Proc. Natl. Acad. Sci. USA
Vol. 93, pp. 13416–13420, November 1996
Economic Sciences

Sources of selection bias in evaluating social programs: An interpretation of conventional measures and evidence on the effectiveness of matching as a program evaluation method

JAMES J. HECKMAN*†, HIDEHIKO ICHIMURA‡, JEFFREY SMITH§, AND PETRA TODD¶

*Department of Economics, University of Chicago, 1126 East 59th Street, Chicago, IL 60637; ‡Department of Economics, University of Pittsburgh, Forbes Quad, 230 South Bouquet Street, Pittsburgh, PA 15260; §Department of Economics, University of Western Ontario, Social Sciences Centre, London, ON Canada, N6A 5C2; and ¶Department of Economics, University of Pennsylvania, 215 Locust Walk, Philadelphia, PA 19104

Contributed by James J. Heckman, July 25, 1996

ABSTRACT **This paper decomposes the conventional measure of selection bias in observational studies into three components. The first two components are due to differences in the distributions of characteristics between participant and nonparticipant (comparison) group members: the first arises from differences in the supports, and the second from differences in densities over the region of common support. The third component arises from selection bias precisely defined. Using data from a recent social experiment, we find that the component due to selection bias, precisely defined, is smaller than the first two components. However, selection bias still represents a substantial fraction of the experimental impact estimate. The empirical performance of matching methods of program evaluation is also examined. We find that matching based on the propensity score eliminates some but not all of the measured selection bias, with the remaining bias still a substantial fraction of the estimated impact. We find that the support of the distribution of propensity scores for the comparison group is typically only a small portion of the support for the participant group. For values outside the common support, it is impossible to reliably estimate the effect of program participation using matching methods. If the impact of participation depends on the propensity score, as we find in our data, the failure of the common support condition severely limits matching compared with random assignment as an evaluation estimator.**

This paper uses data from a large-scale social experiment conducted on a prototypical job training program to decompose conventional measures of selection bias into a component corresponding to selection bias, precisely defined, and into components arising from failure of a common support condition and failure to weight the data appropriately. We demonstrate that a substantial fraction of the conventional measure of selection bias is not due to selection, precisely defined, and we conjecture that this is a general finding. We find that the conventional measure of selection bias is misleading. We also provide mixed evidence on the effectiveness of the matching methods widely used for evaluating programs. The selection bias remaining after matching is a substantial percentage—often over 100%—of the experimentally estimated impact of program participation.

Our analysis is based on the Roy (1) model of potential outcomes, which is identical to the Fisher (2) model for experiments and to the switching regression model of Quandt (3). This class of models has been popularized (and renamed) in statistics as the "Rubin" (4) model. In this model, there are two potential outcomes (Y_0, Y_1), where Y_0 corresponds to the no-treatment state and Y_1 corresponds to the treatment state. The indicator D equals 1 if a person participates in a program, and equals 0 otherwise. The probability that $D = 1$ given X, $\Pr(D = 1 \mid X)$, is sometimes called the propensity score in statistics [see Rosenbaum and Rubin (5)].

The parameter of interest considered in this paper is the mean effect of treatment on the treated. It is not always the parameter of interest in evaluating social programs [see Heckman and Robb (6), Heckman (7), Heckman and Smith (8) and Heckman *et al.* (9)], but it is commonly used. It gives the expected gain from treatment for those who receive it. For covariate vector X, it is defined as

$$\Delta(X) = E(Y_1 - Y_0 \mid X, D = 1) = E(Y_1 \mid X, D = 1) - E(Y_0 \mid X, D = 1).$$

Sometimes interest focuses on the average impact for X in some region K, e.g.,

$$\overline{\Delta}(K) = \int_K \Delta(X)\, dF(X \mid D = 1) / \int_K dF(X \mid D = 1),$$

where $F(X \mid D = 1)$ is the distribution of X conditional on $D = 1$. The term $E(Y_1 \mid X, D = 1)$ in the definition of $\Delta(X)$ can be identified and consistently estimated from data on program participants. Missing from ordinary observational studies is the data required to estimate the counterfactual term $E(Y_0 \mid X, D = 1)$.

