estimator, $\hat{f}_{\theta_0}(\cdot \mid D = d)$, applied to the estimated values of P for each group.[63]

The estimated density is evaluated at all observed data points. For both the $D = 0$ and $D = 1$ distributions, all points with zero density and the points corresponding to the lowest two percent of estimated density values are eliminated or "trimmed."[64] $\hat{S}_P$ is the subset of the points from both densities that survive trimming and share a common support. In our application, roughly 50% of the control observations ($D = 1$) and 80% of the ENPs ($D = 0$) lie in the overlap region. We estimate $E(Y_{it} \mid P_i, D_i = 0)$ by local linear regression using the estimated values of P.

The sample analogue estimators of B_1, B_2, and B_3 defined below equation (20) in the text are, for period t,

$$\hat{B}_{1t} = N_1^{-1} \sum_{i \in \{D=1\}} Y_{it}\hat{I}_i^c - N_0^{-1} \sum_{i \in \{D=0\}} Y_{it}\hat{I}_i^c,$$

$$\hat{B}_{2t} = N_1^{-1} \sum_{i \in \{D=1\}} \hat{E}(Y_{it} \mid \hat{P}_i, D_i = 0)\hat{I}_i - N_0^{-1} \sum_{i \in \{D=0\}} Y_{it}\hat{I}_i,$$

$$\hat{B}_{3t} = N_1^{-1} \sum_{i \in \{D=1\}} [Y_{it} - \hat{E}(Y_{it} \mid \hat{P}_i, D_i = 0)]\hat{I}_i,$$

where $\hat{I}_i = 1\{\hat{P}_i \in \hat{S}_P\}$, $\hat{I}_i^c = 1\{\hat{P}_i \in \hat{S}_P^c\}$, the superscript c denotes complement, N_d denotes the number of observations in the set $\{D = d\}$ for $d \in \{0,1\}$, and $N = N_0 + N_1$.[65]

A.4.2. *Asymptotic Distribution of the Estimators*

Heckman, Ichimura, and Todd (1996) establish that $\hat{B}_{1t}$, $\hat{B}_{2t}$, and $\hat{B}_{3t}$ are consistent and asymptotically normal nonparametric estimators when estimated regressors are used to estimate unknown conditional mean functions. Define $\rho_d = \lim_{N \to \infty} N_d/N$ for $d \in \{0,1\}$ and $\psi_{0t}(p) = E(Y_{it} \mid P_i = p, D_i = 0)$, and let $P_i' = \partial P(Z_i'\theta)/\partial(Z_i'\theta)$ and $\psi_{0t}'(p) = \partial\psi_{0t}(p)/\partial p$, the asymptotic variance-

[63] For all nonparametric estimates, we use the biweight kernel defined earlier.

[64] In estimating the density, we find that it is important to use a kernel that is zero outside a finite interval. With a normal kernel, or any other kernel with unbounded support, no points are estimated to have zero density. This makes it difficult to choose a trimming level that will eliminate the low density points. With a kernel supported over a finite interval, some points are estimated to have a density of zero, so that they can be eliminated along with 2% of the observations with positive estimated densities. With a kernel that has unbounded support, estimates of mean bias tend to be sensitive to the trimming level but with a kernel supported on a finite interval they are not. For further discussion, see Heckman, Ichimura, and Todd (1996).

[65] If we allow for random attrition, as we do in our empirical work, the sets $\{D = 1\}$ and $\{D = 0\}$ and the values of N_1 and N_0 are time indexed.

covariance matrix of $\sqrt{N}(\hat{\theta}-\theta_0)$ as V_θ and

$$c = E\{\psi'_{0t}(P_i)P'_i[Z_i - E(Z_i \mid P_i, D_i = 0)]' \mid D_i = 1\}.$$

The following theorem holds.

THEOREM A.1: *Suppose Assumptions* 1–8 *hold and* $\rho_d > 0$ *for* $d \in \{0,1\}$. *If the deterministic or stochastic bandwidth satisfies* $\operatorname{plim}_{N\to\infty} a_{N_0}/h_{N_0} = \alpha_0$ *for a positive constant* α_0 *and a deterministic sequence* h_N *for which*

$$\lim_{N_0\to\infty} N_0 h_N^2/\log N_0 = \infty \quad \text{and} \quad \lim_{N_0\to\infty} N_0 h_{N_0}^4 = 0,$$

then $N^{1/2}(\hat{B}_{1t} - B_{1t})$ *converges in distribution to* $\mathcal{N}(0, \sigma_{1t}^2)$ *where*

$$\sigma_{1t}^2 = \operatorname{var}(Y_{it} I_i^c \mid D = 0)/\rho_0 + \operatorname{var}(Y_{it} I_i^c \mid D = 1)/\rho_1,$$

$N^{1/2}(\hat{B}_{2t} - B_{2t})$ *converges in distribution to* $\mathcal{N}(0, \sigma_{2t}^2)$ *where*

$$\begin{aligned}\sigma_{2t}^2 = {} & \operatorname{var}\{E(Y_{it} I_i \mid P_i, D_i = 0) \mid D_i = 0\}/\rho_0 + \operatorname{var}\{E(Y_{it} I_i \mid P_i, D_i = 0) \mid D_i = 1\}/\rho_1 \\ & + E\Big\{[(f_{\theta_0}(P_i \mid D_i = 1)/f_{\theta_0}(P_i \mid D_i = 0)) - 1]^2 \\ & \quad \times [Y_{it} I_i - E(Y_{it} I_i \mid P_i, D_i = 0)]^2 \mid D_i = 0\Big\}/\rho_0 + c' V_\theta c,\end{aligned}$$

and $N^{1/2}(\hat{B}_{3t} - B_{3t})$ *converges in distribution to* $\mathcal{N}(0, \sigma_{3t}^2)$ *where*

$$\begin{aligned}\sigma_{3t}^2 = {} & E\{\operatorname{var}(Y_{it} I_i \mid P_i, D_i = 0) \mid D_i = 1\}/\rho_1 \\ & + E\Big\{[f_{\theta_0}(P_i \mid D_i = 1)/f_{\theta_0}(P_i \mid D_i = 0)]^2 \operatorname{var}(Y_{it} I_i \mid P_i, D_i = 0) \mid D_i = 0\Big\}/\rho_0 + c' V_\theta c.\end{aligned}$$

For simplicity the above expression assumes that the score vector of $\hat{\theta}$ and Y_{it} are not correlated.[66] We estimate the asymptotic variances using bootstrap methods, so we do not discuss estimation of the variances by the plug-in method. Modifications to allow for random attrition are straightforward and for the sake of brevity are deleted.

A.5. ESTIMATION OF THE MODEL WITH REGRESSORS

We next present results on the asymptotic distributions of our estimators of β, K_{0t} and K_{1t}, and $\overline{B}_{tS_P}(adj)$.

A.5.1. *Asymptotic Distribution of* $\hat{\beta}$

Let $\tilde{r}_{itd} = R_{it} - E(R_{it} \mid P_i, D_i = d)$ and $\tilde{z}_{id} = Z_i - E(Z_i \mid P_i, D_i = d)$. Throughout we assume that $a_{N_0} = a_{N_1} = a_N$ and $h_{N_0} = h_{N_1} = h_N$.

THEOREM A.2: *Under Assumptions* 1–8, *if the* (*deterministic or stochastic*) *bandwidth* a_N *satisfies* $\operatorname{plim}_{N\to\infty} a_N/h_N = \alpha_0 > 0$ *for some deterministic sequence* h_N *for which* $\lim_{N\to\infty} Nh_N^2/\log N = \infty$ *and* $\lim_{N\to\infty} Nh_N^8 = 0$, *and* H_1, *defined below*, *is nonsingular*, *and* β_0 *is the true value of* β, *then*

$$\sqrt{N}(\hat{\beta} - \beta_0) = H_1^{-1} \sum_{t\in\mathcal{T}} \sum_{d\in\{0,1\}} (N/N_d)^{1/2} N_d^{-1/2} \sum_{i\in\{D=d\}} (\tilde{r}_{itd}\varepsilon_{it} Q_{di} + H_{2d}\psi_i) + o_p(1),$$

[66] Note that when P_i has the same distribution under $D_i = 0$ and $D_i = 1$ this assumption is not necessary because $c = 0$ in that case. The derivation for the more general case is available on request from the authors.

