Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training

JAMES J. HECKMAN and V. JOSEPH HOTZ*

The recent literature on evaluating manpower training programs demonstrates that alternative nonexperimental estimators of the same program produce an array of estimates of program impact. These findings have led to the call for experiments to be used to perform credible program evaluations. Missing in all of the recent pessimistic analyses of nonexperimental methods is any systematic discussion of how to choose among competing estimators. This article explores the value of simple specification tests in selecting an appropriate nonexperimental estimator. A reanalysis of the National Supported Work Demonstration data previously analyzed by proponents of social experiments reveals that a simple testing procedure eliminates the range of nonexperimental estimators at variance with the experimental estimates of program impact.

KEY WORDS: Evaluation; Model-selection tests; Selection bias.

1. INTRODUCTION

Evaluation is now an accepted component of most government programs. Federal-government support for evaluating social programs has ranged between 500 million and a billion dollars per year for the past decade. Rossi and Freeman (1985) describe a sampling of such studies. Manpower training programs, designed to upgrade the employment and earnings of the poor, have been a frequent subject of evaluation.

Virtually all of these manpower training evaluations are based on the following principle. Earnings and other outcome measures of trainees are compared with earnings and other outcome measures of nontrainees. In an experimental evaluation, the *comparison group* consists of individuals who applied and were accepted into the program but were randomized out before the program began. A comparison group selected in this fashion is sometimes called a *control group*. In a nonexperimental evaluation, the comparison group consists of individuals judged to be "comparable" to trainees, except for not having received training. A variety of matching and statistical-adjustment procedures has been proposed to account for discrepancies in observed and unobserved characteristics between trainees and candidate comparison-group members that might distort earnings comparisons. Failure to control for such characteristics in comparing outcomes of trainees with comparison-group members may lead to substantial bias in the estimate of program impacts. Such bias is called *selection bias*.

Nonexperimental estimators differ in the assumptions they invoke to justify various statistical adjustments. Precisely because data from a properly conducted experiment do not require such adjustments, experimental estimates of program impact are inherently less controversial.

Several recent influential articles argue that alternative nonexperimental estimators of manpower training program impact produce a wide range of estimates for the same program. LaLonde (1986) and Fraker and Maynard (1984, 1987) use experimental data from the National Supported Work Demonstration project (NSW) combined with nonexperimental data to compare experimental and nonexperimental estimates of the program. These authors find that nonexperimental estimates vary widely and differ greatly from the experimental estimates. Their findings have led respected scholars to conclude that

> estimates of program effects that are based on nonexperimental comparisons can be subject to substantial misspecification uncertainty (Burtless and Orr 1986, p. 613),

and

> randomized clinical trials are necessary to determine program effects (Ashenfelter and Card 1985, p. 648).

Barnow (1987) argues that

> experiments appear to be the only method available at this time to overcome the limitations of nonexperimental evaluations (p. 190).

The LaLonde and Fraker and Maynard studies stimulated the U.S. Department of Labor to fund a $20 million evaluation of the Job Training Partnership Act using an experimental approach [see the recommendations of the Job Training Longitudinal Survey Research Advisory Panel in the report of Stromsdorfer, Boruch, Bloom, Gueron, and Stafford (1985)].

As noted by Rivlin (1971), randomized assignment of applicants is politically unpopular. There is considerable difficulty in gaining acceptance for such randomization by training program managers and local political officials. Moreover, as noted by Burtless and Orr (1986), there are

* James J. Heckman is A. Whitney Griswold Professor of Economics, Department of Economics, Yale University, New Haven, CT 06520. V. Joseph Hotz is Associate Professor, Graduate School of Public Policy Studies, University of Chicago, Chicago, IL 60637. This research was supported by National Science Foundation Grant SES-87-39151 and a contract from the U.S. Department of Labor. Opinions expressed in this article are those of the authors and not necessarily those of the U.S. Department of Labor or the National Science Foundation. The authors thank Burt Barnow, Ricardo Barros, Kyungsoo Choi, Robert LaLonde, Ernst Stromsdorfer, and the referees for helpful comments; Steve Cameron for able research assistance; and Rebecca Maynard of MPR, Inc., for providing her data.

© 1989 American Statistical Association
Journal of the American Statistical Association
December 1989, Vol. 84, No. 408, Applications & Case Studies

practical difficulties in conducting experiments in social contexts. Individuals assigned to treatments often do not show up. Individuals randomized out of a program at one trial may subsequently cross over and become program participants. Experimental evaluation of multistage programs requires costly multistage randomization (Heckman, Hotz, and Dabos 1987, p. 423). These difficulties compromise the ability of experiments to provide unbiased estimates of program impact without resorting to the nonexperimental statistical-adjustment methods that experiments were designed to avoid. Such costs make it likely that most evaluations will continue to be conducted on nonexperimental data. In light of the conclusions of the recent literature, however, it would seem that such evaluations are unlikely to be credible.

Two unstated premises underlie the recent negative assessments of nonexperimental evaluation procedures. First, alternative nonexperimental estimation procedures should produce approximately the same program estimate. Second, there is no objective way to choose among alternative nonexperimental estimators.

The first premise is invalid if there are systematic differences between trainees and comparison-group members in observed and unobserved characteristics affecting outcome measures. Different nonexperimental estimators make different assumptions about the distribution of these differences. Only in the absence of systematic differences in characteristics between trainees and comparison-group members would alternative nonexperimental estimators produce the same estimate of program impact up to sampling variation. Evidence of striking differences in estimates produced from alternative nonexperimental estimators merely confirms the existence of systematic differences between trainees and comparison-group members in characteristics affecting outcome measures. It sheds no light on the credibility of any particular estimator or class of estimators.

The truth of the second premise hinges on both the content of the available nonexperimental data and the assumptions invoked to justify various nonexperimental estimators. We present a test for model specification that can be applied to any nonexperimental model, provided that there is access to preprogram outcome measures and regressor variables for trainees and comparison-group members. Such data are widely available. We also present tests based on overidentifying aspects of certain evaluation models that impose restrictions beyond those required to estimate program impacts. We then present specification tests that exploit less commonly available experimental data.

This article demonstrates the value of such tests for choosing among alternative nonexperimental estimators. Previous empirical studies documenting the sensitivity of program-impact estimates to alternative nonexperimental estimation procedures either do not test the fitted models against the available data or disregard the inference from such tests.

We reanalyze the NSW data used by LaLonde (1984) and Fraker and Maynard (1984, 1987) in their influential studies. We demonstrate the value of our proposed tests in (a) rejecting nonexperimental estimators that produce estimates that conflict with the experimental evidence, and (b) not rejecting estimators that produce estimates that accord with the experimental evidence. Our evidence tempers the recent pessimism about nonexperimental evaluation procedures that has become common in the evaluation community.

Section 2 defines the problem of selection bias in conducting nonexperimental evaluations of program impact. Section 3 discusses the nonexperimental estimators used in our reanalysis of the NSW data. We are limited because we only have access to grouped rather than individual-level data, so a variety of available nonlinear nonexperimental estimators could not be employed. Nonetheless, the estimators we consider are representative of methods that have previously been used to evaluate training programs. Section 4 outlines procedures for testing the appropriateness of alternative nonexperimental evaluation methods. Section 5 reports the empirical performance of the proposed testing strategy. Section 6 presents concluding remarks.

2. THE PROBLEM OF SELECTION BIAS

Selection bias arises in evaluating the impact of training on mean earnings, when the mean earnings of trainees differ from the mean earnings of comparison-group members even in the absence of training. Letting Y_{it}^* be the earnings of individual i in period t in the absence of training, and letting $d_i = 1$ if a person receives training and 0 otherwise, selection bias is present if

$$E(Y_{it}^* \mid d_i = 1) \neq E(Y_{it}^* \mid d_i = 0),$$

where we assume that comparison-group members would not be trainees ($d_i = 0$).