Many methods exist for constructing this counterfactual or an averaged version of it [see Heckman and Robb (6)]. One common method uses the outcomes of nonparticipants, $E(Y_0 \mid X, D = 0)$, to proxy for the outcomes that participants would have experienced had they not participated. The selection bias $B(X)$ that results from using this proxy is defined as

$$B(X) = E(Y_0 \mid X, D = 1) - E(Y_0 \mid X, D = 0). \quad \textbf{[1]}$$

We have data from a social experiment in which some persons are randomly denied treatment. Let $R = 1$ for persons randomized into the experimental treatment group and $R = 0$ for persons randomized into the experimental control group. Randomization is conditional on $D = 1$, where $D = 1$ now indicates that the person would have participated in the absence of random assignment. Assuming no randomization bias, as defined in Heckman (7) or Heckman and Smith (8), one can use the experimental control group to consistently estimate $E(Y_0 \mid X, D = 1, R = 0) = E(Y_0 \mid X, D = 1)$ under standard conditions. In this paper, we use data on experimental controls and on a companion sample of eligible nonparticipants (persons for whom $D = 0$) to estimate $B(X)$ in order to understand the sources of bias that arise in nonexperimental evaluation studies.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviation: JTPA, Job Training Partnership Act.

†To whom reprint requests should be addressed.

Downloaded by guest on November 30, 2020

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. Heckman and colleagues (10) report that for the data analyzed in this paper

$$S_{1X} = Support\{X \mid D = 1\} \neq Support\{X \mid D = 0\} = S_{0X}.$$

Unequal supports are also found for a particular scalar measure of X, $P(X) = \Pr(D = 1 \mid X)$, which plays an important role in many evaluation methods. We find that

$$S_{1P} = Support\{P(X) \mid D = 1\} \neq Support\{P(X) \mid D = 0\} = S_{0P}.$$

Using the X distribution of participants, we define the mean selection bias $\overline{B}_{S_X}$ as

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

where $S_X = S_{1X} \cap S_{0X}$, the set of X in the common support.

Decomposing the Conventional Measure of Bias

The conventional measure of selection bias B used, e.g., in LaLonde (11), does not condition on X and is defined as $B = E(Y_0 \mid D = 1) - E(Y_0 \mid D = 0)$. It can be decomposed into a portion corresponding to a properly weighted average of $B(X)$ and two other components. First note that

$$B = \int_{S_{1X}} E(Y_0 \mid D = 1, X)dF(X \mid D = 1) - \int_{S_{0X}} E(Y_0 \mid D = 0, X)dF(X \mid D = 0). \quad \textbf{[2]}$$

Further decomposition yields

$$B = E(Y_0 \mid D = 1) - E(Y_0 \mid D = 0) = B_1 + B_2 + B_3, \quad \textbf{[3]}$$

where

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

$$B_2 = \int_{S_X} E(Y_0 \mid X, D = 0)[dF(X \mid D = 1) - dF(X \mid D = 0)], \text{ and } B_3 = P_X \overline{B}_{S_X},$$

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, $S_{1X} \backslash S_X$ is the support of X given $D = 1$ that is not in the overlap set S_X, and $S_{0X} \backslash S_X$ is the support of X given $D = 0$ that is not in the overlap set S_X.

Term B_1 in Eq. **3** does not arise from selection bias precisely defined but rather from the failure to find counterparts to $E(Y_0 \mid D = 1, X)$ in the set $S_{0X} \backslash S_X$ and the failure to find counterparts to $E(Y_0 \mid D = 0, X)$ in the set $S_{1X} \backslash S_X$. Term B_2 arises from the differential weighting of $E(Y_0 \mid D = 0, X)$ by the densities for X given $D = 1$ and $D = 0$ within the overlap set. Only the B_3 term arises from selection bias as precisely defined. The "true" bias $\overline{B}_{S_X}$ may be of a different magnitude and even a different sign than the conventional bias B.

