Evaluating Program Evaluations: New Evidence on Commonly Used Nonexperimental Methods

By DANIEL FRIEDLANDER AND PHILIP K. ROBINS*

Previous research has demonstrated that analysts can use social experiments to evaluate the likely reliability of nonexperimental program-evaluation methods (see Robert LaLonde, 1986; Thomas Fraker and Rebecca Maynard, 1987; LaLonde and Maynard, 1987; James J. Heckman and V. Joseph Hotz, 1989).[1] These studies "evaluate" a nonexperimental evaluation method by applying that method to the data set produced by a true random assignment experiment and comparing the results to the estimates produced by the experimental method. The experimental estimates are presumed to be unbiased estimates of the true program effects. Significant differences between the two sets of estimates are taken to mean that the econometric model generating the nonexperimental estimates is misspecified. Analysts can further use the experimental results to examine whether standard specification tests of the nonexperimental econometric model properly lead them to reject specifications that yield estimates inconsistent with the experimental findings (Orley Ashenfelter, 1978; Ashenfelter and David Card, 1985).

In this study, we extend this approach to assess two conventional nonexperimental strategies for estimating the effect of social programs.[2] The first approach estimates the effects of a policy change in one area by comparing persons in that area to those in a jurisdiction not affected by the policy change.[3] For example, in Long and Wissoker (1992), individuals in several counties in the state of Washington were given a new welfare-to-work program, and their behavior was compared to the behavior of individuals in other counties where

*Manpower Demonstration Research Corporation, 3 Park Avenue, New York, NY 10016, and Department of Economics, University of Miami, Coral Gables, FL 33124-6550, respectively. The research reported in this paper was supported by a grant from the Alfred P. Sloan Foundation. The findings and conclusions do not necessarily represent the official position or policies of the funding organization. Previous versions of this paper were presented at the 1992 meetings of the American Economic Association, the 1991 meetings of the Association for Public Policy and Management, and in seminars at the University of Wisconsin, Penn State University, Rutgers University, the University of North Carolina, Columbia University, the Urban Institute, and the Alfred P. Sloan Foundation. We are indebted to George Cave, whose program created the statistically matched comparison groups used in this study, and to an anonymous referee who provided detailed guidance on structuring the material presented in this paper. Helpful comments were also received on earlier versions of this paper from George Cave, Barbara Goldman, Judith Gueron, James Heckman, V. Joseph Hotz, Robert LaLonde, Robert Meyer, Robert Moffitt, Charles Romeo, Jeffrey Smith, Michael Wiseman, other seminar and workshop participants, and members of the MDRC Committee on Welfare Studies. Outstanding research assistance was provided by Dan Edelstein and Scott Susin.

[1] In this paper, an experimental evaluation is defined as one based on a comparison between a program group and a control group, both of which are created by randomly assigning sample members to one group or the other. A nonexperimental evaluation is one that uses a comparison group not generated by random assignment as a substitute for a control group.

[2] For a general discussion of nonexperimental strategies for evaluating social programs, see Robert Moffitt (1991). Robinson G. Hollister and Jennifer Hill (1995) provide numerous examples of nonexperimental program evaluations in economics and other disciplines.

[3] The comparisons may be made across states (George Farkas et al., 1983, 1984; Denise Polit et al., 1985), across counties (or cities) within states (Randall Brown et al., 1983; Irwin Garfinkel et al., 1992; Sharon K. Long and Douglas A. Wissoker, 1982; Bradley R. Schiller and C. Nielsen Brasher, 1993), or across areas within cities (David A. Long, 1991).

the new program was not available. The second approach compares the behavior of persons in a particular area covered by the policy change to the behavior of individuals in the same area before the change went into effect. For example, in Paul T. Decker (1991), welfare recipients in New Jersey were given access to a new program of employment and training services, and their behavior was compared to the behavior of welfare recipients from a period prior to the implementation of the new program.