Let Y_{it} be the observed value of earnings for individual i at time t and α_{it} be the impact of training on person i at time t. We adopt the convention that training occurs in period k. Then,

$$\begin{aligned} Y_{it} &= Y_{it}^* + d_i\alpha_{it}, & t > k \\ &= Y_{it}^*, & t < k. \end{aligned} \tag{2.1}$$

We focus on estimating the mean impact of training on the trained; that is,

$$E(\alpha_{it} \mid d_i = 1) = E(Y_{it} - Y_{it}^* \mid d_i = 1).$$

The special case of identical training impact for all persons assumes that $\alpha_t = \alpha_{it}$ for all i, so

$$\alpha_t = E(\alpha_{it} \mid d_i = 1) = E(Y_{it} - Y_{it}^* \mid d_i = 1).$$

This case is the focus of most of the attention in the literature.

The mean postprogram earnings of trainees is

$$E(Y_{it} \mid d_i = 1) = E(\alpha_{it} \mid d_i = 1) + E(Y_{it}^* \mid d_i = 1).$$

The mean postprogram earnings of nontrainees is

$$E(Y_{it} \mid d_i = 0) = E(Y_{it}^* \mid d_i = 0).$$

The difference in mean earnings between trainees and nontrainees is

$$E(Y_{it} \mid d_i = 1) - E(Y_{it} \mid d_i = 0)$$
$$= E(\alpha_{it} \mid d_i = 1) + \{E(Y_{it}^* \mid d_i = 1) - E(Y_{it}^* \mid d_i = 0)\}.$$

The expression in braces is the selection-bias term. In the case of random assignment of persons to treatment,

$$E(Y_{it}^* \mid d_i = 1) = E(Y_{it}^* \mid d_i = 0) = E(Y_{it}^*),$$

so the term in braces is 0.

When an experimental control group is not available, evaluations are conducted using various nonexperimental comparison groups. Previous analysts have used the following as comparison groups: (a) individuals who applied to the program and were rejected; (b) individuals who did not apply to the program; or (c) samples of persons similar to trainees, from population samples such as the Current Population Survey or the Panel Survey of Income Dynamics. In samples of type (c), training status is typically unknown, so the comparison group may include some trainees. (This creates the problem of *contamination bias.*) Samples of type (a) and (b) are constructed conditional on $d_i = 0$. A variety of procedures has been proposed for eliminating the effect of selection bias on estimates of program impact.

3. ALTERNATIVE NONEXPERIMENTAL ESTIMATORS FOR MEASURING THE IMPACT OF TRAINING ON EARNINGS IN THE PRESENCE OF NONRANDOM ASSIGNMENT

To conform with the training literature, we confine our review of the available models to linear specifications of earnings equations. Suppose that Y_{it}^* is a linear function of a set of observed characteristics X_{it} and unobserved characteristics represented by U_{it}. Thus

$$Y_{it}^* = X_{it}\beta + U_{it}, \qquad (3.1)$$

where β is a vector of parameters. To simplify exposition and conform with most of the literature, we assume that the training effect is invariant across individuals but not time, so $\alpha_{it} = \alpha_t$. We let $X_i = (X_{i1}, \ldots, X_{iT})$, where T is the number of periods of data on X available for each observation. (We later consider the case where the training effect depends on regressors.) With these assumptions, the equation for observed earnings may be written (using 2.1) as

$$Y_{it} = X_{it}\beta + d_i\alpha_t + U_{it}, \qquad t = 0, \ldots, T. \qquad (3.2)$$

We assume that $E(U_{it} \mid X_i) = 0$ for all i and t.

When assignment to training is nonrandom, selection bias in the estimation of α_t can arise because of dependence between d_i and U_{it}. That is, in a model without regressors,

$$E(U_{it} \mid d_i) \neq 0,$$

which is equivalent to $E(Y_{it}^* \mid d_i = 1) - E(Y_{it}^* \mid d_i = 0) \neq 0$. In a model with regressors, selection bias is present if

$$E(U_{it} \mid d_i, X_i) \neq 0, \qquad (3.3)$$

so

$$E(Y_{it} \mid d_i, X_i) \neq X_{it}\beta + d_i\alpha_t.$$

In this case, an ordinary least squares regression of Y_{it} on X_{it} and d_i does not yield consistent estimates of α_t (or β).

A stochastic relationship between d_i and U_{it} can arise for a variety of reasons. In the absence of random assignment, participation in a training program may be the result of decisions made by individuals eligible for the program, by the program administrators, or both. Whatever the decision-making procedure, it can be described in terms of an index-function framework. Let the index, IN_i, be a function of both observed (Z_i) and unobserved (V_i) variables. (The vector Z_i may include all of the variables in X_i.) For simplicity, we follow standard practice and assume that this function is linear in X_i and V_i:

$$IN_i = Z_i\gamma + V_i. \qquad (3.4)$$

Then, the ith individual's training status is

$$d_i = 1 \quad \text{iff } IN_i > 0$$
$$= 0 \quad \text{otherwise.} \qquad (3.5)$$

For simplicity, and without loss of essential generality, V_i is assumed to be iid across persons, where the distribution function of V_i is denoted by $F(v_i) = \Pr(V_i < v_i)$. Assuming that V_i is distributed independently of Z_i, we may write $\Pr(d_i = 1 \mid Z_i) = E(d_i \mid Z_i) = 1 - F(-Z_i\gamma)$, which Rosenbaum and Rubin (1983) call the propensity score.

Alternative nonexperimental selection estimators of α_t augment the earnings function and selection rule given in Equations (3.2), (3.4), and (3.5), with additional assumptions to undo the dependence between U_{it} and d_i [see (3.3)]. Nonexperimental estimators differ in the assumptions imposed, the data required to implement such estimators, and their robustness to alternative sampling plans and measurement error. Many widely used selection-bias estimators impose more assumptions than are required to recover α_t. When excess identifying conditions are invoked, they can be tested.

Heckman and Robb (1985, 1986) present a comprehensive summary of selection-bias estimators that can be implemented in alternative types of data (e.g., cross-section, repeated cross-section, and longitudinal data), and they consider the robustness of each estimator to alternative sampling plans and errors in measuring training status. Our empirical analysis of alternative nonexperimental estimators for measuring the impact of the NSW demonstration on earnings considers only a subset of the available estimators that can be implemented on the available grouped data.

Dependence between U_{it} and d_i can arise for one of two not necessarily mutually exclusive reasons: (a) dependence between Z_i and U_{it} or (b) dependence between V_i and U_{it}. We refer to the first case as *selection on observables* and the second case as *selection on unobservables*. The source of selection bias for any particular problem depends on the actual process used to select individuals; we consider each case in turn. Throughout, we assume that different

i-subscripted random variables are independently distributed.

3.1 Selection on Observables

Selection on observables occurs when the dependence between U_{it} and d_i is due to observed variables, Z_i, that influence selection into the program. More formally,

$$E(U_{it} \mid d_i, X_i) \neq 0$$
$$E(U_{it} \mid d_i, X_i, Z_i) \neq 0, \qquad (3.6)$$

but

$$E(U_{it} \mid d_i, X_i, Z_i) = E(U_{it} \mid X_i, Z_i).$$

In this case, controlling for the observed selection variables (Z_i) solves the selection-bias problem, that is, removes the dependence between latent earnings and the training dummy, d_i. In particular, one can form estimators by noting that the appropriate conditional expectation function for earnings, conditioned on d_i, X_i, and Z_i, is given by

$$\begin{aligned} E(Y_{it} \mid d_i, X_i, Z_i) &= X_{it}\beta + d_i\alpha_t + E(U_{it} \mid d_i, X_i, Z_i) \\ &= X_{it}\beta + d_i\alpha_t + E(U_{it} \mid X_i, Z_i). \end{aligned} \qquad (3.7)$$

Assuming knowledge of the functional form of $E(U_{it} \mid X_{it}, Z_i)$, this term can be inserted in (3.2), and the resulting equation can be estimated by regression methods to obtain consistent estimates of α_t. Such estimators are members of the class of control function estimators proposed by Heckman and Robb (1985), where $E(U_{it} \mid X_i, Z_i)$ are called *control functions.* Nonparametric matching procedures that contrast the earnings of trainees and comparison-group members based on the X_i and Z_i characteristics are predicated on assumption (3.6).