Reducing the Dimension of the Conditioning Set and a Nonparametric Test of the Validity of Matching

For samples with only a few thousand observations, such as the one we use here, nonparametric estimation of $E(Y_0 \mid X, D = 1)$ and $E(Y_0 \mid X, D = 0)$ for high-dimensional X is impractical. Instead, we estimate conditional means as functions of $P(X)$ using the orthogonal decomposition

$$E(Y_0 \mid X, D = 1) = E(Y_0 \mid P(X), D = 1) + V$$
$$V = E(Y_0 \mid X, D = 1) - E(Y_0 \mid P(X), D = 1),$$

where $E(V \mid P(X), D = 1) = 0$. Heckman *et al.* (12) show that forming the mean conditional on $P(X)$ permits consistent, but possibly inefficient, estimation of terms analogous to those in Eq. **3** but conditioned on $P(X)$ rather than X and with the conditional means integrated against the empirical distributions for $P(X)$, $F(P(X) \mid D = 1)$ and $F(P(X) \mid D = 0)$.

Another advantage of conditioning on $P(X)$ in constructing the conditional means is that we can test the validity of matching as a method of evaluating programs. If

$$Y_0 \perp\!\!\!\perp D \mid X, \quad \textbf{[4]}$$

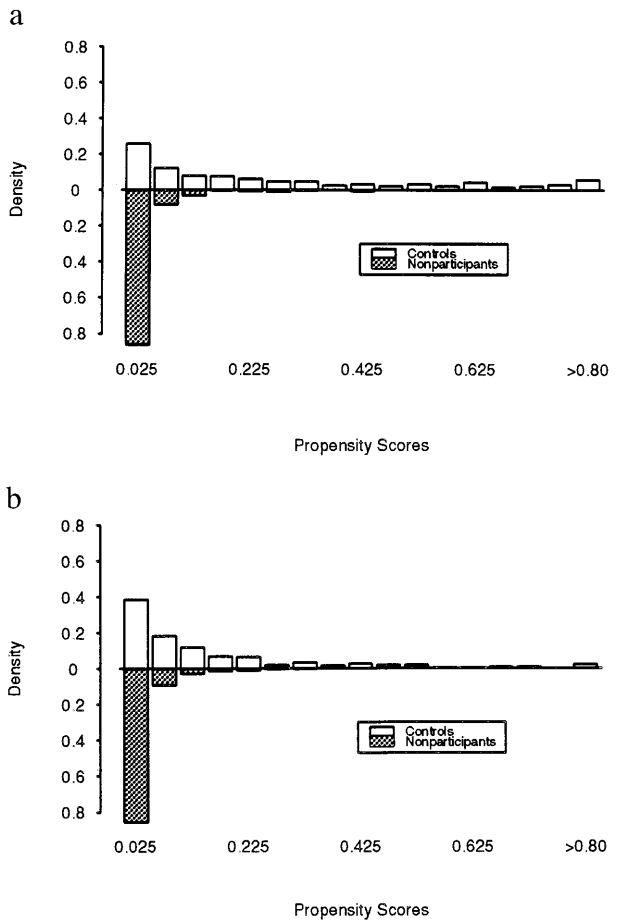

FIG. 1. (*a*) Density of estimated propensity scores for adult male controls and eligible nonparticipants. (*b*) Density of estimated propensity scores for adult female controls and eligible nonparticipants.

Downloaded by guest on November 30, 2020

meaning that Y_0 is independent of D given X, then

$$Y_0 \perp\!\!\!\perp D \mid P(X),$$

for $P(X) \in H \subseteq (0, 1)$, where H is some set in the unit interval [see Rosenbaum and Rubin (5)]. Two implications of Eq. **4** are that

$$E(Y_0 \mid P(X), D = 1) = E(Y_0 \mid P(X)), \quad P(X) \in H, \quad \textbf{[5a]}$$

and

$$E(Y_0 \mid P(X), D = 0) = E(Y_0 \mid P(X)), \quad P(X) \in H, \quad \textbf{[5b]}$$

so that $B(P(X)) = E(Y_0 \mid D = 1, P(X)) - E(Y_0 \mid D = 0, P(X)) = 0$ for all $P(X) \in H$ and hence $\bar{B}_{S_P} = 0$. A test that $B(P(X)) = 0$ for all $P(X) \in H$ is a test of the validity of the matching method as an estimator of treatment effects in the region H.

Provided that condition **5a** is met, matching is a very attractive method for estimating Δ conditional on $P(X)$. Under the condition given by Eq. **4**, or the weaker condition **5a**, the difficulty of finding matches for high-dimensional X is avoided by conditioning only on $P(X)$. Furthermore, matching methods using observations with common support eliminate two of the three sources of bias in Eq. **3**. The bias arising from regions of nonoverlapping support, term B_1 in Eq. **3**, is eliminated by matching only over regions of common support. The bias due to different density weighting is eliminated because matching on participant propensity scores effectively reweights the nonparticipant data. Thus $P_X \bar{B}_{S_P}$ is the only component in Eq. **3** that is not necessarily eliminated by matching.