The obvious shortcoming of these conventional nonexperimental evaluation strategies is the inherent difficulty in controlling for differences in local conditions between the program and comparison sites or for changes in these conditions over time. Nonetheless, it is important to evaluate such comparison strategies because, in practice, variants of them are often the only ones available to an analyst. Data limitations, opposition to social experiments, or the nature of certain programs as an "entitlement" could prevent analysts from randomly assigning some persons in the program locality to a no-program control group.

In this paper, we follow previous research by using experimental data to assess the two nonexperimental evaluation approaches. Our data are from a series of social experiments conducted in several states during the 1980's to evaluate programs aimed at helping welfare recipients find jobs. We simulate the two nonexperimental approaches by creating comparison groups from the true control groups and by comparing the resultant nonexperimental estimates of program effects with experimentally derived estimates of program effects. Although experimental data are required to provide an assessment of the nonexperimental approaches, they are not required to create the comparison groups in practice. Thus, our results have direct implications for choosing a nonexperimental evaluation strategy when an experiment is not feasible.

The remainder of this paper is organized as follows. Section I describes the social experiments and the various comparison groups created for the analysis. Section II discusses the methods used to generate and assess the nonexperimental estimates. Section III presents the empirical results, and Section IV offers some conclusions.

I. Data

In the early 1980's, a number of states undertook changes in their welfare-to-work programs. The aim of these program changes was to increase employment and to decrease welfare receipt. These new programs could require welfare recipients, as a condition for receiving the full amount of their monthly welfare payment, to look for work and attend job-search assistance groups, to participate in job training, or to work at an unpaid, government-sponsored job.[4] To estimate the effects of the new programs, several states relied on social experiments: welfare recipients eligible for a program were randomly assigned either to a program group, which was subject to the program requirements, services, and penalties, or to a control group, which was not. Among their various goals, these experiments were designed to estimate effects on the employment and welfare receipt of the program group.

The four experiments analyzed in this paper are the evaluations of the Arkansas WORK Program, the Baltimore Options Program, the San Diego Saturation Work Initiative Model (SWIM), and the Virginia Employment Services Program (ESP).[5] All

[4] Unpaid work assignments were usually part-time and limited to three months.

[5] These four experiments are from a set of nine that were stimulated by federal welfare reform legislation in 1981. They were selected because they were evaluations of large-scale programs and had at least two years of follow-up data. Summaries of the research findings for all nine experiments (plus some others) are presented in Judith M. Gueron (1990), Gueron and Edward Pauly (1991), Friedlander and Gueron (1992), and Greenberg and Wiseman (1992). Details of the four experiments used in this study are given in Friedlander et al. (1985a, b), Gayle Hamilton and Friedlander (1989), and James Riccio et al. (1986).

of these experiments were initiated within a three-year period, 1982–1985. All were implemented for single-parent families (headed mostly by women) who were receiving benefits from the Aid to Families with Dependent Children (AFDC) program. Each experimental program was mandatory in that failure to participate could result in a partial, temporary reduction in AFDC payments. Individuals with children under the age of six (three in Arkansas) were exempted from the participation requirement and were not part of the research samples.

Typically, about half of all persons enrolled in a program group actually participated in a formal employment and training activity. The remainder found work, left AFDC before participating, were penalized for failure to participate, or in some cases, remained on AFDC and were not reached by the employment program at all. Owing largely to differences in the levels of participation and the mix of activities, the programs' net costs per program group member varied from $118 in Arkansas to $953 in Baltimore. Local environments varied considerably, too. State monthly AFDC grant levels ranged from $140 in Arkansas to $526 in San Diego. Unemployment rates varied from 6.6 percent in Virginia to 11.0 percent in Arkansas.

The experimental evaluations found that all four programs increased employment, and two of the four programs (Arkansas and San Diego) reduced welfare receipt.[6] In this study, we focus on employment effects.[7] We define short-term employment as having any earnings during the third quarter of the experiment, where the first quarter is the quarter of random assignment. We define long-term employment as having any earnings during quarters 6–9 of the experiment (the second year after random assignment).