where $H_1 = \Sigma_{t \in \mathcal{T}} \Sigma_{d \in \{0,1\}} \rho_d E(\tilde{r}_{itd} \tilde{r}'_{itd} Q_{di} \mid D_i = d)$ *and for* $d \in \{0,1\}$,

$$H_{2d} = E(K'_d(P_i) P'_i \tilde{r}_{itd} \tilde{z}'_{id} Q_{di} \mid D_i = d),$$

where Q_{di} *is as defined in Section* A.2.1, *except in this expression true rather than estimated values are used. We estimate the variance-covariance matrix of* $\hat{\beta}$ *by*

$$\hat{V}_\beta = \sum_{d \in \{0,1\}} \sum_{\tau \in \mathcal{T}} \sum_{t \in \mathcal{T}} \sum_{i \in \{D=d\}} \hat{\omega}_{i\tau d} \hat{\omega}'_{itd} (\rho_d N)^{-1}, \tag{A-3}$$

where

$$\hat{\omega}_{i\tau d} = \hat{H}_1^{-1} \left[\hat{R}_{i\tau d} \hat{\varepsilon}_{i\tau} \hat{Q}_{di} + \hat{H}_{2d} \psi(Z_i, D_i) \right],$$

$$\hat{H}_1 = \sum_{d \in \{0,1\}} \sum_{t \in \mathcal{T}} \sum_{i \in \{D=d\}} \hat{R}_{itd} \hat{R}'_{itd} \hat{Q}_{di},$$

$$\hat{H}_{2d} = N^{-1} \sum_{t \in \mathcal{T}} \sum_{i \in \{D=d\}} \hat{K}'_{dt}(\hat{P}_i) P'(Z'_i \hat{\theta}) \hat{R}_{itd} \hat{z}'_{id} \hat{Q}_{di},$$

and where $\hat{z}_{kd} = Z_k - \hat{E}(Z_k \mid \hat{P}_k, D_k = d)$, $\hat{R}_{itd} = R_{itd} - \hat{E}(R_{itd} \mid \hat{P}_i, D_i = d)$, $\hat{\varepsilon}_{itd} = (Y_{it} - R'_{it} \hat{\beta}) - \hat{E}(Y_{it} - R'_{it} \beta \mid \hat{P}_i, D_i = d)$,

$$\hat{K}'_{dt}(P) = \frac{\partial \hat{E}(Y_{it} - R'_{it} \hat{\beta} \mid \hat{P}_i, D_i = d)}{\partial \hat{P}_i},$$

and $\psi(Z_i, D_i)$ *is defined below Assumption* 3.

In an extensive Monte Carlo analysis, Heckman, et al. (1996) show that the asymptotic theory for $\hat{\beta}$ is very reliable for samples of the size used in this paper and that bootstrap and asymptotic standard errors agree.

A.5.2. *Asymptotic Distributions of* K_{0t} *and* K_{1t}

We prove the following central limit theorem for the estimator $\hat{K}_{dt}(\cdot)$ in Heckman, Ichimura, and Todd (1996). Let $K_d(P_0) = (K_{d,1}(P_0), \ldots, K_{d,T}(P_0))'$, and $\hat{K}_d(P_0) = (\hat{K}_{d,1}(P_0), \ldots, \hat{K}_{d,T}(P_0))'$ for $d \in \{0,1\}$.

THEOREM A.3: *Under Assumptions* 1–8, *if the bandwidth satisfies* $\operatorname{plim}_{N \to \infty} a_N / h_N = \alpha_0 > 0$ *for some deterministic sequence* h_N *for which* $\lim_{N \to \infty} N h_N^2 / \log N = \infty$ *and* $\lim_{N \to \infty} N h_N^5 = c$ *for some* $c > 0$, *then*

$$(N a_N)^{1/2} [\hat{K}_d(P_0) - K_d(P_0)] \Rightarrow \mathcal{N}(0, V_d) + \frac{1}{2} K''_d(P_0) \frac{C_2(G)}{C_1(G)} (N a_N)^{1/2} a_N^2 + o_p(1), \tag{A-4}$$

where the (s,t) *element of* V_d *is*

$$\frac{E(\varepsilon_{is} \varepsilon'_{it} \mid P_i = P_0, D_i = d)}{f_{\theta_0}(P_0 \mid D_i = d) \rho_d^2 C_1(G)},$$

and where C_1, C_2, *and* C_3 *are defined just before Assumption* 5.

The asymptotic bias is

$$\Psi_d = \frac{1}{2} K_d''(P_0) \frac{C_2(G)}{C_1(G)} (c\alpha_0)^{1/2}.^{67}$$

A.5.3. *Asymptotic Distribution Theory of Estimation of* $\overline{B}_{tS_P}(adj)$

We next discuss the asymptotic properties of our estimator of the regression-adjusted average bias $\overline{B}_{tS_P}(adj)$, which is defined as

$$\overline{B}_{tS_P}(adj) = \int_{S_P} [K_{1t}(P) - K_{0t}(P)]\, dF(P \mid D = 1) \Big/ \int_{S_P} dF(P \mid D = 1),$$

where S_P is the common support defined earlier. $\overline{B}_{tS_P}(adj)$ is consistently estimated by

$$\hat{\overline{B}}_{tS_P}(adj) = \sum_{i \in \{D=1\}} [\hat{K}_{1t}(\hat{P}_i) - \hat{K}_{0t}(\hat{P}_i)]\hat{I}_i \Big/ \sum_{i \in \{D=1\}} \hat{I}_i,$$

where $\hat{I}_i$ for $i \in \{D=0\} \cup \{D=1\}$ is defined in Section A.4.1. These terms ensure that the estimated control functions K_{0t} and K_{1t} are compared at common points of support and keep the denominators of $\hat{K}_{1t}$ and $\hat{K}_{0t}$ from becoming too small so that the statistical properties of this average are well defined. Denote the conditional expectation of a random variable given $P \in S_P$ by E_{S_P} and let $\varphi_{it} = [K_{1t}(P_i) - K_{0t}(P_i)]I_i - E\{[K_{1t}(P_i) - K_{0t}(P_i)] \mid I_i\}$.

THEOREM A.4: *Under Assumptions* 1–8 *if the bandwidth satisfies* $\operatorname{plim}_{N\to\infty} a_N/h_N = \alpha_0 > 0$ *for some deterministic sequence* h_N *for which* $\lim_{N\to\infty} Nh_N^2/\log N = \infty$ *and* $\lim_{N\to\infty} Nh_N^4 = 0$, *then the asymptotic distribution of* $\sqrt{N}(\hat{\overline{B}}_{tS_P}(adj) - \overline{B}_{tS_P}(adj))$ *is the same, to the first order, as*

$$\begin{aligned} N_1^{-1/2} \Bigg[& \sum_{i \in \{D=1\}} \varepsilon_{it} I_i - \sum_{i \in \{D=0\}} \varepsilon_{it} I_i [f_{\theta_0}(P_i \mid D_i = 1)/f_{\theta_0}(P_i \mid D_i = 0)] + \sum_{i \in \{D=1\}} \varphi_{it} \Bigg] \Big/ \\ & [\rho_1^{1/2} \Pr(P \in S_P \mid D = 1)] \\ & - [E_{S_P}(X_{it} \mid D_i = 1) - E_{S_P}(E_{S_P}(X_{it} \mid P_i, D_i = 0) \mid D_i = 1)]' \sqrt{N}(\hat{\beta} - \beta_0) \\ & - [E_{S_P}(K_{0t}'(P_i) P_i' Z_i \mid D = 1) \\ & \quad - E_{S_P}(E_{S_P}(K_{0t}'(P_i) P_i' Z_i \mid P_i, D_i = 0) \mid D_i = 1)]' \sqrt{N}(\hat{\theta} - \theta_0). \end{aligned}$$

Note that if the distributions of P for the ENP and control groups are the same, then the estimation of β and θ does not affect the first order asymptotic distribution since the latter two terms in this expression are zero. In this paper, we bootstrap to estimate the standard errors, so we do not present details of how to construct plug-in estimates of the variances.

A.6. JUSTIFYING THE TEST STATISTICS USED IN THIS PAPER

Testing for the absence of selection bias, $B_{1t}(P) = 0$ for all t, or the equivalent hypothesis of mean independence of U_{0it} conditional on P, $E(U_{0it} \mid P_i = P, D_i = d) = E(U_{0it} \mid P_i = P)$, and testing for index sufficiency are central tasks of this paper. All of the required test statistics are derived from the results presented in Theorem A.3. An important consequence of this theorem is that if the same kernel G and bandwidth a_N are used to estimate K_{1t} and K_{0t}, the associated bias terms (the

[67] In samples with a few thousand observations, estimation of β affects the sampling error of the estimated functions. Since $\hat{\beta}$ converges at rate $N^{1/2}$, and the bias functions converge more slowly, a conventional argument assumes that "N is big enough" to ignore the effect of estimating $\hat{\beta}$ in deriving the asymptotic distribution of the estimated K functions. This assumption turns out to be quite misleading in samples of the size at our disposal.

second term on the right-hand side of equation (A-4)) cancel when $K_{1t} = K_{0t}$, as is postulated under the null hypothesis of no selection bias. Because we use the local linear regression estimator, the bias does not depend on the distribution of the regressors on which K_{1t} and K_{0t} are estimated. These convenient properties allow us to avoid having to adjust the test statistics for noncentrality parameters. Exactly the same elimination of noncentrality parameters occurs when testing for conditional mean independence, which in this context is the same as testing for the absence of bias conditional on P. A similar simplification emerges in testing for index sufficiency. In that case, it is postulated that $K_1(P, J_l) - K_0(P, J_l)$ are the same for all discrete-valued J_l, $l = 1, \ldots, L$. Under the null hypothesis of index sufficiency, the bias term that arises from forming $\hat{K}_1(\hat{P}, J_l) - \hat{K}_0(\hat{P}, J_l)$ is the *same* for all J_l. The test for index sufficiency is based on differences $[\hat{K}_1(\hat{P}, J_l) - \hat{K}_0(\hat{P}, J_l)] - [\hat{K}_1(\hat{P}, J_{l'}) - \hat{K}_0(\hat{P}, J_{l'})]$ for $J_l \neq J_{l'}$. The bias term is the same and differences out under the null hypothesis.