In our empirical analysis, we present estimates of NSW program impact using *linear control function estimators.* First proposed by Barnow, Cain, and Goldberger (1980, pp. 47–48), they are a natural starting point and, if (3.7) holds along with the linearity of $E(U_{it} \mid X_i, Z_i)$, they produce consistent estimators of program impacts. We use two variants of their estimator. Variant 1 corresponds to the assumption that $E(\alpha_{it} \mid d_i = 1, X_i, Z_i) = \alpha_t$. In this case, inserting a linear version $E(U_{it} \mid X_i, Z_i)$ in (3.2) yields

$$Y_{it} = C_i\delta_t + d_i\alpha_t + \tilde{U}_{it}, \qquad (3.8)$$

where C_i denotes the vector of all variables included in either X_i or Z_i, $\tilde{U}_{it} = U_{it} - E(U_{it} \mid d_i, C_i) = U_{it} - E(U_{it} \mid C_i)$, and δ_t is a parameter vector. Consistent estimators of α_t can be obtained by using either ordinary or weighted (by the square root of cell size) least squares to estimate (3.8). Variant 2 allows the training impact to depend on person-specific values of the regressors; that is, $E(\alpha_{it} \mid d_i = 1) = C_i\theta_t$. The appropriate estimating equation is

$$Y_{it} = C_i\delta_t + d_i(C_i\theta_t) + \tilde{U}_{it}. \qquad (3.9)$$

3.2 Selection on Unobservables

The dependence between U_{it} and the training indicator variable d_i may not be eliminated even after controlling for Z_i. That is,

$$E(U_{it} \mid d_i, X_i) \neq 0$$
$$E(U_{it} \mid d_i, X_i, Z_i) \neq E(U_{it} \mid X_i, Z_i). \qquad (3.10)$$

Selection then depends on unobservables. Several estimation procedures have been proposed to deal with selection bias when selection is on unobservables. Such estimators are formed by invoking assumptions about the distributions, or moments of the distributions, of V_i, Z_i, and U_{it}. In our analysis of the NSW data, we consider the fixed-effect and random-growth estimators. Each is appropriate under a specific set of assumptions about the form of the dependence between V_i and U_{it}.

First, consider the *fixed-effect* (or first-difference) *estimator*. Suppose that even though (3.10) holds the conditional expectation of the difference in pretraining and posttraining sets of U_{it}'s does not depend on d_i. That is, assume that the following condition holds:

$$E(U_{it} - U_{it'} \mid d_i, X_i) = 0 \quad \text{for all } t, t', t > k > t'. \qquad (3.11)$$

This specification is motivated by a model in which U_{it} is of the form

$$U_{it} = \phi_{1i} + \nu_{it},$$

where ϕ_{1i} is a zero-mean person-specific component or fixed effect and ν_{it} is a zero-mean random component independent of all other values of $\nu_{it'}$ ($t \neq t'$) and ϕ_{1i}. In this specification, selection is assumed to occur on the permanent component ϕ_{1i}; that is, this component accounts for the dependence between V_i and U_{it}.

For this model, consistent estimates of the impact of training can be obtained by regressing the difference between Y_{it} and $Y_{it'}$ on d_i and $X_{it} - X_{it'}$. As with the linear control function estimators, we estimate two variants of the fixed-effect estimator. Variant 1 of the fixed-effect estimator is derived by estimating the following regression:

$$Y_{it} - Y_{it'} = d_i\alpha_t + (X_{it} - X_{it'})\beta + (\nu_{it} - \nu_{it'}), \quad t > k > t'. \qquad (3.12)$$

If (3.11) holds, under standard rank conditions this produces a consistent estimator of α_t. Variant 2, which again allows the training impact to vary as a function of C_{it}, is obtained by estimating the following regression:

$$Y_{it} - Y_{it'} = d_i(C_i\theta_t) + (X_{it} - X_{it'})\beta + (\nu_{it} - \nu_{it'}), \quad t > k > t'. \qquad (3.13)$$

We also use a *random-growth estimator,* which is motivated by a special case of a more general class of models in which U_{it} has a factor structure. In particular, suppose that U_{it} is of the form

$$U_{it} = \phi_{1i} + t\phi_{2i} + \nu_{it}, \qquad (3.14)$$

where ϕ_{1i} is as before and ϕ_{2i} is a person-specific growth rate, where (ϕ_{1i}, ϕ_{2i}) are assumed to have zero means and finite variances and to be independent of v_{it} for all i and t. The dependence between U_{it} and d_i is assumed to arise because of dependence between d_i and (ϕ_{1i}, ϕ_{2i}).

Given (3.14), one can transform the earnings equation to eliminate ϕ_{1i} and ϕ_{2i}. Values of Y_{it} in two consecutive pretraining periods can be used to proxy these components. We obtain variant 1 of the random-growth estimator by estimating

$$\begin{aligned}(Y_{it} - Y_{it'}) - (t - t')(Y_{it'} - Y_{i,t'-1}) \\ = d_i\alpha_t + [(X_{it} - X_{it'}) - (t - t')(X_{it'} - X_{i,t'-1})]\beta \\ + [(v_{it} - v_{it'}) - (t - t')(v_{it} - v_{i,t'-1})], \qquad (3.15)\end{aligned}$$

for $t > k > t'$, by least squares. The resulting estimator of α_t is consistent under standard conditions. [Pudney (1982) proves that the asymptotic distribution of the estimator of α_t is invariant to the choice of earnings from other periods used to proxy for ϕ_{1i} and ϕ_{2i}, provided that all of the v_{it}'s have nonzero variances and the equation is estimated by generalized least squares.] As before, variant 2 of the random-growth estimator is obtained by estimating the following regression:

$$\begin{aligned}(Y_{it} - Y_{it'}) - (t - t')(Y_{it'} - Y_{i,t'-1}) \\ = d_i(C_i\theta_t) + [(X_{it} - X_{it'}) - (t - t')(X_{it'} - X_{i,t'-1})]\beta \\ + [(v_{it} - v_{it'}) - (t - t')(v_{it} - v_{i,t'-1})], \qquad (3.16)\end{aligned}$$

for $t > k > t'$.

The fixed-effect, random-growth, and linear control function estimators all yield consistent estimators of the training effect when applied to choice-based samples, because they are based on conditional (on d_i) moment restrictions. Choice-based samples are samples that mix trainees and comparison-group members in different proportions than they are found in the population (e.g., see Manski and McFadden 1981). Our sample is choice-based, so the robustness of estimators to this sampling plan is a desirable feature.

4. TESTING ALTERNATIVE SPECIFICATIONS

We propose and implement three types of model-specification tests. The first is based on access to data on preprogram earnings and regressor variables for future program participants (trainees and, when available, controls) and comparison-group members. Such data are widely available. Ignoring contamination bias, a candidate selection-correction procedure for program evaluation applied to preprogram data should make the adjusted-earnings equation of future trainees and comparison-group members alike, provided that the equation for preprogram earnings is like that for postprogram earnings (except for the additive training effect). If a candidate selection-bias adjustment does not align the preprogram earnings equations for future participants (trainees and controls) and comparison-group members, and if it is plausible to assume that the source of preprogram differences in earnings between the two types of individuals is the same as for the postprogram differences, the candidate correction procedure is rejected.

Our second test is based on additional restrictions implied by certain models. Even in the absence of preprogram data on earnings for participants and comparison-group members, it is sometimes possible to test the validity of a particular selection-adjustment procedure. Heckman and Robb (1985) note that the assumption of normality or symmetry for U_{it} that underlies many adjustment procedures can be tested in a single postprogram cross-section of program trainees and comparison-group members. Many other selection estimators are based on assumptions that can be tested. Rejection of the testable assumptions underlying a procedure would cause rejection of a candidate selection-correction method.