II. Construction of Comparison Groups

For this study, we constructed four comparison groups for each of the welfare-to-work programs. The first utilizes the control group from another state (or from all three other states combined) as a comparison group for the program group in the original state.[8] The program and comparison groups are similar in that they are all AFDC case heads who meet the same eligibility criteria for a mandatory welfare-to-work program, although local labor-market conditions varied considerably across the states and the procedures for bringing individuals into the samples differed in some details.

The second comparison group is based on statistical matching. The observed background characteristics of each program group member are matched against those of every candidate for the comparison group; the candidate most closely resembling the program group member is selected to be in the comparison group (see e.g., Katherine P. Dickinson et al., 1984, 1987). In this study, we use a procedure in which each program group member is matched non-

[6]A complete set of experimental estimates of program effects is presented in an unpublished appendix available from the authors upon request.

[7]We analyze employment rather than earnings to eliminate the need to adjust for state differences in wage levels and the cost of living. To simplify the exposition, we do not present results for welfare receipt. In an earlier version of this paper, available from the authors upon request, we report limited comparisons of experimental and nonexperimental estimates for the welfare outcomes. Generally, the welfare estimates performed worse than the employment estimates, and the two often produced conflicting results (i.e., nonexperimental welfare estimates were accurate when nonexperimental employment estimates were not, and vice versa).

[8]For example, in the case of the Arkansas WORK program, we generate three separate nonexperimental estimates of the program effect using the control group from the Baltimore Options program, the San Diego SWIM program, and the Virginia ESP program. A fourth nonexperimental estimate is produced by creating a composite comparison group consisting of the control groups from Baltimore, San Diego, and Virginia combined. We compare the nonexperimental estimates produced by each of these four comparison groups with the one experimental estimate produced for Arkansas.

parametrically to a "nearest neighbor" in the control groups from the other states using a Mahalanobis distance metric.[9]

The first two comparison groups are subject to confounding interstate differences. The third and fourth comparison groups are not. The third is based on comparisons of two areas or offices within a state or locality. It may be feasible to use this approach for cases in which a new program (or a major program innovation) can be implemented earlier in one place than in another. For example, in the case of welfare-to-work programs, advance implementation may be planned for one or a group of welfare offices out of several in an urban area. Pilot testing of this sort is common, although such testing is usually undertaken to identify problems in program design rather than to estimate program effects. Nonetheless, a sample collected from the nonprogram offices could be used as a comparison group for a sample collected under the same procedures and during the same time period as in the program offices. In our study, we simulate this cross-site strategy in the Arkansas and San Diego samples.[10] We split each of these samples into two local office samples. For each office, we use the program group and produce nonexperimental estimates of the program effect using controls from the other office as a comparison group. These estimates are then compared to the experimental estimates computed from the program and control groups from the same office.

The fourth comparison group is based on a before–after comparison of outcomes within a particular area or office and can be utilized when a new program or a significant program change is about to be implemented. For several months or a year, or even for several years prior to implementation, a sample of persons who would be eligible for the program can be drawn to serve as a comparison group. Once the program begins, the same sampling procedures are continued, but the new sample members constitute the program group. As in the cross-site strategy, state differences are eliminated from the comparison, although cyclical labor-market differences may well be important, as may other events occurring at the locality. To simulate this approach, we partition our samples roughly in half, into an early and a late cohort according to date of random assignment.[11] Then, for each of the four programs (i.e., within each state), program group members from the late cohort become the program group for the analysis, and controls from the early cohort

[9] The Mahalanobis distance is given by

$$d = (\mathbf{x}_p - \mathbf{x}_c)'\mathbf{T}^{-1}(\mathbf{x}_p - \mathbf{x}_c)$$

where $\mathbf{x}$ is the vector of characteristics for a program group member (p) and a matched comparison group member (c), and $\mathbf{T}$ is the total sample covariance matrix. The data are sorted randomly, and for each $\mathbf{x}_p$ an $\mathbf{x}_c$ is chosen for which d is smallest. Because a comparison-group observation is not used again after it is matched, the results can vary depending on how the data are initially sorted. For a general description of nearest-neighbor techniques, see D. J. Hand (1981). It should be noted that there are several variants of the nearest-neighbor technique as well as other methods of statistical matching. The matched comparison samples were created with a program written by George Cave of the Manpower Demonstration Research Corporation. The matches were made on a set of 16 observed baseline characteristics, including prior employment and welfare history, age, number and ages of children, education, type of welfare case (applicant or recipient), marital status, and ethnicity.