Nonparametric estimates of each of the components in Eq. **3** are obtained from Eq. **6**, below, where n_1 denotes the size of the $D = 1$ sample and n_0 denotes the size of the $D = 0$ sample. Let $\hat{}$ indicate an estimate and let $\{D = 1\}$ be the set of indices i for persons with $D = 1$, $\{D = 0\}$ be the set of indices i for $D = 0$, and $P_i = P(X)$ for person i. Then we may decompose $\hat{B}$ into the sample analogs of the three terms in Eq. **3**,

$$\hat{B} = \hat{E}(Y_0 \mid D = 1) - \hat{E}(Y_0 \mid D = 0) = \hat{B}_1 + \hat{B}_2 + \hat{B}_3 \quad \textbf{[6]}$$

where

$$\hat{B}_1 = \frac{1}{n_1} \sum_{\substack{i \in \{D=1\} \\ P_i \in S_{1P} \setminus S_P}} Y_0(P_i) - \frac{1}{n_0} \sum_{\substack{i \in \{D=0\} \\ P_i \in S_{0P} \setminus S_P}} Y_0(P_i)$$

$$\hat{B}_2 = \frac{1}{n_1} \sum_{\substack{i \in \{D=1\} \\ P_i \in S_P}} \hat{E}(Y_0 \mid D = 0, P_i) - \frac{1}{n_0} \sum_{\substack{i \in \{D=0\} \\ P_i \in S_P}} Y_0(P_i)$$

Table 1. Decomposition of mean earnings difference between experimental controls and comparison sample of eligible nonparticipants

Quarter	(2) Mean earnings difference ($\hat{B}$)	(3) Nonoverlapping support* ($\hat{B}_1$) [%]	(4) Different density weighting ($\hat{B}_2$) [%]	(5) Selection bias ($\hat{B}_3$) [%]	(6) Average bias ($\hat{\bar{B}}_{S_P}$)	(7) Selection bias ($\hat{\bar{B}}_{S_P}$) as a % of treatment impact†
Adult men, experimental controls, and comparison sample of eligible nonparticipants‡						
t = 1	−418	240	−676	18	36	713
	(38)	(29) [−57]	(35) [162]	(26) [−4]		
t = 2	−349	294	−658	15	30	83
	(47)	(37) [−84]	(43) [188]	(31) [−4]		
t = 3	−337	305	−649	7	13	23
	(55)	(38) [−90]	(44) [192]	(30) [−2]		
t = 4	−286	323	−644	35	69	117
	(57)	(37) [−113]	(47) [225]	(32) [−12]		
t = 5	−305	320	−671	45	89	201
	(57)	(39) [−105]	(52) [220]	(38) [−15]		
t = 6	−328	303	−655	24	47	78
	(63)	(44) [−93]	(50) [200]	(42) [−7]		
Postprogram average	−337	298	−659	24	48	109
	(47)	(35) [−88]	(42) [195]	(28) [−7]		
Adult women, experimental controls, and comparison sample of eligible nonparticipants§						
t = 1	−26	83	−144	35	46	302
	(24)	(11) [−316]	(18) [548]	(24) [−132]		
t = 2	29	100	−120	49	64	261
	(25)	(13) [344]	(20) [−411]	(28) [167]		
t = 3	38	105	−120	54	70	151
	(26)	(14) [272]	(22) [−312]	(30) [139]		
t = 4	55	108	−107	54	70	206
	(30)	(16) [195]	(23) [−193]	(29) [97]		
t = 5	62	117	−102	47	62	212
	(34)	(18) [188]	(25) [−164]	(33) [76]		
t = 6	40	122	−114	33	44	158
	(36)	(18) [301]	(24) [−283]	(29) [82]		
Postprogram average	33	106	−118	45	59	202
	(26)	(13) [318]	(20) [−355]	(26) [136]		

Bootstrapped standard errors are shown in parentheses; percentages of mean difference attributable to components are shown in square brackets. Quarterly earnings expressed in monthly dollars.