Recall that $\hat{K}_{1t}$ and $\hat{K}_{0t}$ are estimated via local linear regression of the residuals $\hat{U}_{0it} = Y_{it} - R'_{it}\hat{\beta}$ on $\hat{P}_i$ for the samples for which $D = 1$ and $D = 0$, respectively. In constructing the tests, the asymptotic theory suggests that estimation of β should not affect the distribution of the test statistics, because $\hat{\beta}$ converges at rate $\sqrt{N}$ but $\hat{K}_1$ and $\hat{K}_0$ converge at rates $\sqrt{Na_N}$, which are lower. However, Heckman, Ichimura, and Todd (1996) report in a simulation study that for samples of the size used in this paper, failure to account for the effects of estimated β on the variance of the test statistics produces tests that reject at too high a rate relative to the nominal significance level and hence are conservative.

A.6.1. *Test Of No Bias Or Conditional Mean Independence*

Under the conditions of Theorem A.3, and under the null hypothesis $B_t(P) = K_{1t}(P) - K_{0t}(P) = 0$, if the same kernel and bandwidth are used to estimate K_{1t} and K_{0t}, then

$$(\hat{K}_{1t} - \hat{K}_{0t})' \left[(\hat{V}_{1t}/(N_1 a_{N_1})) + (\hat{V}_{0t}/(N_0 a_{N_0})) \right]^{-1} (\hat{K}_{1t} - \hat{K}_{0t}) \xrightarrow{d} \chi^2(1).$$

Arraying the K_{1t} and K_{0t} into a $T \times 1$ vector $\hat{K}_1 - \hat{K}_0$, under the conditions of Theorem A.3 applied to all t,

$$(\hat{K}_1 - \hat{K}_0)' \left[(\hat{V}_1/(N_1 a_{N_1})) + (\hat{V}_0/(N_0 a_{N_0})) \right]^{-1} (\hat{K}_1 - \hat{K}_0) \xrightarrow{d} \chi^2(T),$$

where $\hat{V}_{dt}$ is a consistent estimator of V_{dt}, $d \in \{0, 1\}$, and $a_{N_0} = a_{N_1}$.

We now present methods for estimating the variances V_0 and V_1. To conserve on notation, and to anticipate the expression for the variances required in the test of index sufficiency, we present expressions for the variances conditional on strata $J_l, l = 1, \ldots, L$. In testing for mean independence, there is only one stratum—the whole sample. We first present the estimator of the variance that does not adjust for higher order terms.

A.6.1.1. *Unadjusted Variance Estimator.* Define

$$\hat{V}_t(P_0) = \operatorname{diag}(\hat{V}_{1t}(P_0, J_1), \hat{V}_{0t}(P_0, J_1), \ldots, \hat{V}_{1t}(P_0, J_L), \hat{V}_{0t}(P_0, J_L)).$$

For $d \in \{0, 1\}$,

$$\hat{V}_{dt}(P_0, J_l) = \frac{C_G^{-1} \widehat{\operatorname{var}}(Y_t - R'_t \beta_0 \mid P = P_0, J = J_l, D = d)}{\hat{f}_{\theta_0}(P_0 \mid D = d)\hat{P}(J = J_l \mid P = P_0, D = d)},$$

consistently estimates V_{dt} where $C_G = C_3(G)/C_1^2(G)$, G is the same kernel density function used in the local regression estimator, and L equals the number of discrete values of J_l with $L = 1$ in the test for mean independence. Further, $\widehat{\operatorname{var}}(Y_{it} - R'_{it}\beta_0 \mid P = P_0, J = J_l, D = d)$, $\hat{f}_{\theta_0}(P_0 \mid J = J_l, D = d)$, and $\hat{P}(J = J_t \mid P = P_0, D = d)$ are consistent estimators of the conditional variance of ε_{it}, the conditional density of P_i, and the conditional probability of $J = J_l$, respectively. To test mean independence at S different values of P simply add the test statistics over all points separated by at least $2a_N$. Each test statistic is independent and thus the overall asymptotic distribution is $\chi^2(ST)$.

To estimate one diagonal component of V_d, Heckman, Ichimura, and Todd (1996) justify the estimator

$$\hat{V}_{dt}(P_0) = \sum_{i \in \{D=d\}} \tilde{\varepsilon}_{idt}^2 \hat{W}_{id}^2(P_0),$$

where $\tilde{\varepsilon}_{idt} = (Y_{it} - R_{it}'\hat{\beta} - \hat{K}_{dt}(P_i))\hat{Q}_{di}$ and the weights $\hat{W}_{id}(P_0)$ are

$$\hat{W}_{id}(P_0) = \frac{G_{i0}\Sigma_{k \in \{D=d\}} G_{k0}(P_k - P_0)^2 - G_{i0}(P_0 - P_i)[\Sigma_{k \in \{D=d\}} G_{k0}(P_k - P_0)]}{\Sigma_{j \in \{D=d\}} G_{j0}\Sigma_{k \in \{D=d\}} G_{k0}(P_k - P_0)^2 - \left[\Sigma_{k \in \{D=d\}} G_{k0}(P_k - P_0)^2\right]^2},$$

where $G_{jk} = G((P_j - P_k)/a_N)$. Although they show that this is a consistent estimator, their Monte Carlo study reveals that this estimator underestimates the true variance by about 50%. This evidence motivates our proposal to use the adjusted variance defined in A.6.1.2 below.

Over all time periods, the natural estimator of the variance-covariance matrix for the full $T \times 1$ vector K_d at $P = P_0$ is

$$\hat{V}_d(P_0) = \sum_{i \in \{D=d\}} \tilde{\varepsilon}_{id}\tilde{\varepsilon}_{id}'\hat{W}_{id}^2(P_0),$$

where $\tilde{\varepsilon}_{id} = (\tilde{\varepsilon}_{id1}, \ldots, \tilde{\varepsilon}_{idT})$. However, this estimator is not feasible when the panel is not balanced as is the case for our data.[68] Consistent estimates of each component in the variance-covariance matrix are not guaranteed to produce an estimated covariance matrix that is positive definite. Instead, we use an alternative consistent estimator that is guaranteed to be positive semi-definite:

$$\hat{V}_d(P_0) = \sum_{i \in \{D=d\}} \hat{e}_{id}\hat{e}_{id}',$$

where

$$\hat{e}_{id} = [\tilde{\varepsilon}_{id1}\eta(i,1)\hat{W}_{id}(P_0), \ldots, \tilde{\varepsilon}_{idT}\eta(i,T)\hat{W}_{id}(P_0)]',$$

and where $\eta(i,t) = 1$ if observation i has data available in period t and $\eta(i,t) = 0$ otherwise.

A.6.1.2. *Adjusting for Estimating β.* To adjust for higher-order variance terms, we apply the delta method to add two terms to $\hat{V}_{dt}$:

$$\hat{\hat{V}}_{dt} = \hat{V}_{dt} + a_{N_d}A_t'\hat{V}_\beta A_t - 2a_{N_d}\left[\sum_{i \in \{D=d\}}\sum_{t \in \mathcal{T}} \tilde{\varepsilon}_{idt}\hat{\omega}_{idt}\hat{W}_{id}\hat{Q}_{di}\right]\left[\sum_{i \in \{D=d\}} R_{it}\hat{W}_{id}\hat{Q}_{di}\right]$$

where

$$A_t = \sum_{i \in \{D=d\}} R_{it}\hat{W}_{id}\hat{Q}_{di}.$$

A.6.2. *Testing Index Sufficiency*

Testing index sufficiency is a central goal of this paper. Unlike the test proposed by Aït-Sahalia, Bickel, and Stoker (1994), we test for index sufficiency of a subfunction rather than of an entire function. We ask if the K_{dt} functions can be written solely in terms of P_i, so that we can represent equation (A-2) as $Y_{it} = R_{it}'\beta + D_i[K_{1t}(P_i) - K_{0t}(P_i)] + K_{0t}(P_i) + \varepsilon_i$. We are not interested in the question of whether the conditional mean function for Y_{it} can be expressed solely as a function of P_i, which is the question addressed by Aït-Sahalia, et al. (1994).