A third test of the validity of a nonexperimental estimator is based on access to experimental data. Controls from the experiment, who do not receive training, are pooled with comparison-group members. Controls are like trainees, except they do not receive training. Ignoring contamination bias, a valid selection-correction procedure should make the adjusted-earnings equations for controls and comparison-group members alike. Unlike the first test, this test does not assume temporal stability in the earnings equation, and thus it is more robust than a test based on preprogram earnings.

The third test is of no direct use in any particular nonexperimental evaluation, because by assumption an experimental estimate is available. Its value comes in evaluating an estimator that might be suitable for nonexperimental evaluation of the same program when experimental data are not available, or in picking an estimator for a similar program.

4.1 The Preprogram Tests

The linear control function version of this test is based on equations (3.8) and (3.9), where $t < k$, $d_i = 1$ if an observation is a future participant (trainee or control), and $d_i = 0$ if an observation is from the comparison group. If a valid linear control function has been used, estimated values of α_t and θ_t $(t < k)$ should not be statistically significantly different from 0, since no observation has undertaken training.

The fixed-effect [(3.12) and (3.13)] and random-growth [(3.15) and (3.16)] versions of this test modify these equations and use preprogram information for periods t and t'. Defining d_i as in the preceding paragraph, estimated values of α_t and θ_t should not be statistically significantly different from 0 for any correctly specified selection-correction model.

4.2 The Postprogram Tests

These tests are identical in structure to the preprogram tests, except now $t > k$ and $d_i = 1$ if an observation is a member of the experimental control group and $d_i = 0$ if an observation is a member of the comparison group. Since neither group receives training, estimated values of α_t and θ_t should not be statistically significantly different from 0 for a valid nonexperimental selection-correction estimator.

4.3 Tests of Model Restrictions

In the absence of strong beliefs about the functional form of the outcome equation and the appropriate regressor variables, there are no testable restrictions (apart from those already discussed) implied by the linear control function estimator. The situation is different for the fixed-effect and random-growth estimators. Under the assumptions that justify those estimators, values of Y from periods other than those specified by Equations (3.8), (3.9), (3.15), and (3.16) should not appear as regressors in those equations. A test that the coefficients on these extraneous Y values are equal to 0 is a test of the restrictions implied by these models. To conduct these tests, there must be enough periods of panel data for there to be extraneous Y variables. Thus in samples with two periods of panel data the fixed-effect estimator is just identified and has no testable restrictions apart from those already presented.

5. A REANALYSIS OF THE NSW DATA

In this section, we apply our specification tests to models estimated on data from the NSW experiment previously analyzed by LaLonde (1986), Fraker and Maynard (1984, 1987), and LaLonde and Maynard (1987) in their critiques of nonexperimental evaluation procedures. All NSW participants (both trainees and controls) used in our analysis were enrolled in the program in either 1976 or 1977. We focus on two demographic subgroups: high-school dropouts and women who participated in the Aid to Families with Dependent Children (AFDC) program. Fraker and Maynard obtain both pretraining and posttraining data on earnings from Social Security Administration (SSA) records for the NSW participants and comparison groups drawn from the March 1976 and March 1977 Current Population Survey (CPS). The comparison group is temporally aligned with the NSW participant group, so the preprogram and postprogram periods for this group refer to the time periods corresponding to the period of operation of the NSW program.

To protect the confidentiality of the SSA earnings information on individuals, only mean values of earnings for cells of individuals for both the NSW participants and those in the CPS were provided by the SSA. The means are for cells consisting of 7–10 sample members from the NSW trainees and controls and the CPS comparison groups, respectively, where sample members were assigned to cells on the basis of (a) date of enrollment (for NSW participants) or date of interview (for CPS comparison groups), (b) whether they had been employed in the year prior to enrollment or interview, and (c) geographical location. We used the *grouped* version of these data because they provide extensive longitudinal earnings histories. The price of using these histories is that the grouping of the data precludes the use of many nonlinear nonexperimental estimators. Many control functions are nonlinear functions of X_i and Z_i. Such estimators are not generally consistent when using grouped rather than individual-level data. The mean of a nonlinear function of X's is not equal to the nonlinear function evaluated at their mean values. Estimators that are linear functions of regressors are consistent using either grouped or individual-level data. Therefore, we restrict our investigation to linear nonexperimental estimators.

For high-school dropouts, we use the comparison group constructed by Fraker and Maynard from the CPS that consists of individuals who were (a) between ages 16 and 20 in the interview or enrollment year, (b) not in school in the interview month or at enrollment, and (c) high-school dropouts. For AFDC females, we use a CPS-based comparison group consisting of adult women who (a) are between ages 18 and 64 in the interview or enrollment year, (b) are AFDC recipients, and (c) have dependents of age 16 or less. The criteria used by Fraker and Maynard to define both comparison groups mimic the eligibility requirements for the program. Given the highly geographically concentrated nature of NSW and the small fraction of the U.S. population living in the NSW target areas, it is plausible to assume that few (if any) persons in either comparison group were eligible to participate in the NSW program, so contamination bias is unimportant in our data.

The variables used in our analysis are defined in Table 1. Mean values for trainees, controls, and the comparison group for both youths and AFDC women are given in Table 2. Table 2 demonstrates how random assignment of both youths and AFDC female NSW participants produces trainees and controls with virtually identical pretraining characteristics. In contrast, the means of most of the variables differ substantially between the CPS comparison group and the NSW participant groups (trainees and controls) for youths. For AFDC females there is closer agreement between the means of the CPS comparison group and the means of the NSW participants, but the discrepancies are still sizeable.

5.1 Estimates of the Impact of Training

Tables 3 and 4 present estimates of program impact for high-school dropouts and AFDC women, respectively, produced from the two variants of each of the three basic models presented in Section 3. The format is the same for each table. For both variants of each model, we present estimates (and their standard errors) of the impact of training on posttraining earnings in 1978 and 1979. For each year, the variant 1 columns record estimates of α_t and the variant 2 columns present estimates of the model in which training impact is of the form $\alpha_{it} = C_i\theta_t$ and is evaluated at the sample means for C_i from the NSW trainee sample $(\overline{C})$. The rows in these tables give the training impact estimates, using the following:

1. the linear control function estimator [Eqs. (3.8) and (3.9)]
2. the fixed-effect estimator [Eqs. (3.12) and (3.13)], with both 1972 and 1974 earnings for the preprogram earnings used to construct the preprogram and postprogram differences
3. the random-growth estimator [Eqs. (3.15) and (3.16)],

Table 1. Definition of Variables

Variable	*Description*
Earnings variables	
SSEARN72	SSA earnings in 1972 (in 1978 dollars)
SSEARN73	SSA earnings in 1973 (in 1978 dollars)
SSEARN74	SSA earnings in 1974 (in 1978 dollars)
SSEARN75	SSA earnings in 1975 (in 1978 dollars)
SSEARN78	SSA earnings in 1978 (in 1978 dollars)
SSEARN79	SSA earnings in 1979 (in 1978 dollars)
Background variables in *B*1	
BLKHIS	1 if black or Hispanic and 0 otherwise
SEX	1 for men and 0 for females
MARRIAGE	1 if married at enrollment for NSW participants or at March interview for CPS respondents and 0 otherwise
AGE	Age in years at enrollment for NSW participants or at March interview for CPS respondents
EDUC	Years of schooling completed at enrollment for NSW participants or at March interview for CPS respondents
URBAN	1 if in central-city standard metropolitan statistical area and 0 otherwise
7677ENR	1 if enrolled in 1977 for NSW participants or if interviewed in March 1977 for CPS respondents and 0 otherwise
Background variables in *B*2	
BLACK	1 if black and 0 otherwise
HISPANIC	1 if Hispanic and 0 otherwise
AGESQ	AGE squared
HOUSESIZE	Number of household members at enrollment for NSW participants or at March interview for CPS respondents
DEPEND	Number of dependents at enrollment for NSW participants or at March interview for CPS respondents [used only for AFDC recipient (women) results]
AGEKID	Age of youngest dependent at enrollment for NSW participants or at March interview for CPS respondents [used only for AFDC recipient (women) results]
Work history variables in *W*1	
SSEARNL1	Annual earnings (from SSA data) one year prior to enrollment for NSW participants or one year prior to interview of CPS respondents
SSEARNL2	Annual earnings (from SSA data) two years prior to enrollment for NSW participants or two years prior to the interview for CPS respondents
WORKWKS	Number of weeks worked in year prior to enrollment for NSW participants or in year prior to interview for CPS respondents
UEWKS	Number of weeks unemployed in year prior to enrollment for NSW participants or in year prior to interview for CPS respondents
AVEHRS	Average hours per week (when worked) in year prior to enrollment for NSW participants or in year prior to interview for CPS respondents
WELFARE	Per capita benefit the household received from welfare and earnings of the sample member in the month prior to enrollment for NSW participants or the interview for CPS respondents
Work history variables in *W*2	
SSEARNL3	Annual earnings (from SSA data) three years prior to enrollment for NSW participants or three years prior to interview for CPS respondents
SSEARNL4	Annual earnings (from SSA data) four years prior to enrollment for NSW participants or four years prior to interview for CPS respondents
CLERSALE	1 if job prior to enrollment for NSW participants or if current/most recent job for CPS respondents was a clerical or sales occupation and 0 otherwise
SERVICE	1 if job prior to enrollment for NSW participants or if current/most recent job for CPS respondents was in service sector and 0 otherwise
PROFESSION	1 if job prior to enrollment for NSW participants or if current/most recent job for CPS respondents was a professional occupation and 0 otherwise
AFDC	1 if AFDC received in year prior to enrollment for NSW participants or in year prior to interview for CPS respondents and 0 otherwise [used only for high-school dropout (youth) results]