* Two percent trimming rule used to determine overlapping support region (S_P) following [12]. For adult males, proportion of controls in $S_P = 0.51$. Proportion of eligible nonparticipants in $S_P = 0.97$. For adult females, proportion of controls is 0.76 and proportion of nonparticipants is 0.96.

† Ratio of absolute value of $\hat{\bar{B}}_{S_P}$ to absolute value of experimentally determined impact.

‡ Adult male sample contains 508 controls and 388 eligible nonparticipants.

§ Adult female sample contains 696 controls and 866 eligible nonparticipants.

Downloaded by guest on November 30, 2020

Table 2. Selection bias estimates at P deciles

Quarter	Propensity score decile									
	1	2	3	4	5	6	7	8	9	10
Adult men, experimental controls, and comparison sample of eligible nonparticipants										
t = 1	−276	−78	44	−43	−92	−5	120	135	137	283
	(145)	(111)	(115)	(106)	(131)	(112)	(126)	(117)	(160)	(206)
t = 2	−177	−19	72	−107	−117	1	148	180	267	240
	(140)	(108)	(132)	(134)	(136)	(98)	(131)	(124)	(160)	(319)
t = 3	−183	−105	118	−37	−76	29	161	269	296	−200
	(143)	(110)	(144)	(132)	(141)	(118)	(125)	(144)	(157)	(400)
t = 4	−171	107	251	−13	−77	−13	179	186	188	51
	(150)	(126)	(154)	(143)	(144)	(102)	(132)	(136)	(144)	(312)
t = 5	−229	205	303	−78	−76	70	215	225	202	250
	(176)	(118)	(136)	(141)	(142)	(127)	(150)	(150)	(147)	(264)
t = 6	−306	−44	47	−133	−70	73	129	192	263	247
	(131)	(134)	(156)	(132)	(134)	(128)	(141)	(136)	(156)	(243)
Postprogram average	−224	11	139	−69	−85	26	159	198	225	145
	(61)	(48)	(57)	(54)	(56)	(47)	(55)	(55)	(63)	(121)
Adult women, experimental controls, and comparison sample of eligible nonparticipants										
t = 1	119	8	−9	18	54	84	82	−85	16	302
	(80)	(54)	(66)	(48)	(70)	(53)	(80)	(75)	(67)	(71)
t = 2	170	65	95	37	113	55	−34	−21	51	192
	(92)	(56)	(94)	(53)	(71)	(74)	(86)	(87)	(76)	(104)
t = 3	170	89	158	136	109	13	−37	−35	46	96
	(92)	(65)	(78)	(71)	(80)	(72)	(83)	(90)	(81)	(111)
t = 4	124	91	97	83	82	50	38	−88	30	126
	(93)	(56)	(64)	(58)	(83)	(61)	(88)	(99)	(78)	(119)
t = 5	141	129	89	88	70	38	−42	−121	−9	192
	(92)	(60)	(70)	(67)	(90)	(66)	(79)	(101)	(80)	(98)
t = 6	115	111	32	36	−29	52	−2	−96	3	185
	(90)	(69)	(81)	(59)	(92)	(74)	(90)	(94)	(83)	(103)
Postprogram average	140	82	77	66	67	49	1	−74	23	182
	(37)	(25)	(31)	(24)	(33)	(27)	(34)	(37)	(32)	(42)

Deciles of the distribution of P for $D = 1$ group of experimental controls. Asymptotic standard errors in parentheses; quarterly earnings stated in monthly dollars.

$$\hat{B}_3 = \frac{1}{n_1} \sum_{\substack{i \in \{D=1\} \\ P_i \in S_P}} [Y_0(P_i) - \hat{E}(Y_0 \mid D = 0, P_i)],$$

and where the imputed outcome in the no-treatment state for an observation with propensity score P_i, $\hat{E}(Y_0 \mid D = 0, P_i)$, is estimated by a local linear regression of Y_0 on P_i using data on persons for whom $D = 0$. We use the local linear regression methods of Fan (13) with optimal data-dependent bandwidths. Each term under the summations on the right-hand side of Eq. **6** is self-weighted by averaging over the empirical distribution of propensity scores in either the $D = 1$ or $D = 0$ sample. Heckman *et al.* (12) show that under random sampling each term is consistently estimated and $\sqrt{N}$ times each term centered around its probability limit is asymptotically normal. That work extends the analysis in Rosenbaum and Rubin (5) by presenting a rigorous asymptotic distribution theory for the matching estimator.