[10] The two local offices for San Diego SWIM are in the same city (Service Center and San Diego West). In Arkansas, the two offices are in different cities (Little Rock and Pine Bluff). In Baltimore and Virginia, the presence of many offices with small sample sizes in each made it infeasible to perform a one-against-one cross-site analysis.

[11] The further apart in time the cohorts are, the greater will be the risk of confounding environmental events, but the longer will be the follow-up for the early cohort before program start-up. Once the program begins, either follow-up for the early cohort must end or the program must delay working with any early cohort members who still remain eligible for the program. In practice, a systematic and gradual phase-in of enrollment and participation, beginning with the late cohort, could leave an early cohort untouched for a year or longer.

become the comparison group. We compare these nonexperimental cross-cohort estimates with the experimental estimates computed from the program and control groups of the late cohort at each site.

III. The Econometric Specification

The nonexperimental estimates presented in this paper are derived from a linear regression model that adjusts for differences in the characteristics of sample members observed just prior to their entry into the sample.[12] Several of these characteristics were also used to create the statistically matched cross-state comparison samples.[13] The same regression specification is applied to produce experimental estimates of program effects. Although regression adjustment is not strictly required for experimental estimates, it improves statistical precision somewhat by reducing residual variation and correcting for minor chance differences that exist between the program and control groups after randomization.

The dependent variables in the regression model are the short-term and long-term employment measures defined earlier. The independent variables include lagged employment (measured in the first quarter prior to random assignment),[14] a vector of demographic characteristics of the sample member prior to random assignment, and a dummy variable equal to 1 if the sample member is in a program group and 0 if the sample member is in a comparison group (nonexperimental estimate) or a control group (experimental estimate).

The coefficient of the program-group dummy is the estimate of the program's effect. For a nonexperimental estimate, it may or may not be an unbiased estimate of the true program effect, depending on whether the error term is uncorrelated or correlated with the program-group dummy. Only if the preprogram values of the outcome variable and demographic characteristics remove any correlation with the error term will the nonexperimental estimate be unbiased. Differences in motivation, labor markets, events, or other factors not accounted for in the set of demographic variables but present in the program group dummy and the error term will induce bias. The experimental estimate of the program effect is unbiased because randomization (assuming it is implemented properly) ensures that the error term is uncorrelated with the program-group dummy.

In experimental designs, "internal" validity is based on the randomized assignment of sample members to program and control groups. Ensuring the validity of experiments generally rests on careful implementation of the randomization process and continued monitoring of the different treatment of the research groups during the follow-up period. Internal validity is not a subject for statistical testing, although it is often "confirmed" after the fact by verifying that observable pre-random-assignment demographics are similar across research groups.

In nonexperimental designs, internal validity necessarily rests on the validity of the econometric specification and the statistical methods used to estimate it. As a consequence, it has been argued that statistically testing the econometric model should be an integral part of any nonexperimental estimation procedure.[15] If a model fails a speci-

[12] Heckman and Hotz (1989) found that the linear regression model produced the best nonexperimental estimates of training effects on earnings for AFDC recipients.

[13] In some sense, the statistical-matching technique serves the same function as the regression adjustment using observed preprogram personal characteristics. Indeed, as we will see, the two cross-state methods perform about the same despite the fact that the matched comparison samples bear closer resemblance to the program samples on the basis of these preprogram characteristics than the unmatched comparison samples.

[14] Inclusion of additional periods of preprogram employment data has little impact on the nonexperimental estimates.