with both 1972–1973 and 1973–1974 earnings for the preprogram earnings used to construct the dependent variable for this estimator

4. the experimental estimator (formed by the difference in weighted means of NSW trainees and NSW controls, where the weights used to compute these means are the square roots of the number of individuals in the grouped data cells).

Estimates are presented for each nonexperimental estimator controlling for alternative sets of X_i (and C_i), described in Table 1 and labeled $B1$, $B2$, $W1$, and $W2$. The notation $B1 + B2$ denotes the combined set of variables from $B1$ and $B2$. All estimates are obtained using $(N_j)^{1/2}$-weighted least squares to account for the grouped nature of the data, where N_j is the number of observations in cell j. The standard errors are produced from conventional formulas. The same inferences are obtained from jackknifed standard errors, which are not reported here.

The bottom rows in Tables 3 and 4 present the experimental estimates for high-school dropouts and AFDC women, respectively. For high-school dropouts, the weighted mean experimental differences are -48 for 1978 and 9 for 1979. Neither estimate is statistically significantly different from 0. For AFDC women, there is evidence of a statistically significantly positive impact of training on 1978 earnings (440) and a weaker positive effect on 1979 earnings (267). The weighted mean difference between trainees and the nonexperimental comparison group are found in the first row of the linear control function esti-

Table 2. Sample Means

	High-school dropouts (youths)			AFDC recipients (women)		
	NSW samples			NSW samples		
Variable	Trainees	Controls	CPS sample	Trainees	Controls	CPS sample
Earnings variables						
SSEARN72	192.7	228.9	201.3	971.3	1,085.6	1,041.0
SSEARN73	329.9	401.1	548.4	1,087.6	1,206.3	1,192.7
SSEARN74	581.1	630.6	1,036.6	895.3	1,000.8	1,201.9
SSEARN75	532.4	504.9	1,455.9	541.4	638.9	1,045.8
SSEARN78	1,704.0	1,751.5	3,654.8	2,007.8	1,588.9	1,841.8
SSEARN79	1,838.2	1,825.5	3,787.0	2,039.9	1,798.3	1,959.6
Background variables in *B*1						
BLKHIS	.918	.909	.196	.955	.945	.500
SEX	.883	.864	.483	.000	.000	.000
MARRIAGE	.044	.033	.285	.023	.042	.186
AGE	18.200	18.347	18.080	33.375	33.615	31.460
EDUC	9.616	9.677	10.658	10.307	10.272	11.133
URBAN	1.000	1.000	.240	.979	.981	.440
7677ENR	.645	.651	.433	.729	.719	.485
Background variables in *B*2						
BLACK	.736	.706	.110	.835	.817	.381
HISPANIC	.182	.203	.086	.120	.128	.120
HOUSESIZE	4.704	4.746	3.335	3.613	3.779	3.636
DEPEND				2.167	2.292	2.506
AGEKID				9.341	9.215	10.124
AGESQ				1,169.06	1,181.43	1,078.44
Work history variables in *W*1						
SSEARNL1	559.0	539.3	1,545.4	459.0	462.1	905.3
SSEARNL2	436.4	447.4	944.8	508.8	607.6	862.8
WORKWKS	9.3	9.3	21.4	3.3	3.2	11.0
UEWKS	10.3	11.2	2.2	11.7	13.3	1.9
HOURS	3.3	3.2	15.9	1.3	.9	7.2
WELFARE	33.8	33.7	121.6	93.9	91.6	169.7
Work history variables in *W*2						
SSEARNL3	357.1	400.6	521.9	711.8	816.5	872.5
SSEARNL4	180.6	198.1	194.7	711.9	769.1	741.1
CLERSALE	.101	.108	.126	.072	.081	.109
SERVICE	.256	.212	.228	.100	.084	.164
PROFESSION	.057	.040	.016	.013	.016	.022
AFDC	.045	.043	.146			
Number of observations	566	678	2,368	800	802	1,995
Number of cells	69	87	321	110	107	266

mates, labeled "No control variables" in the two tables. The estimates for high-school dropouts are statistically significantly negative for each year ($-1{,}910$ and $-1{,}917$ for 1978 and 1979, respectively), whereas for AFDC women they are small and statistically insignificant (157 and 79 for 1978 and 1979, respectively).

The nonexperimental estimates presented in these two tables exhibit the same kind of instability chronicled by LaLonde (1986). Different nonexperimental estimators produce very different inferences about the effect of the program. For youths, the nonexperimental estimates are always negative—often statistically significantly so. For AFDC women, the nonexperimental estimates are generally positive and often statistically significantly so, especially for the 1978 earnings measure.

Tables 3 and 4 suggest that selection bias is an empirically important problem in using nonexperimental data to evaluate the impact of training on earnings. If selection bias were not present, alternative nonexperimental methods would generate the same inference about the impact of training. The fact that alternative nonexperimental estimators produce different inferences about training indicates that some (perhaps all) of the models are misspecified. To see if it is possible to detect misspecified models, we now turn to the results from our specification tests. We consider results for youth and women in turn. Note that a limited set of the preprogram tests presented is found in Heckman et al. (1987).

5.2 Results of Model-Selection Tests for High-School Dropouts (youths)

Table 5 reports probability values (*P* values) for the specification tests described in Section 4. Under the heading "Preprogram tests using preprogram earnings," we present *P* values for the hypotheses $\alpha_t = 0$ (variant 1) and $C\theta_t = 0$ (variant 2) ($t < k$), for earnings models fit on a pooled sample of future participants (trainees and controls) and comparison-group members. Recall that for this test, $d_i = 1$ if an observation is a trainee (or control) and 0 otherwise. Tests for the vector hypotheses $\theta_t = 0$ are always consistent with tests for the more restricted hy-

Table 3. Estimates of Training Effects for High-School Dropouts (youths)