Failure of a Common Support Condition: A Major Component of Measured Selection Bias

A major finding reported in our research [see Heckman *et al.* (10, 12)] is that using a variety of conditioning variables, the support condition

$$Support\{P(X) \mid D = 1\} = Support\{P(X) \mid D = 0\}$$

is not satisfied over large intervals of $0 \leq P(X) \leq 1$ in our sample. Fig. 1 *a* and *b* present histograms showing on the same graph the distributions of the estimates of $P(X)$ for the control and comparison groups for adult men and women, respectively. The propensity scores were estimated using the covariates X reported in Heckman *et al.* (10). These covariates are chosen to minimize 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 is some cutoff value of $P(X)$. Recent (last 6 month) unemployment and earnings histories turn out to be the key predictors of participation for both groups. We find that the set of X that is chosen is robust to wide variations in P_c around the (known) population mean of P_i, $E(P(X))$. Our estimation method corrects for the overrepresentation of the experimental control group ($D = 1$) relative to the eligible nonparticipants ($D = 0$) in the available data using ideas developed in the analysis of weighted distributions by Rao (14, 15). A universal finding in our research using a variety of covariates is the failure of the common support condition. For both male and female comparison groups, there are substantial stretches of the control group values of P for which there are no comparison group members. This is an essential and hitherto unnoticed source of selection bias as conventionally measured.

Estimating the Components of the Conventional Measure of Selection Bias

Table 1 presents consistent and asymptotically normal estimates of the three components of the decomposition in Eq. **3** estimated using the formula in Eq. **6**. The data are from the National Job Training Partnership Act (JTPA) Study (NJS), a recent experimental evaluation of the training programs funded under the JTPA [see Orr *et al.* (16)]. The JTPA program is the largest federal training program in the United States and is similar both to earlier federal training programs in the United States and to many other programs throughout

Downloaded by guest on November 30, 2020

the world. Lessons from our study are likely to apply to other training programs.

In the JTPA evaluation, accepted applicants were randomly assigned into treatment and control groups, with the control group prohibited from receiving JTPA services for 18 months. A sample of persons eligible for JTPA in the same localities as the experiment who chose not to participate in the program was collected as a nonexperimental comparison group. The same survey instrument was administered to the control and comparison groups.

In the notation defined earlier, 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 $D = 0$. Following the experimental analysis, we use quarterly earnings and total earnings in the 18 months after random assignment as our outcome measures.

Table 1 reports estimates of the components of the decomposition in Eq. **3** with earnings as the outcome variable for the adult men and women in our data. The first column in each table indicates the quarter (3-month period) over which the estimates are constructed. These quarters are defined relative to the month of random assignment. Each row corresponds to one quarter, with the bottom row reporting totals over the first six quarters (18 months) after random assignment. The second column reports the estimated mean selection bias $\hat{B}$. The next three columns report estimates of the components of the decomposition in Eq. **3**. 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 the third column of each table. The component arising from misweighting of the data, $\hat{B}_2$, is given in the fourth column and the component due to true selection bias, $\hat{B}_3$, appears in the fifth column. The sixth column presents $\hat{\overline{B}}_{S_P}$, the estimated selection bias for those in the overlap set S_P. The final column expresses $\hat{\overline{B}}_{S_P}$ as a fraction of the experimental impact estimate. All of the values in Table 1 are reported as monthly dollars. Thus, the value of −418 in the first row and first column of Table 1 indicates a mean earnings difference of −$418 per month over the 3 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 notes to each table.

A remarkable feature of the tables is that for the overall 18 month earnings measure, terms $\hat{B}_1$ and $\hat{B}_2$ are generally substantially larger than the selection bias term $\hat{B}_3$ for both groups. For adult males, the selection bias is a tiny fraction (only two percent) of the conventional measure of selection bias and is not statistically significantly different from zero. This is surprising since a majority of both the control and comparison group samples are in the overlap set, S_P, for both groups. For adult women, selection bias is proportionately higher although the conventional measure $\hat{B}$ is lower than for adult males. For them the bias measures $\hat{B}$ and $\hat{B}_3$ are of the same order of magnitude. Results for male and female youth reported in Heckman *et al.* (12) are similar to those for adult women. These overall results appear to provide a strong endorsement for matching on the propensity score as a method of program evaluation, especially for males. However, the bias $\hat{\overline{B}}_{S_P}$ that is not eliminated by matching on a common support is still large relative to the treatment effects, as is shown in the seventh column of Table 1.