[15] An argument for specification testing and a variant of the test we employ appeared first in connection with employment and training program evaluations in Ashenfelter (1978). The rationale was developed further, along with additional tests, by Heckman and Hotz (1989).

fication test, then it may be declared the wrong model, and we should not accept the estimates of program effects it produces. If a model passes, then we have grounds for accepting the validity of the estimated program effects. According to this argument, it is quite possible that the recent pessimistic appraisals of nonexperimental methods for estimating program effects have resulted largely from failure to apply specification tests (see Heckman and Hotz, 1989).

One kind of specification test is concerned with whether the econometric model will account for all differences in outcomes between program and comparison groups except those induced by the program. Operationally, such a test generally looks at whether the model "correctly" predicts no differences in outcomes between the program and comparison groups during the period *before the program group enters the program* (i.e., before they receive any "treatment"). In this study, we test whether the estimated program effect is different from zero when the model is applied to the preprogram period. We do this by estimating a linear regression model similar to the one used to estimate the experimental and nonexperimental program effects, except that the dependent variable is the individual's employment status in the first quarter *prior* to entry into the sample (i.e., for program group members, the first quarter before enrollment in the program). All independent variables are the same except that the lagged value of the dependent variable is measured three quarters prior to the dependent variable.[16] The specification test is based on the coefficient of the program-group dummy. If this coefficient is significantly different from zero, then the nonexperimental estimator fails the test; otherwise it passes.[17] We use 10 percent as the critical level for determining statistical significance.[18]

This test has the advantage of being readily understandable and quite easy to apply in practice. Specification testing does have a serious shortcoming, however. The fit of the model for the preprogram period has no necessary logical connection with the validity of the model for the follow-up period. Becoming eligible for an employment and training program frequently occurs at a life transition, such as dissolution of a family group, loss of support, a long stint of joblessness, entry of a single mother's youngest child into school, and the like. An empirical specification that adequately describes behavior before such a transition, and therefore passes the specification test, may not provide an adequate control function for outcomes following the transition. Or the procedure may work well with short-term follow-up but not with long-term follow-up. For this reason, we examine outcomes measured at short and long lengths of time after the program begins.

If we find, through studies like the present one, that the specification test generally leads to nonexperimental estimates that are similar to experimental estimates for outcomes measured at different points in time after enrollment, then we may have confidence in using the procedure, but only through repeated validation—not because any theory indicates that the test necessarily must work. It is quite possible that empirical studies will show certain kinds of nonexperimental results to be generally invalid even when they pass a specification test. Moreover, the test may work well only for some groups or under some conditions.

[16] Ideally, more preprogram data should be used in specification testing. For example, Heckman and Hotz (1989) and others have argued for using a long preprogram period in the specification tests. Our ability to do testing on this scale is limited by the amount of preprogram data we have (one year).

[17] A more conservative test would posit that all of the coefficients, not only the program group dummy, are the same in preprogram and follow-up equations. Clearly, a number of more complicated specifications and more complex testing procedures than those we utilize could be adopted.

[18] We also tested a larger critical level and found that the results were insensitive to the level chosen.

IV. Empirical Results

A large number of nonexperimental estimates were produced for comparison with the experimental estimates. For the cross-state comparisons, there were 120 pairs of experimental and nonexperimental estimates (96 unmatched and 24 matched).[19] For the within-state comparisons, there were 40 pairs of experimental and nonexperimental estimates (16 across sites and 24 across cohorts).[20] To facilitate interpretation of our results, we first present detailed estimates for a selected subset of these experimental and nonexperimental pairs. We then present a summary of the results for all 160 pairs of estimates.

Table 1 presents a number of selected experimental and nonexperimental estimates of longer-term employment effects.[21] Each cell of the table is devoted to one estimate and shows the following information (from the top down): (i) the estimated program effect, (ii) the standard error of the estimated program effect, (iii) the probability value (p value) of the specification test, and (iv) the sample size. A specification test fails for a p value of less than 0.10. Such a result indicates that the program-group dummy is different from zero when the dependent variable is employment in the first quarter prior to program entry. This is taken to mean that the comparison group does not provide a valid representation of the preprogram behavior of the particular program group in question and that the regression adjustment has not succeeded in making the comparison valid.