[68] Recall that for simplicity we have ignored the unbalanced case in presenting the asymptotic theory. Modifying it to account for random attrition is straightforward but notationally burdensome.

Using discrete regressors, the null hypothesis of index sufficiency is as follows. Letting J_l be the value of the discrete regressor in the lth group,

$$K_{1t}(P,J_l)-K_{0t}(P,J_l)=K_{1t}(P,J_{l'})-K_{0t}(P,J_{l'}) \quad \text{all for } l \text{ and } l' \quad (l,l'=1,\ldots,L)$$

and for all P and $t\in\mathcal{T}$. We test for equality of the conditional mean bias functions for different subgroups within the population.

We estimate $K_{dt}(P,J_l)$ for some fixed finite number of points $P=P_s$, $s=1,\ldots,S$, all of which are in the support of P given J_l for $l=1,\ldots,L$, and compare the estimated functions at the chosen points. We construct the test at points where the conditional densities are bounded away from zero, to guarantee that estimators of each $K_{dt}(P,J_l)$ are uniformly consistent and converge at the same rate.[69]

The statistic for testing the null hypothesis of index sufficiency at a point P_0 in period t is

$$\hat{\tau}_t(P_0)=\hat{\Delta}_t(P_0)'\hat{\Omega}_t^{-1}(P_0)\hat{\Delta}_t(P_0)\xrightarrow{d}\chi^2(L-1),$$

where

$$\hat{\Delta}_t(P_0)=M\cdot[\hat{K}_{1t}(P_0,J_1),\hat{K}_{0t}(P_0,J_1),\ldots,\hat{K}_{1t}(P_0,J_L),\hat{K}_{0t}(P_0,J_L)]',$$

and

$$M=\begin{bmatrix}1&-1&-1&1&0&0&\cdots&0&0&0&0\\0&0&1&-1&-1&1&\cdots&0&0&0&0\\\vdots&\vdots&\vdots&\vdots&\vdots&\vdots&\ddots&\vdots&\vdots&\vdots&\vdots\\0&0&0&0&0&0&\cdots&1&-1&-1&1\end{bmatrix},$$

so that $\hat{\Delta}_t(P_0)$ is a vector of $(L-1)$ contrasts and $\hat{\Omega}_t(P)=M(\widehat{VV}(P))_tM'$, where

$$\begin{aligned}(\widehat{VV}(P))_t=\operatorname{diag}(&\widehat{VV}_{1t}(P,J_1)/(N_{1t1}a_N),\widehat{VV}_{0t}(P,J_1)/(N_{0t1}a_N),\\&\ldots,\widehat{VV}_{1t}(P,J_l)/(N_{1tl}a_N),\widehat{VV}_{0t}(P,J_l)/(N_{0tl}a_N),\\&\ldots,\widehat{VV}_{1t}(P,J_L)/(N_{1tl}a_N),\widehat{VV}_{0t}(P,J_L)/(N_{0tL}a_N)),\end{aligned}$$

$$\widehat{VV}_{dt}(P,J_l)=\sum_{i\in\{D=d\}_l}\tilde{\varepsilon}_{idt}^2\hat{W}_{id}^2(P),$$

and where $\{D=d\}_l$ is the set of i values for $D=d$ associated with discrete regressor J_l, and N_{dtl} is the number of observations in the cell d,t,l.

A test of the hypothesis over T time periods at point P under the null is based on the entire vector $\hat{\Delta}(P_0)$ of length $(L-1)T$ where $\hat{\Delta}(P_0)=(\hat{\Delta}_1(P_0),\ldots,\hat{\Delta}_T(P_0))'$. The test statistic is

$$\hat{\tau}(P_0)=\hat{\Delta}(P_0)\hat{\Omega}^{-1}(P_0)\hat{\Delta}'(P_0)\xrightarrow{d}\chi^2((L-1)T),$$

where

$$\hat{\Omega}(P_0)=[I_T\otimes M][\widehat{VV}(P_0)][I_T\otimes M]'$$

and where $\widehat{VV}(P_0)=\operatorname{diag}((\widehat{VV}(P_0))_1,(\widehat{VV}(P_0))_2,\ldots,(\widehat{VV}(P_0))_T)$, I_T is the $T\times T$ identity matrix and $\otimes$ denotes a Kronecker product. For values of P that are at least two bandwidths a_N apart, the chi-square tests are independent when the kernel is supported on $[-1,1]$, and we can perform an overall test for S values as a sum of the χ^2 statistics over P. The resulting statistic is $\chi^2(S(L-1)T)$. To adjust for estimation error in β replace $\widehat{VV}_{dt}(P_0,J_l)$ by the appropriate adjustment for cell d,t,l analogous to the adjustment given in A.6.1.2.

[69] The same "trimming rule" discussed in Section A.4.1 is used to estimate the densities for the different subgroups on the J_l, $l=1,\ldots,L$.

APPENDIX B

B.1. SAMPLES USED IN THE ANALYSIS

Our data consist of four samples: the experimental control group sample, the experimental treatment group sample, the eligible nonparticipant (ENP) sample, and the Survey of Income and Program Participation (SIPP) sample. The first three of these samples were collected at four of the training centers participating in the National JTPA Study: Corpus Christi, Texas, Fort Wayne, Indiana, Jersey City, New Jersey, and Providence, Rhode Island.

The control and treatment group samples consist of persons who took part in the JTPA experiment. They applied to the JTPA program, were determined eligible for JTPA services under Title II-A of the Act, were accepted into the program, and were recommended for particular JTPA services. About one third were assigned to the control group and excluded from JTPA services for 18 months while the rest were assigned to the treatment group and given access to JTPA services.

Nonexperimental data were collected on a sample of eligible nonparticipants residing in the same geographic areas as the experimental groups. The ENP sample is composed of individuals who were, on the basis of a screening interview, determined to be (i) eligible for JTPA due to economic disadvantage; (ii) 22 to 54 years of age; (iii) not in junior high or high school; and (iv) not permanently disabled.[70] Our other nonexperimental comparison group sample is drawn from the 1988 SIPP Full Panel. We treat month 12 of the panel as a single cross-section in constructing the sample.

To match the ENP sample, we impose criteria (ii) and (iii) on the remaining samples. We are unable to impose criterion (iv) due to data limitations. Criterion (i) is imposed on the SIPP eligible subsample used in Tables XIV. The other SIPP subsamples are defined in the notes to that table. In all of the samples individuals missing data on key variables such as race or date of eligibility screening are omitted. Table B-1 summarizes the number of individuals omitted due to each criterion in the ENP and control samples.

We also impose a rectangular sample restriction based on our outcome variable, quarterly earnings. For the ENP and control samples, this restriction requires (i) at least one month of valid earnings data prior to random assignment (for the controls) or eligibility screening (for the ENPs) (hereafter the date of random assignment or eligibility determination is denoted as RA/EL); (ii) valid earnings data in the month of RA/EL; and (iii) at least one month of valid earnings data in months 13–18 after RA/EL. Table B-I indicates the number of additional observations lost due to this restriction for the ENP and control samples. Due to data limitations, only restricton (i) is applied to the treatment group sample. For the SIPP, we require valid earnings data in the first and final month of the panel for sample inclusion.

B.2. SURVEY INSTRUMENTS

The Long Baseline Survey (LBS) gathered five years of retrospective data on earnings and employment, demographic characteristics, household composition, recent training history, and transfer program participation for the ENP and control group samples. Controls completed the LBS within one or two months after random assignment. For them, the survey covers the five years prior to random assignment. The ENPs completed the LBS from 0 to 24 months after eligibility screening. For them, the survey covers the five years prior to the survey date. The response rate on the LBS was 90 percent for the controls and 78 percent for the ENPs.

Both the first and second follow-up surveys collected detailed retrospective data on job spells, hours and rates of pay, social program participation, training and job search activities, as well as background and demographic information. The surveys are basically identical except for the time periods covered. The first follow-up survey was administered to treatments, controls, and ENPs and covered the period from 12 to 24 months after random assignment for the experimental groups and from 12 to 48 months after the LBS interview date for the ENPs. The second follow-up survey was

[70] For more information on the sampling frame for the ENP sample, see Smith (1994).