The decompositions for quarterly earnings tell a somewhat different story. There is considerable evidence of selection bias for adult males in quarter t = 5, although even in this quarter the selection bias is still dwarfed by the other components of Eq. **3**. However, expressed as a fraction of the experimental impact estimate, the bias is substantial in most quarters.

The evidence for the empirical importance of selection bias that is not removed by the matching estimator used in this paper is even stronger when we examine the bias at particular deciles of the P_i distribution. This is done in Table 2. For adult males, the bias tends to be large, negative and statistically significant at the lowest decile, with a large positive bias in the upper deciles. For adult women, the pattern is U-shaped with the smallest bias at the lowest deciles. The apparent success of the matching method in eliminating selection bias in the overall estimates is a fortuitous circumstance that masks substantial bias within quarters and over particular subintervals of P_i. These patterns are found for many different specifications of P (see ref. 10).

The Failure of Matching to Estimate the Full Treatment Effect

Fig. 1 demonstrates that the support of P_i in the overlap set, S_P, is substantially different from the support of P_i for participants in the program, S_{1P}. This evidence implies that even if matching eliminates selection bias for P_i in the common support, the matching estimator cannot estimate the impact of participation over the entire set S_{1P}. In Heckman *et al.* (10), we report that the treatment effect varies with P_i; thus, failure of the common support condition $S_{0P} = S_{1P}$ means that the matching estimator cannot identify the full treatment effect. At best, the matching estimator provides a partial description of the impact of participation on outcomes.

We thank Derek Neal and José Scheinkman for critical readings of this manuscript. We thank the Bradley Foundation, the Russell Sage Foundation, and the National Science Foundation (SBR-93-21-048) for research support.

1. Roy, A. D. (1951) *Oxford Economic Papers* **3,** 135–146.
2. Fisher, R. A. (1935) *Design of Experiments* (Hafner, New York).
3. Quandt, R. (1972) *J. Am. Stat. Assoc.* **67,** 306–310.
4. Rubin, D. (1978) *Ann. Stat.* **7,** 34–58.
5. Rosenbaum, P. & Rubin, D. B. (1983) *Biometrika* **70,** 41–55.
6. Heckman, J. & Robb, R. (1985) in *Longitudinal Analysis of Labor Market Data,* eds. Heckman, J. & Singer, B, (Cambridge Univ. Press, Cambridge, U.K.), pp. 156–245.
7. Heckman, J. (1992) in *Evaluating Welfare and Training Programs*, eds. Manski, C. & Garfinkel, I. (Harvard Univ. Press, Cambridge, MA), pp. 62–95.
8. Heckman, J. & Smith, J. (1995) *J. Econ. Perspect.* **9,** 85–110.
9. Heckman, J., Smith, J. & Taber, C. (1996) *Rev. Econ. Stat.*, in press.
10. Heckman, J., Ichimura, H., Smith, J. & Todd, P. (1996) *Econometrica*, in press.
11. LaLonde, R. (1986) *Am. Econ. Rev.* **76,** 604–620.
12. Heckman, J., Ichimura, H. & Todd, P. (1996) *Rev. Econ. Studies*, in press.
13. Fan, J. (1992) *J. Am. Stat. Assoc.* **87,** 998-1004.
14. Rao, C. R. (1965) in *Classical and Contagious Discrete Distributions*, ed. Patil, G. P. (Stat. Publ. Soc., Calcutta), pp. 320–333.
15. Rao, C. R. (1986) in *A Celebration of Statistics*, ed. Feinberg, S. (Springer, Berlin), pp. 543–569.
16. Orr, L., Bloom, H., Bell, S., Lin, W., Cave, G. & Doolittle, F. (1995) *The National JTPA Study: Impacts, Benefits and Costs of Title II-A* (Abt Assoc., Bethesda).

Downloaded by guest on November 30, 2020