	1978 earnings		1979 earnings	
Model and control variable sets	Variant 1 (α_t)	Variant 2 ($\bar{C}\theta_t$)	Variant 1 (α_t)	Variant 2 ($\bar{C}\theta_t$)
Nonexperimental estimates				
Linear control function estimates				
No control variables	-1,910		-1,917	
	(243)		(191)	
*B*1	-1,884	-1,827	-2,119	-2,092
	(247)	(246)	(342)	(300)
*B*1 + *B*2	-1,279	-1,079	-1,569	-1,498
	(273)	(295)	(239)	(319)
*B*1 + *W*1	-1,117	-1,146	-1,539	-1,447
	(246)	(263)	(343)	(372)
*B*1 + *B*2 + *W*1 + *W*2	-889	-889	-996	-1,331
	(328)	(380)	(442)	(388)
Fixed-effect estimates constructed with $t' = 1972$ pretraining earnings				
No control variables	-1,904		-1,910	
	(236)		(266)	
*B*1	-1,886	-1,831	-2,172	-2,070
	(242)	(201)	(277)	(275)
*B*1 + *B*2	-1,360	-1,227	-1,644	-1,647
	(270)	(291)	(309)	(330)
Fixed-effect estimates constructed with $t' = 1974$ pretraining earnings				
No control variables	-1,456		-1,462	
	(203)		(166)	
*B*1	-1,411	-1,370	-1,663	-1,636
	(227)	(228)	(301)	(269)
*B*1 + *B*2	-1,035	-964	-1,330	-1,383
	(255)	(276)	(326)	(312)
Random-growth estimates constructed with $t' = 1973$ and $t' - 1 = 1972$ pretraining earnings				
No control variables	-649		-446	
	(336)		(386)	
*B*1	-231	-235	-241	-236
	(414)	(416)	(475)	(477)
*B*1 + *B*2	-23	76	-85	-126
	(476)	(515)	(547)	(589)
Weighted average of estimates	-24		-154	
	(185)		(212)	
Random-growth estimates constructed with $t' = 1974$ and $t' - 1 = 1973$ pretraining earnings				
No control variables	-499		-267	
	(328)		(307)	
*B*1	-614	-589	-701	-659
	(431)	(436)	(510)	(515)
*B*1 + *B*2	-624	-777	-806	-850
	(497)	(537)	(586)	(630)
Weighted average of estimates	-616		-724	
	(426)		(502)	
Experimental estimates	-48		9	
	(144)		(173)	

NOTE: Standard errors are in parentheses.

pothesis $\bar{C}\theta_t = 0$ and for the sake of brevity are not reported.

Under the headings "Postprogram tests," we present P values for tests of the hypotheses $\alpha_t = 0$ and $\bar{C}\theta_t = 0$ ($t > k$), for earnings models fit on a pooled sample of experimental controls from the experiment ($d_i = 1$) and comparison-group members ($d_i = 0$). Again, tests of the vector hypothesis $\theta_t = 0$ are consistent with the test based on $\bar{C}\theta_t$ and are not reported here.

Under the heading "Model-restriction tests," we report P values for the hypotheses that extraneous Y values do not have statistically significant coefficients in the fixed-

Table 4. Estimates of Training Effects for AFDC Recipients (women)

	1978 earnings		1979 earnings	
Model and control variable sets	Variant 1 (α_t)	Variant 2 ($\bar{C}\theta_t$)	Variant 1 (α_t)	Variant 2 ($\bar{C}\theta_t$)
Nonexperimental estimates				
Linear control function estimates				
No control variables	157		79	
	(164)		(155)	
*B*1	686	726	494	534
	(192)	(194)	(193)	(195)
*B*1 + *B*2	231	638	-195	546
	(282)	(358)	(286)	(360)
*B*1 + *W*1	653	715	496	370
	(203)	(260)	(230)	(289)
*B*1 + *B*2 + *W*1 + *W*2	937	907	441	586
	(263)	(335)	(303)	(386)
Weighted average of estimates	374		238	
	(146)		(152)	
Fixed-effect estimates constructed with $t' = 1972$ pretraining earnings				
No control variables	231		153	
	(152)		(156)	
*B*1	699	736	508	544
	(185)	(188)	(193)	(195)
*B*1 + *B*2	938	1,124	512	1,032
	(275)	(353)	(287)	(362)
Fixed-effect estimates constructed with $t' = 1974$ pretraining earnings				
No control variables	475		397	
	(135)		(150)	
*B*1	713	693	522	500
	(168)	(170)	(179)	(184)
*B*1 + *B*2	946	800	520	708
	(250)	(321)	(268)	(328)
Random-growth estimates constructed with $t' = 1973$ and $t' - 1 = 1972$ pretraining earnings				
No control variables	494		460	
	(367)		(433)	
*B*1	78	44	-217	-263
	(463)	(473)	(546)	(557)
*B*1 + *B*2	486	-936	-15	-1,372
	(692)	(898)	(816)	(1,965)
Random-growth estimates constructed with $t' = 1974$ and $t' - 1 = 1973$ pretraining earnings				
No control variables	1,276		1,398	
	(356)		(471)	
*B*1	1,183	981	1,109	860
	(453)	(453)	(576)	(576)
*B*1 + *B*2	1,278	880	935	808
	(677)	(869)	(778)	(918)
Experimental estimates	440		267	
	(142)		(162)	

NOTE: Standard errors are in parentheses.

Table 5. Specification Tests of Nonexperimental Estimators for High-School Dropouts (youths)

	Probability values											
			Model-restriction tests						Postprogram tests			
	Preprogram tests using preprogram earnings		Preprogram earnings		1978 earnings		1979 earnings		1978 earnings		1979 earnings	
Control variable set	$\alpha_t = 0$	$\bar{C}\theta_t = 0$	$\alpha_t = 0$	$\bar{C}\theta_t = 0$	$\alpha_t = 0$	$\bar{C}\theta_t = 0$	$\alpha_t = 0$	$\bar{C}\theta_t = 0$	$\alpha_t = 0$	$\bar{C}\theta_t = 0$	$\alpha_t = 0$	$\bar{C}\theta_t = 0$
	1975 earnings as dependent variable				*1978 or 1979 earnings as dependent variable*							
Linear control function estimators												
No control variables	.000								.000		.000	
*B*1	.000	.000							.000	.000	.000	.000
*B*1 + *B*2	.000	.012							.000	.000	.000	.000
*B*1 + *W*1	.000	.208							.000	.012	.000	.000
*B*1 + *B*2 + *W*1 + *W*2	.016	.336							.005	.632	.033	.000
	t = 1974 and t′ = 1972 earnings				*t = 1978 or 1979 and t′ = 1972 earnings*							
Fixed-effect estimators												
No control variables	.000		.000		.000		.000		.000		.000	
*B*1	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
*B*1 + *B*2	.000	.019	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	t = 1975 and t′ = 1972 earnings				*t = 1978 or 1979 and t′ = 1974 earnings*							
Fixed-effect estimators												
No control variables	.000		.000		.502		.715		.000		.000	
*B*1	.000	.000	.000	.000	.076	.081	.661	.629	.000	.002	.000	.001
*B*1 + *B*2	.000	.000	.000	.000	.023	.021	.524	.817	.000	.002	.000	.001
	t = 1975, t′ = 1973, t′ − 1 = 1972 earnings				*t = 1978 or 1979, t′ = 1973, t′ − 1 = 1972 earnings*							
Random-growth estimators												
No control variables	.000		.035		.000		.000		.007		.042	
*B*1	.375	.173	.316	.080	.000	.000	.000	.000	.329	.113	.427	.443
*B*1 + *B*2	.558	.128	.614	.042	.000	.000	.000	.000	.798	.695	.622	.608
	t = 1975, t′ = 1974, t′ − 1 = 1973 earnings				*t = 1978 or 1979, t′ = 1974, t′ − 1 = 1973 earnings*							
Random-growth estimators												
No control variables	.000		.126		.000		.003		.139		.474	
*B*1	.301	.172	.809	.817	.090	.105	.281	.353	.567	.398	.821	.417
*B*1 + *B*2	.352	.121	.909	.659	.070	.146	.276	.546	.696	.312	.698	.169

effect and random-growth models fit on preprogram earnings [for a pooled sample of future trainees and controls ($d_i = 1$) and comparison-group members ($d_i = 0$)] and postprogram earnings [for a pooled sample of controls from the experiment ($d_i = 1$) and comparison-group members ($d_i = 0$)]. We do not see any compelling model restrictions for the linear control function estimator, so no test is reported.