TABLE B-I

NUMBERS OF OBSERVATIONS OMITTED DUE TO SAMPLE RESTRICTIONS
Experimental Control and Elig. Nonparticipant (ENP) Samples, Adult Males

Restriction	ENP Sample	Control Sample
Total number of observations	827	864
Number dropped due to missing date of eligibility screening	4	0
Number dropped due to missing value for race	7	0
Number dropped due to having no valid earnings observations	56	54
Number dropped due to rectangular sample restriction	372	302
Final analysis sample size	388	508

Note: The rectangular sample restriction requires that each observation included have at least one month of valid earnings data in the 18 months prior to random assignment or eligibility screening (RA/EL), valid earnings data in the month of RA/EL, and at least one month of valid earnings data in months 13 to 18 after RA/EL.

administered to a random sample of experimentals, including approximately one quarter of the adults. This survey covered the period from the first follow-up survey to the second follow-up survey date, which was from 24 to 48 months after random assignment.[71] The response rate to the first follow-up survey was 81 percent for the experimental groups and 79 percent for the ENPs. The second follow-up survey, administered only to the experimental groups, had a response rate of 80 percent. Experimentation with alternative methods for dealing with attrition and nonresponse sustains the findings reported in the text.

At the time of random assignment, control and treatment group members completed a Background Information Form (BIF) that collected information on demographic characteristics, social program participation, training and schooling activity, and recent labor market experience. We use the BIF data only to fill in background variables missing on the other surveys due to item nonresponse.

B.3. GENERATING THE VARIABLES USED IN OUR ANALYSIS

For the experimental and ENP samples, we use the monthly total earnings variables constructed by Abt Associates, the firm hired by the U.S. Department of Labor to produce public use data files for the JTPA experiment. These variables are based on information about average hours worked and average rates of pay on individual job spells. These variables include tips, bonuses and overtime, which are smoothed over each job spell. For the seam month between the LBS and the follow-up surveys, we calculate earnings by weighting up the information from the LBS survey.

The monthly earnings data from the LBS and follow-up surveys are combined to form a panel of up to 90 months for each individual. We organize the data by month relative to RA/EL rather than by calendar time. Since the ENPs were screened for eligibility prior to completing the LBS, we realign the data so that the month of eligibility screening for the ENPs corresponds to the month of random assignment for the controls.

The monthly earnings data from the SIPP are based on direct responses to questions about earnings on up to two jobs and from up to two businesses in each month of each four month SIPP survey reference period. Earnings on additional jobs or from additional businesses, as well as casual earnings, are collected from an additional survey question. The SIPP earnings variables also include tips, bonuses and overtime, but they are not smoothed over job spells as in the data from the JTPA experiment.[72]

[71] For control and treatment group members not responding to the first follow-up survey, the second follow-up collected information on the entire period from random assignment to the second follow-up survey interview date.

[72] For more information about the monthly earnings variables, see Smith (1995).

For our regression analyses, the monthly data were converted to quarterly data. Average monthly earnings per quarter are formed from the monthly data by taking an average over the three months that comprise each quarter. If there are missing data on earnings, the quarterly average is taken over the available months. In calculating the quarterly data, the quarters begin with the first month after RA/EL. The top one percent of the quarterly earnings values are trimmed in each quarter from the combined sample of ENPs and controls. No trimming is performed on the SIPP earnings data as they appear to be less prone to outliers.

We align the ENP data relative to the controls in the month in which we know with certainty that both the controls and ENPs are eligible for JTPA. Aligning the groups in this way requires individual realignment of the ENPs due to differences across persons in the lag between measured eligibility and administration of the baseline survey. All of the regression-adjusted estimates include variables for calendar time. However, these variables are not substantively important.

The calendar year and month of each observation in the panel are determined from variables giving the date of random assignment for control (and treatment) group members and the date of eligibility screening for the ENPs. We construct the monthly age variables using each individual's date of birth. For the control group, the date of birth is taken from the BIF while for the ENPs it is taken from the LBS.

Demographic and background variables, such as race, marital status, and education, are usually obtained from the LBS. Missing values are replaced using information from the BIF or from a follow-up survey where possible. Missing values due to item nonresponse on the variables used to estimate the possibilities of participation for the ENP and control samples are imputed. For continuous variables, values are imputed from a linear regression with the following regressors: indicators for race/ethnicity, indicators for age categories, an indicator for receipt of a high school diploma or GED, and site indicators. These variables had no missing values after imposing the initial sample restrictions. All covariates are interacted with a control group indicator. Missing values of dichotomous variables are replaced with the predicted probabilities estimated using a logit equation with the same covariates. Missing values of indicator variables with more than two categories, such as the five indicators for highest grade completed, are replaced by the predicted probabilities from a multinomial logit model where the underlying categorical variable used to construct the indicators is the dependent variable. No imputed values were generated for the SIPP sample as the rates of item nonresponse in that sample are very low. Table B-II presents descriptive statistics on the variables used to analyze the ENP and control samples. Further details on the construction of the variables and the samples appear in an expanded version of this appendix, and are available on request from the authors.

APPENDIX C

SELECTION OF VARIABLES FOR USE IN ESTIMATING THE PROBABILITY OF PARTICIPATION, P

This appendix presents the criteria used to select the Z variables in the probability of participation, $\Pr(D = 1 \mid Z)$. We have richer data than that available to previous analysts. Human capital theory suggests that younger people are more likely to benefit from training. Previous research suggests the importance of marital status, household size, and family income in affecting schooling and training decisions. Ashenfelter's (1978) analysis demonstrates the importance of recent earnings in determining participation in training programs.

To select among the variables suggested by theory, we use the two criteria discussed in Section 4.3 of the text: (a) the fraction of observations correctly predicted using the population proportion of controls as a cutoff value; and (b) statistical significance. For (a), we look at both the simple mean of the control and ENP correct prediction rates and the control correct prediction rate by itself. For (b) we "test up" by iteratively adding variables starting with the training center indicators and demographic variables. Variables which are statistically significant at conventional levels and which increase the prediction rates by a substantial amount are retained in the final specification.

Table C-I presents the control and ENP correct prediction rates, along with the simple average of the two rates, for five alternative models of P. The first three rows correspond to the three "coarse"

TABLE B-II

DESCRIPTIVE STATISTICS FOR VARIABLES USED IN THE PAPER
Experimental Control and Elig. Nonparticipant (ENP) Samples
Adult Males, 508 Controls and 388 ENPs

Variable Names (Effects)	ENPs Mean	Controls Mean	ENPs Std Error	Controls Std Error
Corpus Christi, TX	0.418	0.165	0.025	0.016
Fort Wayne, IN	0.317	0.530	0.024	0.022
Jersey City, NJ	0.121	0.156	0.017	0.016
Providence, RI	0.144	0.150	0.018	0.016
White	0.387	0.524	0.025	0.022
Black	0.119	0.272	0.016	0.020
Hispanic	0.441	0.169	0.025	0.017
Other Races	0.054	0.035	0.012	0.008
Age 25 to 29	0.173	0.220	0.019	0.018
Age 30 to 39	0.397	0.380	0.025	0.022
Age 40 to 49	0.216	0.138	0.021	0.015
Age 50 to 54	0.052	0.026	0.011	0.007
Less than 10th Grade	0.341	0.196	0.023	0.017
10th–11th Grade	0.183	0.230	0.019	0.019
12th Grade	0.270	0.361	0.022	0.021
1–3 Years College	0.131	0.168	0.017	0.016
4 + Years College	0.075	0.046	0.013	0.009
Last Married 1–12 Months Prior to RA/EL	0.020	0.040	0.007	0.008
Last Married > 12 Months Prior to RA/EL	0.038	0.131	0.009	0.014
Single, Never Married	0.255	0.508	0.021	0.022
Children Age Less than 6	0.332	0.179	0.023	0.016
Quarter 1	0.277	0.251	0.018	0.015
Quarter 2	0.227	0.207	0.018	0.013
Quarter 3	0.174	0.281	0.014	0.016
Quarter 4	0.321	0.260	0.018	0.015
Year 1986	0.000	0.000	0.000	0.000
Year 1987	0.126	0.322	0.015	0.020
Year 1988	0.715	0.544	0.020	0.021
Year 1989	0.147	0.129	0.016	0.014
Year 1990	0.012	0.006	0.005	0.003
Year 1991	0.000	0.000	0.000	0.000
Ever had Vocational Training	0.247	0.349	0.022	0.021
Currently Having Vocational Training	0.016	0.071	0.006	0.011
In School or Training in the Month of RA/EL	0.097	0.063	0.015	0.011
Last in School or Training 1–3 Months before RA/EL	0.019	0.047	0.007	0.009
Last in School or Training 4–6 Months before RA/EL	0.015	0.028	0.006	0.007
Local Unemployment Rate	7.719	6.287	0.169	0.120
Employed → Employed	0.731	0.210	0.022	0.018
Unemployed → Employed	0.067	0.106	0.012	0.013
OLF → Employed	0.019	0.047	0.007	0.009
Employed → Unemployed	0.042	0.273	0.010	0.019
Unemployed → Unemployed	0.042	0.174	0.010	0.016
OLF → Unemployed	0.014	0.060	0.006	0.010
Employed → OLF	0.012	0.057	0.005	0.010
Unemployed → OLF	0.006	0.017	0.004	0.006
OLF → OLF	0.067	0.058	0.012	0.010
One Job Spell in 18 Months Prior to RA	0.580	0.348	0.025	0.021
Two Job Spells in 18 Months Prior to RA	0.229	0.287	0.021	0.020
Three or More Job Spells in 18 Months Prior to RA	0.095	0.250	0.015	0.019
Total Number of Household Members	4.132	3.072	0.083	0.076