The first four columns of Table 1 provide information about the cross-state estimates. In the first two columns, two sets of experimental estimates are given, unadjusted and regression-adjusted. Comparing these two provides a check on the implementation of the experimental design. If randomization is properly implemented, then in most cases regression adjustment will affect only the standard errors and not the numerical estimate of the program effect.[22] In Table 1, the unadjusted and regression-adjusted experimental estimates of the program effect are quite close to one other for all four programs, and all but one of the experimental comparisons pass the specification test. The one test failure occurs in Baltimore, where the unadjusted experimental estimate is not statistically significant and the p value of the specification test is less than 0.10.[23] This is the only experimental comparison to fail the specification test in Table 1. Regression adjustment for observed baseline characteristics in that experiment slightly increases the experimental estimate of program effect and makes it statistically significant. Regression adjustment also changes

[19] In the unmatched analyses, equations were estimated for each of the four states times four comparison groups per state (each of the other states plus the combined other three states) times two dependent variables (short-term and long-term employment) times three sample subgroups (short-term recipients of welfare, long-term recipients of welfare, and both groups combined). In the matched analyses, equations were estimated for each of the four states times one comparison group per state times two dependent variables times three sample subgroups.

[20] In the analyses across sites, equations were estimated for two welfare offices times two dependent variables times three sample subgroups in San Diego plus two welfare offices times two dependent variables in Arkansas (sample subgroup analyses were not possible in Arkansas because sample sizes were too small). In the analyses across cohorts, equations were estimated for four states times two dependent variables times three sample subgroups.

[21] For the cross-state/unmatched column, the pair with the largest comparison group is shown for each state (the pair in which controls from the other three states combined are used). For the cross-state/matched column, all pairs are shown. For the cross-site column, one pair for each state was selected at random. For the cross-cohort column, all pairs are shown.

[22] True randomization will occasionally create program and control groups that, by chance, have background characteristics that differ to a statistically significant degree. A control group created by a well-implemented random assignment procedure, therefore, will occasionally fail the specification test.

[23] The coefficient on the program-group dummy in the baseline period is significantly negative, suggesting that the unadjusted experimental estimate is biased downward.

TABLE 1—SELECTED ESTIMATES OF PROGRAM EFFECT ON FRACTION EMPLOYED 6–9 QUARTERS AFTER BASELINE

	Cross-state estimates				Within-state estimates			
	Experimental estimates		Nonexperimental estimates		Across site		Across cohort	
Program	Unadjusted	Adjusted	Unmatched	Matched	Experimental	Nonexperimental	Experimental	Nonexperimental
Arkansas, WORK	0.050†	0.057*	−0.139**	−0.152**	0.064*	0.101**	0.009	0.067†
	(0.027)	(0.025)	(0.024)	(0.028)	(0.032)	(0.038)	(0.034)	(0.035)
	[0.119]	[0.230]	[0.000]	[0.896]	[0.523]	[0.849]	[0.956]	[0.600]
	N = 1,127	N = 1,127	N = 4,593	N = 1,120	N = 692	N = 543	N = 605	N = 570
Baltimore, OPTIONS	0.030	0.041*	0.132**	0.127**	—	—	0.041†	0.081**
	(0.019)	(0.018)	(0.015)	(0.018)			(0.023)	(0.026)
	[0.064]	[0.191]	[0.165]	[0.681]			[0.140]	[0.280]
	N = 2,757	N = 2,757	N = 4,567	N = 2,724			N = 1,725	N = 1,380
San Diego, SWIM	0.091**	0.090**	0.020	0.034*	0.078**	0.123**	0.083**	0.067**
	(0.017)	(0.017)	(0.016)	(0.017)	(0.024)	(0.024)	(0.026)	(0.024)
	[0.773]	[0.951]	[0.883]	[0.285]	[0.678]	[0.201]	[0.709]	[0.303]
	N = 3,211	N = 3,211	N = 4,597	N = 3,208	N = 1,603	N = 1,620	N = 1,453	N = 1,620
Virginia, ESP	0.050**	0.060**	0.115**	0.127**	—	—	0.058*	0.090**
	(0.019)	(0.018)	(0.013)	(0.014)			(0.024)	(0.026)
	[0.440]	[0.891]	[0.017]	[0.230]			[0.492]	[0.000]
	N = 3,150	N = 3,150	N = 5,688	N = 4,238[a]			N = 1,673	N = 1,620