The preprogram and postprogram tests and the model-restriction tests generally produce consistent findings. Linear control function and fixed-effect models are decisively rejected; the random growth model is not. (Though not reported in Table 5, the coefficient estimates associated with the preprogram and postprogram tests range from −2,128 to −167 for the linear control function models, from −2,186 to −274 for the fixed-effect models, and from −894 to −37 for the random-growth models.) Nevertheless, the tests for the model restrictions applied to the random-growth model fit on postprogram data are mixed. Using 1973 and 1972 earnings to proxy the unobserved components, ϕ_{1i} and ϕ_{2i}, the model is rejected on the postprogram data. Using 1973 and 1974 earnings to proxy the unobserved components, the model is not rejected on the postprogram data. Neither version of the random-growth model with regressors is rejected when it is fit on the preprogram sample that combines future participants (trainees and controls) and comparison-group members.

The rejected version of the model uses the longest lags in preprogram earnings of any of the fitted models to eliminate the permanent and random-growth components in U_{it} ($t - t' = 6$ and 7 for 1978 and 1979, respectively). Earnings functions are well known to be concave in age or experience. The linear growth specification (3.14) may become a progressively poorer approximation as the lag length increases between the dependent variable and the proxy variables. A better model might augment (3.14) to include a third component, ϕ_{3i}, multiplied by $(t - t')^2$. To find sufficient proxy variables for this model requires a third year of preprogram earnings data, which is not available to us. Models with autoregressive specifications for U_{it} failed specification tests.

A slight extension of (3.14) produces a model that passes specification tests and produces estimates of program impact very close to those obtained from the random-growth model reported in Table 3. In place of (3.14), we write

$$U_{it} = \phi_{1i} + b_t\phi_{2i} + \nu_{it}, \qquad b_t \neq b_{t'}, t \neq t',$$

which permits the growth component to be unrestricted. The coefficient on $Y_{it'} - Y_{i,t'-1}$ in (3.15) and (3.16), defined $\omega_{t',t'-1} = (b_t - b_{t'})/(b_{t'} - b_{t'-1})$, now becomes a parameter to be estimated. Moving $(Y_{it'} - Y_{i,t'-1})$ multiplied by this coefficient to the right side of those equations and using instrumental variables to account for endogeneity of the regressor produces a model that is not rejected by any of the postprogram specification tests (see Table 6).

The endogeneity in this variable is due to its dependence on $v_{it'}$ and $v_{i,t'-1}$, which appear as disturbances in the equations. Variables in $B2$ are used as instruments assuming that the variables in $B1$ belong in the earnings equation. Empirical results based on these instruments must be qualified in light of the finding by Heckman and Robb (1985) that unweighted instrumental variables procedures are not robust to choice-based sampling. We do not have access to the data required to construct the appropriate weights to guarantee consistency of the instrumental variables estimator.

The estimates of the free parameters on $(Y_{it'} - Y_{i,t'-1})$ ($\omega_{t',t'-1}$ in Table 7) are not exactly equal to $(t - t')$, but one cannot reject the hypothesis that they are equal to $(t - t')$ at conventional levels of significance. Reading across the first row of Table 7, the estimated values of $\omega_{t',t'-1}$ for 1978 earnings are close to $(t - t') = 6$, whereas the estimated values of $\omega_{t',t'-1}$ for 1979 earnings are close to $(t - t') = 7$. The true specification for youths in this program seems to be quite close to the random-growth model. The estimated program impacts in this table are close to those reported in Table 3 using $B1$ variables as regressors.

Returning to Table 3, note that the random-growth model produces the same inference about program impact as the experiment—that training has no effect on earnings. The estimates from the random-growth model are more negative than the estimates produced by the experiment, but the standard errors for the nonexperimental estimator are bigger.

Since we have no basis for distinguishing among the various versions of the random-growth model that control for alternative sets of variables in X_i and survive our battery of tests, we compute a weighted average of estimates for the random-growth models fit using alternative sets of preprogram years for t' and $t' - 1$. The weighted average is formed by applying generalized least squares to the regression coefficients of each model, treated as observations, using the variance–covariance matrix of each regression coefficient to form the variance for each observation. Such weighted averages can be given a Bayesian justification (see Leamer 1978). The weighted averages, based on the alternative random-growth estimators not rejected by the postprogram and model-restriction tests using 1978 and 1979 earnings, are presented at the bottom of Table 3. Although the point estimates differ from those obtained from experimental data, they produce the same conclusion as the experiment, namely that there is no statistically significant impact of the NSW program on youth earnings. The estimates from the modified random-growth estimator presented in Table 7 produce the same inference and are closer to the experimental results. (Note that the same random-growth estimators are rejected by the preprogram tests and tests of model restrictions based on preprogram earnings data. Therefore, even in the absence of the experimental data, one would have chosen the same set of random-growth estimators and thus obtained the same weighted average estimates as in Table 3.)

Observe that the higher the P value for an estimator displayed in Tables 5 and 6, the lower (on average) the discrepancy between the nonexperimental and experimental estimates. Low P values indicate model misspecification. Such misspecification should widen the difference between the estimate obtained from the misspecified model and the experimental estimate.

For youths our testing strategies lead to a conclusion quite different from that reported by Fraker and Maynard and LaLonde. The models that survive our tests yield the same conclusion about the impact of the NSW training program on the earnings of youths as the experiment. The rejected models are the source of the discrepancy in inference between experimental and nonexperimental estimates reported in the literature. Note, however, that the nonexperimental estimators not rejected by these specification tests have much larger standard errors than the experimental estimators. Experimental data produce a sharper inference. Nonetheless, our results for youths suggest that pessimism concerning the use of nonexperimental

Table 6. Specification Tests of Modified Random-Growth Estimators for High-School Dropouts (youths)

	Probability values							
	Model-restriction tests				Postprogram tests			
	1978 earnings		1979 earnings		1978 earnings		1979 earnings	
Control variable set	$\alpha_t = 0$	$\bar{C}\theta_t = 0$	$\alpha_t = 0$	$\bar{C}\theta_t = 0$	$\alpha_t = 0$	$\bar{C}\theta_t = 0$	$\alpha_t = 0$	$\bar{C}\theta_t = 0$
	$t' = 1973, t' - 1 = 1972$ pretraining earnings							
Modified random-growth estimators								
$B1$	.065	.050	.063	.040	.333	.401	.582	.658
	$t' = 1974, t' - 1 = 1973$ pretraining earnings							
Modified random-growth estimators								
$B1$	.095	.137	.321	.447	.881	.696	.948	.840

Table 7. Estimates of Training Effects Using Modified Random-Growth Estimators for High-School Dropouts (youths)

Control variable set	1978 earnings				1979 earnings			
	Variant 1		Variant 2		Variant 1		Variant 2	
	α_t	$\omega_{t',t'-1}$	$\bar{C}\theta_t$	$\omega_{t',t'-1}$	α_t	$\omega_{t',t'-1}$	$\bar{C}\theta_t$	$\omega_{t',t'-1}$
	t′ = 1973, t′ − 1 = 1974 pretraining earnings							
Modified random-growth estimators								
*B*1	−191	5.836	−183	5.942	−277	6.749	−270	6.927
	(329)	(1.021)	(298)	(1.253)	(351)	(.918)	(379)	(1.124)
	t′ = 1974, t′ − 1 = 1973 pretraining earnings							
Modified random-growth estimators								
*B*1	−237	4.691	−201	4.573	−237	5.213	−213	5.011
	(367)	(1.001)	(361)	(1.116)	(385)	(1.023)	(370)	(.927)

NOTE: Standard errors are in parentheses.

estimators may not be well-founded and that a systematic procedure exists to identify estimators that replicate the inferences drawn from experimental methods.