TABLE C-I

PERFORMANCE OF ALTERNATIVE PROBABILITY OF PROGRAM PARTICIPATION LOGIT SPECIFICATIONS COMPARING COARSE AND RICH PROBABILITY OF PROGRAM PARTICIPATION SPECIFICATIONS
(Estimated Standard Errors in Parentheses)
Experimental Control and Elig. Nonparticipant (ENP) Samples
Adult Males, 508 Controls and 388 ENPs

Specification	ENP[b] Prediction Percentage	Control[b] Prediction Percentage	Equal-Weight[c] Prediction Percentage
Coarse Scores I[a]	69.07	70.47	69.77
	(2.35)	(2.02)	(1.55)
Coarse Scores II[a]	72.42	74.21	73.32
	(2.27)	(1.94)	(1.49)
Coarse Scores III[a]	79.38	78.15	78.77
	(2.05)	(1.83)	(1.38)
Best Predictor P[a]	81.96	81.89	81.92
	(1.95)	(1.71)	(1.30)
Best Predictor P Without Earnings[a]	82.47	81.50	81.99
	(1.93)	(1.72)	(1.29)

[a]See the definitions of the variables in these models presented under Table XIII. The variables in the optimal scores are presented in Table III.
[b]The "ENP Prediction Rate" and "Control Prediction Rate" columns give the percentage of ENPs and controls correctly predicted, respectively, using the hit or miss rule.
[c]The "Equal-Weight Prediction Rate" column gives the simple mean of the ENP and control correct prediction rates.

participation probability models used in the analysis reported in Table XIII; these models are defined in the notes to that table. The demographic variables in the coarse scores I model do a surprisingly good job of predicting participation. Adding annual earnings, which has a statistically significant coefficient in the logit, improves the prediction rate for both groups. However, using the recent labor force status history variables defined in Table I, instead of annual earnings, improves the prediction rates even more—over 10 percent for the ENPs and nearly 8 percent for the controls. The coefficients on the labor force status history variables are also statistically significant when added to either Coarse I or Coarse II. It is instructive to compare the Coarse Scores III row with the fourth row, which displays the prediction rates for the best-predictor P specification used to generate the main results in this paper. Over two-thirds of the difference in prediction rates between the best-predicting models is due to the addition of the labor force transition variables. The importance of these variables in predicting participation in this program is a new finding, discussed in detail in Heckman and Smith (1995b).

The last row of Table C-I presents prediction rates for a slightly reduced specification that discards the variable measuring earnings in the month of random assignment or eligibility screening used in the best P predictor model. The coefficient on earnings in the month of random assignment or eligibility screening is highly statistically significant when it is included in any specification, and its inclusion substantially improves the control ($D = 1$) prediction rate. However, as shown in the fourth column of Table C-I, including it in the model decreases the simple average of the ENP and control prediction rates by 0.07. Since a model that includes the earnings variable dominates on two of the three prediction criteria, and since the associated coefficient on the variable is statistically significant, we include the earnings variable in our best predictor equation.

APPENDIX D

MONTE CARLO STUDY OF THE TEST FOR INDEX SUFFICIENCY

To study the size and power of our test for index sufficiency, we perform a Monte Carlo study using 100 generated samples, each with 896 observations—the size of our sample. We investigate the

size and power of our test using an additional variable besides P. We test the hypothesis that $\{K_1(P \mid r=1) - K_0(P \mid r=1)\} - \{K_1(P \mid r=0) - K_0(P \mid r=0)\} = 0$, where $r \in \{0,1\}$ indicates race group.

The data are generated from the following procedure. We first estimate earnings function (16) with $K_0(P)$ and $K_1(P)$ parameterized as quadratic functions of P so that $K_0(P \mid r=1) = K_0(P \mid r=0) = \alpha_{10}P + \alpha_{20}P^2$, and $K_1(P \mid r=1) = K_1(P \mid r=0) = \alpha_{11}P + \alpha_{21}P^2$. We estimate this function using a combined sample of blacks and whites, imposing a common β across groups. This defines the base model for the null hypothesis.

Using the realized data for R_i, and the estimated β, denoted $\hat{\beta}$, we generate residuals for each observation as follows:

$$\begin{aligned}\varepsilon_{idr} = Y_{idr} - R_i'\hat{\beta} - r_i d_i K_1(P \mid r=1) \\ - (1-r_i)d_i K_1(P \mid r=0) - r_i(1-d_i)K_0(P \mid r=1) \\ - (1-r_i)(1-d_i)K_0(P \mid r=0).\end{aligned}$$

Using these residuals we estimate the variance of ε, which is assumed to be common across race groups but allowed to vary across D. We assume that $\varepsilon_{idr} \sim \mathcal{N}(0, \sigma_d^2)$, $d \in \{0,1\}$ in generating the data for the analysis. We assume that the departures from the null operate through the linear term of the $K_1(P)$ function. Thus $K_1(P \mid r=1) = (\alpha_{11} + \gamma)P + \alpha_{21}P^2$ and $K_1(P \mid r=0) = \alpha_{11}P + \alpha_{21}P^2$. The assumption $K_0(P \mid r=1) = K_0(P \mid r=0)$ is maintained throughout.

The specified value of γ determines how far the data deviate from the assumption of index sufficiency. For larger values of γ, the model deviates more from the index sufficient model, and one would expect to see more rejections. We compute the number of rejections as a function of γ for our index sufficiency test using a 5% chi-squared critical value. The results of this analysis are displayed in Figure D-1, which plots the number of rejections against the average deviation from index sufficiency, defined here as $\gamma\bar{P}$, where $\bar{P}$ is the mean of the probabilities of participation taken over the region of common support for all (D, r) subgroups.

At $\gamma = 0$, the null hypothesis of index sufficiency is correct, and we can determine the size of our test for the sample sizes used in the paper. We obtain 25 out of 100 rejections at $\gamma = 0$ despite a

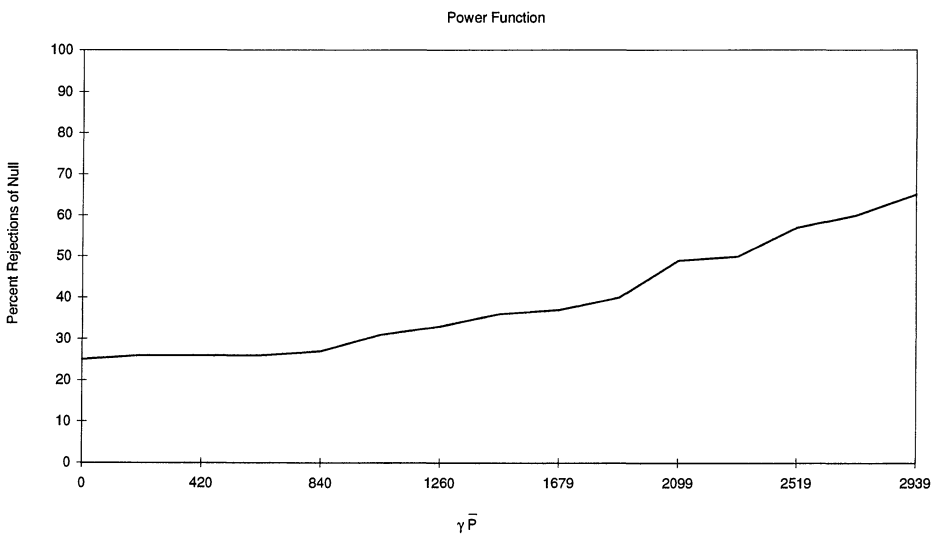


FIGURE D-1.—Power function for joint index sufficiency tests, $K_1(P \mid r=1) = K_1(P \mid r=0) + \gamma P$ and $K_0(P \mid r=1) = K_0(P \mid r=0)$.

nominal size of 5%. Thus there is a tendency to reject the null hypothesis too frequently when it is true and our test is conservative. In addition, the power function is very flat over a broad region of the data.