Notes: From top to bottom, the entries in each cell are (i) estimate of program effect, (ii) standard error of estimated program effect (in parentheses), (iii) p value of specification test (in brackets), and (iv) sample size. We tested the difference between each nonexperimental estimate and the first experimental estimate to the left of it in the table for statistical significance. Such differences were found to be statistically significant for all unmatched and matched cross-state estimates and for none of the cross-site or cross-cohort estimates.

[a] Virginia, unlike the other experiments, had a program-group:control-group size ratio of 2:1; matching therefore yields a sample larger than the original evaluation sample.

†Statistically significant at the 10-percent level.

*Statistically significant at the 5-percent level.

**Statistically significant at the 1-percent level.

the specification-test statistic to a "pass," indicating that regression has controlled for the differences in baseline characteristics. It is of interest to note that the failed (unadjusted) estimate of program effect is quite close to the passed (regression-adjusted) estimate of program effect. That is, the test does not seem to be serving its function of eliminating the inaccurate estimates while keeping those that are close to the best estimate.

Table 1 next shows unmatched and matched nonexperimental cross-state estimates. These estimates use the program group in the state and a comparison group created from the control groups in the other three states. The unmatched and matched cross-state nonexperimental estimates are similar to each other but differ considerably from the regression-adjusted experimental estimates (in the second column). For Arkansas, both of the nonexperimental estimates are of the wrong sign and both differ by about 0.20. In the other states, the signs of the experimental and nonexperimental estimates are the same, but their magnitudes are never closer than about 0.05. Statistical inferences differ (i.e., only one estimate of program effect in a pair is statistically significant at the 10-percent level or both are statistically significant but with opposite signs) in three of the eight pairs of experimental and nonexperimental estimates. All of the differences between the cross-state nonexperimental estimates (both unmatched and matched) and their corresponding regression-adjusted experimental estimates are statistically significant.

These results suggest that using individuals in one state as a comparison group for individuals in another state can lead to quite inaccurate estimates of the size of a program effect, even if the two groups are matched statistically according to a set of baseline characteristics. Moreover, the specification test has difficulty discriminating between accurate and inaccurate nonexperimental estimates. Only two of the unmatched nonexperimental estimates and none of the matched nonexperimental estimates are rejected by the specification test. Moreover, in both Arkansas and Virginia, the specification test rejects the unmatched estimate but accepts a matched estimate that is even further from the experimental estimate.

The situation improves somewhat when switching from cross-state to within-state comparisons, as shown in the right side of Table 1 (last four columns). First, an experimental estimate is shown for one local office in Arkansas and one in San Diego. Next is shown a nonexperimental estimate produced by using as a comparison group the control group in the other local office in the evaluation sample.[24] For both pairs, the experimental and nonexperimental estimates are within 0.05 of each other, the differences are not statistically significant, and the statistical inferences are the same. Both nonexperimental estimates pass the specification test.

The last two columns of Table 1 show the second kind of within-state estimates, namely, the cross-cohort estimates. The next-to-last column shows the experimental estimates of the program effect created by using program and control group members randomly assigned in a late cohort (i.e., late in the period of random assignment). The last column shows nonexperimental estimates using the early-cohort control group as a comparison group for the late-cohort program group in the same locality. For three of the four pairs, the differences between experimental and nonexperimental estimates are less than 0.05, and the statistical inferences are the same. For none of the cohort pairs are the experimental and nonexperimental estimates significantly different. The specification test unfortunately rejects one of these nonexperimental estimates (Virginia). It also fails to reject the Arkansas nonexperimental estimate, which differs more from its paired experimental estimate than any of the other cross-cohort estimates and also yields a different statistical inference.