5.3 Results of Model-Selection Tests for AFDC Recipients (women)

Table 8 reports the results of specification tests applied to alternative earnings equations for AFDC women. The format of this table is the same as that of Table 5. Neither the preprogram tests nor the postprogram model-specification tests are decisive in rejecting any of the models. (The coefficient estimates associated with these tests range from −686 to 476 for the linear control function models, −449 to 750 for the fixed-effect models, and −1,961 to 1,071 for the random-growth models.) The tests of model restrictions have much more bite. For *both* the preprogram and postprogram versions of these tests, the fixed-effect and random-growth models are decisively rejected. By

Table 8. Specification Tests of Nonexperimental Estimators for AFDC Recipients (women)

	Probability values											
			Model-restriction tests						Postprogram tests			
	Preprogram tests using preprogram earnings		Preprogram earnings		1978 earnings		1979 earnings		1978 earnings		1979 earnings	
Control variable set	$\alpha_t = 0$	$\bar{C}\theta_t = 0$	$\alpha_t = 0$	$\bar{C}\theta_t = 0$	$\alpha_t = 0$	$\bar{C}\theta_t = 0$	$\alpha_t = 0$	$\bar{C}\theta_t = 0$	$\alpha_t = 0$	$\bar{C}\theta_t = 0$	$\alpha_t = 0$	$\bar{C}\theta_t = 0$
	1975 earnings as dependent variable				*1978 or 1979 earnings as dependent variable*							
Linear control function estimators												
No control variables	.000								.082		.246	
*B*1	.274	.817							.196	.098	.210	.121
*B*1 + *B*2	.000	.436							.217	.404	.192	.265
*B*1 + *W*1	.199	.021							.204	.358	.402	.806
*B*1 + *B*2 + *W*1 + *W*2	.139	.010							.435	.232	.973	.429
	t = 1974 and t′ = 1972 earnings				*t = 1978 or 1979 and t′ = 1972 earnings*							
Fixed-effect estimators												
No control variables	.000		.000		.000		.000		.031		.141	
*B*1	.608	.622	.000	.000	.000	.000	.000	.000	.455	.251	.490	.311
*B*1 + *B*2	.561	.131	.000	.000	.000	.000	.000	.000	.837	.074	.898	.045
	t = 1975 and t′ = 1972 earnings				*t = 1978 or 1979 and t′ = 1974 earnings*							
Fixed-effect estimators												
No control variables	.000		.000		.144		.014		.574		.893	
*B*1	.128	.824	.000	.000	.022	.047	.004	.009	.383	.340	.430	.423
*B*1 + *B*2	.307	.299	.000	.000	.019	.014	.001	.001	.701	.402	.772	.280
	t = 1975, t′ = 1973, t′ − 1 = 1972 earnings				*t = 1978 or 1979, t′ = 1973, t′ − 1 = 1972 earnings*							
Random-growth estimators												
No control variables	.021		.000		.000		.000		.738		.989	
*B*1	.016	.055	.000	.000	.000	.000	.000	.000	.227	.243	.208	.208
*B*1 + *B*2	.183	.069	.000	.000	.000	.000	.000	.000	.375	.074	.358	.075
	t = 1975, t′ = 1974, t′ − 1 = 1973 earnings				*t = 1978 or 1979, t′ = 1974, t′ − 1 = 1973 earnings*							
Random-growth estimators												
No control variables	.999		.102		.002		.004		.022		.008	
*B*1	.827	.974	.033	.021	.000	.000	.000	.000	.124	.267	.135	.303
*B*1 + *B*2	.686	.985	.040	.037	.000	.000	.000	.000	.267	.659	.277	.627

default, we do not reject the linear control function estimators.

Returning to Table 4, we examine the inference about the impact of the experiment derived from the linear control function specification. A generalized least squares average of the linear control function estimators not rejected by the postprogram tests produces the number shown in the row labeled "Weighted average of estimates." (In the absence of experimental data, a slightly different set of linear control function estimators not rejected by the preprogram tests would be included in this average. In this case, the corresponding average estimates are 702 with a standard error of 168 for 1978 earnings and 515 with a standard error of 179 for 1979 earnings.) The 1978 weighted nonexperimental estimate, although somewhat lower than the experimental estimate, leads to the same strong inference—that training raised the earnings of trainees. The 1979 weighted nonexperimental estimate leads to an inference similar to that obtained from the experiment—a slightly weaker, but still positive, effect of training on earnings. Again, the difference in inference between experimental and nonexperimental estimates arises from the rejected models.

6. CONCLUSIONS

This article considers the problem of assessing the validity of alternative nonexperimental evaluation estimators. We critically examine the claims of LaLonde and Fraker and Maynard concerning the difficulty in using nonexperimental methods to evaluate social programs. A simple model-selection strategy based on easily implemented specification tests eliminates nonexperimental evaluation models that do not produce estimated program impacts close to the experimental results: The models that are not rejected produce impacts close to the experimental results, at least in the case of women on AFDC.

We do not claim that we have found the "true" model for either youths or AFDC women. Such a claim would be premature. As noted previously, a variety of nonlinear nonexperimental estimators could not be implemented in this study because of the grouped nature of the data available. Nonetheless, using the same data set analyzed by critics of nonexperimental data, our analysis demonstrates that simple specification tests eliminate the most unreliable and misleading estimators that give rise to the sensitivity problem recently discussed in the evaluation literature. Thus, while not definitive, our results are certainly encouraging for the use of nonexperimental methods in social-program evaluation.

[*Received November 1987. Revised August 1989.*]

REFERENCES

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

Barnow, B. (1987), "The Impact of CETA Programs on Earnings: A Review of the Literature," *Journal of Human Resources,* 22, 157–193.

Barnow, B., Cain, G., and Goldberger, A. (1980), "Issues in the Analysis of Selectivity Bias," *Evaluation Studies,* 5, 42–59.

Barros, R. (1987), "Two Essays on the Nonparametric Estimation of Economic Models With Selectivity Using Choice-Based Samples," unpublished Ph.D. dissertation, University of Chicago, Dept. of Economics.

Burtless, G., and Orr, L. (1986), "Are Classical Experiments Needed for Manpower Policy?" *Journal of Human Resources,* 21, 606–639.

Fraker, T., and Maynard, R. (1984), *An Assessment of Alternative Comparison Group Methodologies for Evaluating Employment and Training Programs,* Princeton, NJ: MPR, Inc.

——— (1987), "Evaluating Comparison Group Designs With Employment-Related Programs," *Journal of Human Resources,* 22, 194–227.

Heckman, J., Hotz, V. J., and Dabos, M. (1987), "Do We Need Experimental Data to Evaluate the Impact of Manpower Training on Earnings?" *Evaluation Review,* 11, 395–427.

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

——— (1986), "Alternative Identifying Assumptions in Econometric Models of Selection Bias," in *Advances in Econometrics: Innovations in Quantitative Economics, Essays in Honor of Robert L. Basmann* (Vol. 5), ed. D. Slottje, Greenwich, CT: JAI Press, pp. 243–287.

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

LaLonde, R., and Maynard, R. (1987), "How Precise Are Evaluations of Employment and Training Programs: Evidence From a Field Experiment," *Evaluation Review,* 11, 428–451.

Leamer, E. (1978), *Specification Searches: Ad Hoc Inference With Nonexperimental Data,* New York: John Wiley.

Manski, C., and McFadden, D. (1981), "Alternative Estimators and Sample Designs for Discrete Choice Analysis," in *Structural Analysis of Discrete Data With Econometric Applications,* eds. C. Manski and D. McFadden, Cambridge, MA: MIT Press, pp. 2–50.

Pudney, S. (1982), "Estimating Latent Variable Systems When Specification Is Uncertain: Generalized Component Analysis and the Eliminant Method," *Journal of the American Statistical Association,* 77, 883–889.

Rivlin, A. (1971), *Systematic Thinking for Social Action,* Washington, DC: Brookings Institution.

Rosenbaum, P., and Rubin, D. (1983), "The Central Role of The Propensity Score in Observational Studies for Causal Effects," *Biometrika,* 70, 41–55.

Rossi, P., and Freeman, H. (1985), *Evaluation: A Systematic Approach* (3rd ed.), Beverly Hills, CA: Sage Publications.

Stromsdorfer, E., Boruch, R., Bloom, H., Gueron, J., and Stafford, F. (1985), "Recommendations of the Job Training Longitudinal Survey Research Advisory Panel," report to the Employment and Training Administration, U.S. Department of Labor, Washington, DC.