Figures D-2 graph the estimated bias functions for quarter 3 by the site, race, and education categories that underlie the tests on index sufficiency reported in Table VIII. These are typical of

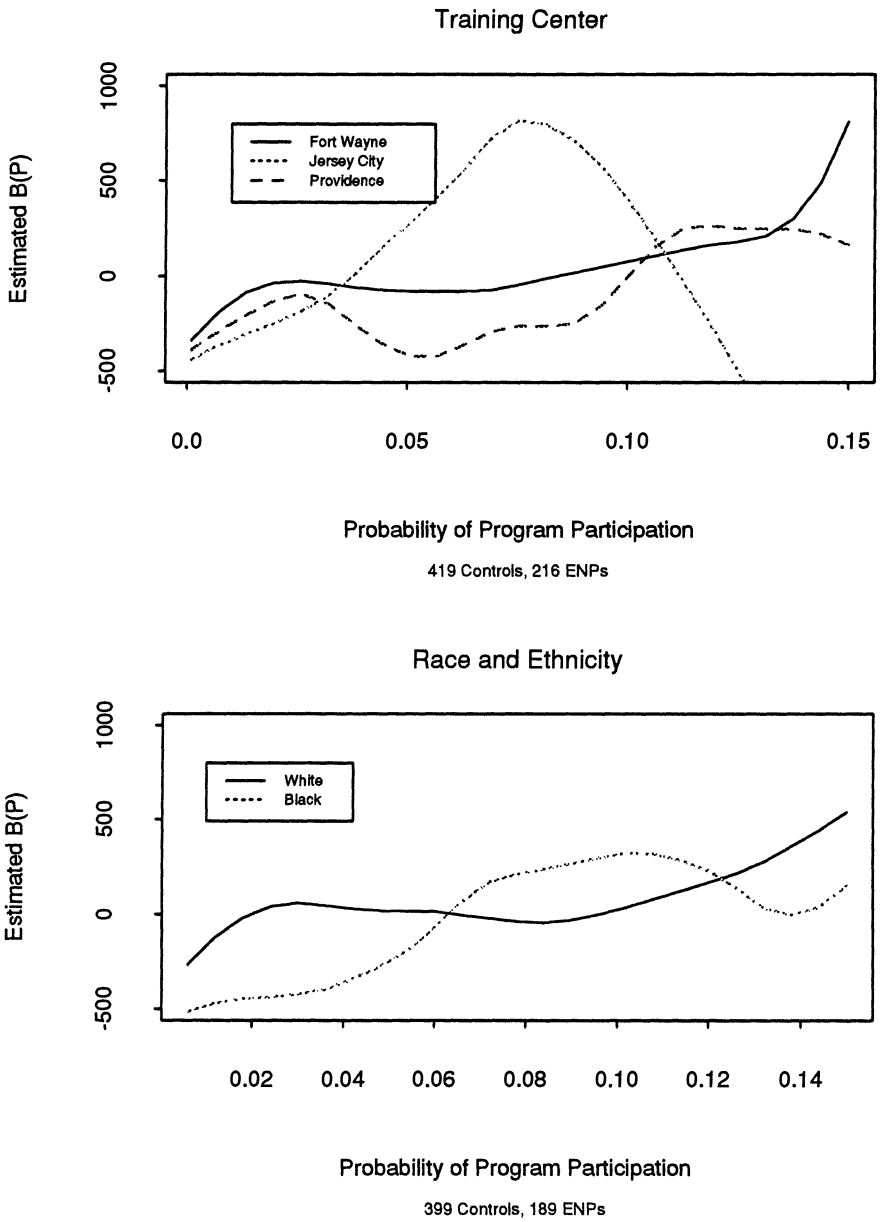


FIGURE D-2.—Estimated bias as a function of P, adult males, best predictor P model for the probability of program participation; bandwidth = 0.06, trimming = 2%.

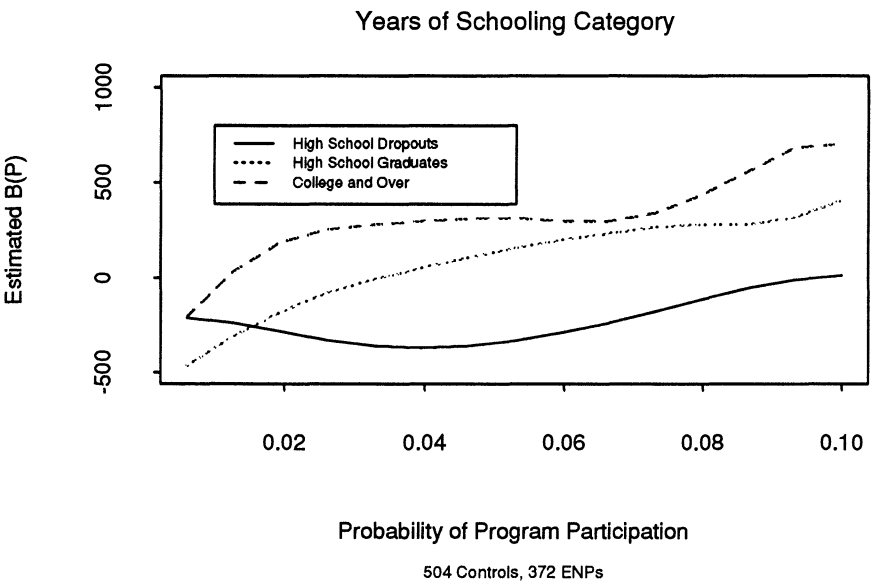


FIGURE D-2.—*Continued*

the biases found in other quarters and of the differences tested in VIII. While the agreement in the estimated bias functions is close for some groups, there are clearly differences in the bias for the other groups. Our failure to reject the null hypothesis of index sufficiency may be a consequence of the low power of our tests. The large disparity between the bias functions for certain groups does not necessarily imply that index sufficiency does not characterize the bias *within* those groups. However, the samples at our disposal are too small to make such a test meaningful. Overall, we do *not* reject the null hypothesis of index sufficiency but in light of the relatively low power of the test, our acceptance of the null hypothesis is necessarily a qualified one.

REFERENCES

AHN, H., AND J. POWELL (1993): "Semiparametric Estimation of Censored Selection Models with a Nonparametric Selection Mechanism," *Journal of Econometrics*, 58, 3–29.

AÏT-SAHALIA, Y., P. BICKEL, AND T. STOKER (1994): "Goodness-Of-Fit Tests for Regression Using Kernel Methods," Mimeo, University of Chicago.

AMEMIYA, T. (1985): *Advanced Econometrics*. Cambridge: Harvard University Press.

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

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.

BARNOW, B., G. CAIN, AND A. GOLDBERGER (1980): "Issues in the Analysis of Selectivity Bias," in *Evaluation Studies Review Annual, Volume 5*, ed. by E. Stromsdorfer and G. Farkas. San Francisco: Sage, pp. 290–317.

BLOOM, H., L. ORR, G. CAVE, S. BELL, AND F. DOOLITTLE (1993): *The National JTPA Study: Title IIA Impacts on Earnings and Employment at 18 Months*. Bethesda, MD: Abt Associates.

BREIMAN, L., J. H. FRIEDMAN, R. OLSHEN, AND C. J. STONE (1984): *Classification and Regression Trees*. Belmont, CA: Wadsworth International Group.

BURTLESS, G. (1995): "The Case for Randomized Field Trials in Economic and Policy Research," *Journal of Economic Perspectives*, 9, 63–84.

CHAMBERS, J., AND T. HASTIE (1993): *Statistical Models in S*. Pacific Grove, CA: Wadsworth and Brooks.

COCHRANE, W., AND D. RUBIN (1973): "Controlling Bias in Observational Studies," *Sankyha*, 35, 417–446.

COSSLETT, S. (1991): "Semiparametric Estimation of a Regression Model with Sample Selectivity," in *Nonparametric and Semiparametric Methods in Econometrics and Statistics*, ed. by W. Barnett, J. Powell, and G. Tauchen. Cambridge: Cambridge University Press, pp. 175–198.

DAWID, A. P. (1979): "Conditional Independence in Statistical Theory," *Journal of the Royal Statistical Society, Series B*, 41, 1–31.

DEVINE, T., AND J. HECKMAN (1996): "The Structure and Consequences of Eligibility Rules for a Social Program: A Study of the Job Training Partnership Act (JTPA)," *Research in Labor Economics, Volume 15*, ed. by S. Polachek. Greenwich, CT: JAI Press, pp. 111–170.

FAN, J. (1992): "Design-adaptive Nonparametric Regression," *Journal of the American Statistical Association*, 87, 998–1004.

——— (1993): "Local Linear Regression Smoothers and Their Minimax Efficiencies," *The Annals of Statistics*, 21, 196–216.

FAN, Y., AND Q. LI (1996): "Consistent Model Specification Tests: Omitted Variable Bias and Semiparametric Functional Forms," *Econometrica*, 64, 865–890.

FECHNER, G. T. (1860): *Elemente der Psychophysik*. Leipzig: Breitkopf and Härtel.

FISHER, R. A. (1935): *Design of Experiments*. New York: Hafner.

HAAVELMO, T. (1944): "The Probability Approach in Econometrics," *Econometrica*, 12, 1–145.

HASTIE, T., AND R. TIBSHIRANI (1990): *Generalized Additive Models*. London: Chapman and Hall.

HECKMAN, J. (1979): "Sample Selection Bias as a Specification Error," *Econometrica*, 46, 931–961.

——— (1980): "Addendum To Sample Selection Bias as a Specification Error," in *Evaluation Studies Review Annual, Volume 5*, ed. by E. Stromsdorfer and G. Farkas. San Francisco: Sage, pp. 970–995.

——— (1990a): "Varieties of Selection Bias," *American Economic Review*, 80, 313–318.

——— (1990b): "Alternative Approaches to the Evaluation of Social Programs," Barcelona Lecture, World Congress of the Econometric Society.

——— (1992): "Randomization and Social Policy Evaluation," in *Evaluating Welfare and Training Programs*, ed. by C. F. Manski and I. Garfinkel. Cambridge, MA: Harvard University Press, pp. 201–230.

——— (1996): "Randomization as an Instrumental Variable," *Review of Economics and Statistics*, 73, 336–340.

——— (1997): "Instrumental Variables: A Study of Implicit Behavioral Assumptions in One Widely-Used Estimator," *Journal of Human Resources*, 32, 442–462.

HECKMAN, J., N. HOHMANN, M. KHOO, AND J. SMITH (1997): "Substitution and Dropout Bias in Social Experiments: A Study of an Influential Social Experiment," Mimeo, University of Chicago.

HECKMAN, J., AND V. J. HOTZ (1989): "Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training," *Journal of the American Statistical Association*, 84, 862–874.

HECKMAN, J., H. ICHIMURA, AND P. TODD (1996; first draft 1994): "Making the Asymptotic Theory of Semiparametric Estimation Empirically Relevant," Mimeo, University of Chicago.

——— (1997; first draft 1993): "Matching As An Econometric Estimator: Evidence from Evaluating a Job Training Programme," *Review of Economic Studies*, 64, 605–654.

——— (1998, first draft 1993): "Matching As An Econometric Evaluation Estimator," *Review of Economic Studies*, 65, 261–294.

HECKMAN, J., M. KHOO, R. ROSELIUS, AND J. SMITH (1996): "The Empirical Importance of Randomization Bias in Social Experiments: Evidence from the National JTPA Study," Mimeo, University of Chicago.

HECKMAN, J., R. LALONDE, AND J. SMITH (1999): " The Economics and Econometrics of Active Labor Market Programs," in *Handbook of Labor Economics, Volume III*, ed. by O. Ashenfelter and D. Card. Amsterdam: North Holland.

HECKMAN, J., L. LOCHNER, AND C. TABER (1997): "Formulating and Estimating Dynamic General Equilibrium Models to Evaluate Policies Designed to Promote Skill Formation," Marschak Lecture, Hong Kong Meetings of the Econometric Society, July 25, 1997.

——— (1998): "General Equilibrium Treatment Effects: A Study of Tuition Policy," *American Economic Review*, 88, 381–386.

HECKMAN, J., AND R. ROBB, JR. (1985): "Alternative Methods for Evaluating the Impact of Interventions," in *Longitudinal Analysis of Labor Market Data*, ed. by J. Heckman and B. Singer. New York: Cambridge University Press, pp. 156–245.

——— (1986): "Alternative Methods for Solving the Problem of Selection Bias in Evaluating the Impact of Treatments on Outcomes," in *Drawing Inferences from Self-Selected Samples*, ed. by H. Wainer. New York: Springer-Verlag, pp. 63–107.

HECKMAN, J., AND P. SIEGELMAN (1993): "The Urban Institute Audit Studies: Their Methods and Findings," in *Clear and Convincing Evidence: Measurement of Discrimination in America*, ed. by M. Fix and R. Struyk. Washington, DC: Urban Institute Press, pp. 260–311.

HECKMAN, J., AND J. SMITH (1993): "Assessing the Case for Randomized Evaluation of Social Programs," in *Measuring Labour Market Measures: Evaluating the Effects of Active Labour Market Policies*, ed. by K. Jensen and P. K. Madsen. Copenhagen: Ministry of Labour, pp. 35–96.

——— (1995a): "Assessing the Case for Randomized Social Experiments," *Journal of Economic Perspectives*, 9, 85–110.

——— (1995b): "Ashenfelter's Dip and the Determinants of Participation in a Social Program: Implications for Simple Program Evaluation Strategies," mimeo, University of Chicago.

——— (1996): "Experimental and Non-Experimental Evaluations," in *International Handbook of Labour Market Policy and Evaluation*, ed. by G. Schmid, J. O'Reilly, and K. Schömann. London: Edward Elgar, pp. 37–88.

——— (1998): "Evaluating the Welfare State," in *Econometrics and Economic Theory in the 20th Century: The Ragnar Frisch Centennial*, ed. by S. Strom. Cambridge: Cambridge University Press for Econometric Society Monograph Series.

HECKMAN, J., J. SMITH, AND N. CLEMENTS (1997; first draft 1993): "Making the Most Out of Program Evaluations and Social Experiments: Accounting for Heterogeneity in Program Impacts," *Review of Economic Studies*, 64, 421–471.

HECKMAN, J., J. SMITH, AND C. TABER (1996): "What Do Bureaucrats Do? The Effects of Performance Standards and Bureaucratic Preferences on Acceptance into the JTPA Program," in *Reinventing Government and the Problem of Bureaucracy. Advances in the Study of Entrepreneurship, Innovation and Economic Growth, Volume 6*, ed. by G. Libecap. Greenwich, CT: JAI Press, pp. 110–130.

——— (1998): "Accounting for Dropouts in Evaluations of Social Programs," *Review of Economics and Statistics*, forthcoming.

HECKMAN, J., AND P. TODD (1994): "Interpreting Standard Measures of Selection Bias," Mimeo, University of Chicago, Department of Economics.

LALONDE, R. (1986): "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," *American Economic Review*, 76, 604–620.

LANCASTER, K. (1971): *Consumer Demand: A New Approach*. New York and London: Columbia University Press.

LEWIS, H. G. (1963): *Unionism and Relative Wages*. Chicago: University of Chicago Press.

MALINVAUD, E. B. (1970): *Statistical Methods of Econometrics*. Amsterdam: North-Holland.

MARSCHAK, J. (1953): "Economic Measurements for Policy and Prediction," in *Studies in Econometric Method*, ed. by W. Hood and T. J. Koopmans. New York: Wiley, pp. 1–26.

MOFFITT, R. (1992): "Evaluation Methods for Program Entry Effects," in *Evaluating Welfare and Training Programs*, ed. by C. Manski and I. Garfinkel. Cambridge: Harvard University Press, pp. 231–252.

ORR, L., H. BLOOM, S. BELL, W. LIN, G. CAVE, AND F. DOOLITTLE (1995): *The National JTPA Study: Impacts, Benefits and Costs of Title II-A*. Bethesda, MD: Abt Associates.

QUANDT, R. (1972): "A New Approach to Estimating Switching Regressions," *Journal of the American Statistical Association*, 67, 306–310.

RAO, C. R. (1965): "On Discrete Distributions Arising Out of Methods of Ascertainment," in *Classical and Contagious Discrete Distributions*, ed. by G. P. Patil. Calcutta: Statistical Publication Society, pp. 320–333.

——— (1986): "Weighted Distributions," in *A Celebration of Statistics*, ed. by S. Feinberg. Berlin: Springer-Verlag, pp. 543–569.

ROBINSON, P. (1988): "Root-*N*-Consistent Semiparametric Regression," *Econometrica*, 56, 931–954.

ROSENBAUM, P., AND D. RUBIN (1983): "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41–55.

ROSELIUS, R. (1996): "New Life for Non-Experimental Methods: Three Essays that Re-examine the Evaluation of Social Programs," Ph.D. Thesis, University of Chicago.

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

SILVERMAN, R. (1986): *Density Estimation*. London: Chapman and Hall.

SMITH, J. (1994): "Sampling Frame for the Eligible Non-Participant Sample," Mimeo, University of Chicago.

——— (1995): "A Comparison of the Earnings Patterns of Two Samples of JTPA Eligibles," Mimeo, University of Western Ontario.

TODD, P. (1995): "Semiparametric Least Squares Estimation of Binary Choice Models with Choice Based Samples Using Local Linear Regression," Mimeo, University of Chicago.

TORP, H., O. RAAUM, E. HERNAES, AND H. GOLDSTEIN (1993): "The First Norwegian Experiment," in *Measuring Labour Market Measures: Evaluating the Effects of Active Labour Market Policies*, ed. by K. Jensen and P. K. Madsen. Copenhagen: Ministry of Labour, pp. 97–140.

WAHBA, G. (1984): "Partial Spline Models for the Semiparametric Estimation of Functions of Several Variables," in *Statistical Analysis of Time Series*. Tokyo: Institute of Mathematical Statistics, pp. 37–95.

WHITE, H. (1980): "Using Least Squares to Approximate Unknown Regression Functions," *International Economic Review*, 21, 149–170.

YATCHEW, A. (1997): "An Elementary Estimator for the Partial Linear Model," Mimeo, Department of Economics, University of Toronto.