Table 2 summarizes the results for all 160 pairs of experimental and nonexperimental estimates (including those in Table 1). As in Table 1, each column represents a different comparison-group specification. There are three panels in the table: the top one for all pairs of experimental and nonexperimental estimates, the middle one for pairs in which the nonexperimental comparison passes the specification test, and the bottom one for pairs in which the nonexperimental comparison fails the specification test. The first row of each panel gives the number of pairs of experimental and nonexperimental estimates in the panel. The second row of each panel gives the mean of the experimental estimates in the panel. The third row in each panel gives the mean of the absolute difference between the experimental and nonexperimental estimates in each pair, our principal summary measure.[25] The fourth row in each panel gives the percentage of pairs for which the experimental and nonexperimental estimates of the program effect yield different statistical inferences (based on the 10-percent level of significance). Finally, the fifth row in each panel gives the

[24] In San Diego, Service Center is the program office and San Diego West is the comparison office. In Arkansas, Little Rock is the program site and Pine Bluff is the comparison site.

[25] The average *absolute* difference is presented, but the average difference is not. The latter is not meaningful for the cross-state estimates (particularly the unmatched sample) or for the cross-site estimates, because the control group for each state or site serves as a comparison group in a nonexperimental estimate, forcing the average difference to be close to zero by definition.

TABLE 2—SUMMARY OF EXPERIMENTAL AND NONEXPERIMENTAL ESTIMATES OF PROGRAM EFFECT ON FRACTION EMPLOYED

	Comparison group specification			
	Cross-state estimates		Within-state estimates	
Statistic	Unmatched	Matched	Across site	Across cohort
All Pairs of Experimental and Nonexperimental Estimates:				
Number of pairs	96	24	16	24
Mean experimental estimate	0.056	0.056	0.069	0.045
Mean absolute experimental – nonexperimental difference	0.090	0.080	0.044	0.034
Percentage with different inference	47	38	13	29
Percentage with statistically significant difference (10-percent level)	70	67	31	4
Pairs That Pass Specification Test:				
Number of pairs	32	18	14	16
Mean experimental estimate	0.052	0.048	0.057	0.046
Mean absolute experimental – nonexperimental difference	0.059	0.080	0.042	0.030
Percentage with different inference	41	39	14	19
Percentage with statistically significant difference (10-percent level)	56	72	29	0
Pairs That Fail Specification Test:				
Number of pairs	64	6	2	8
Mean experimental estimate	0.057	0.077	0.157	0.043
Mean absolute experimental – nonexperimental difference	0.105	0.080	0.063	0.042
Percentage with different inference	50	33	0	50
Percentage with statistically significant difference (10-percent level)	77	50	50	13

percentage of pairs for which the difference between experimental and nonexperimental estimates is statistically significant at the 10-percent level.[26]

The top panel in Table 2 displays results for all pairs of experimental and nonexperimental estimates, ignoring whether the nonexperimental part of the pair passes or fails the specification test. Looking first at the cross-state estimates that do not utilize statistical matching (the first column of results), it is seen that the mean absolute difference

[26] It is important to recognize that the sample sizes vary across the comparison-group specifications. In particular, sample sizes are smaller in the within-state specifications than in the cross-state specifications. In comparing the results across the different comparison-group specifications, potential biases related to sample size should be kept in mind. For the first criterion used to judge the accuracy of the nonexperimental estimates, mean absolute difference, decreasing sample size will tend to produce worse results, assuming that there is no change in any real underlying bias. The other two criteria may tend to show (falsely) improvement as sample size decreases. For a given size of the program effect, "different inference" may occur less often in smaller samples because it is more likely that both experimental and nonexperimental estimates will not be statistically significant. Similarly, fewer statistically significant differences may occur in smaller samples, since *any* bias in the nonexperimental estimate, no matter how slight, will yield "difference statistically significant" if only the samples are large enough.

is large, much larger than the mean experimental estimate itself. Nearly half of the nonexperimental estimates (47 percent) lead to different statistical inferences from those for the corresponding experimental estimates. An even larger share of the pairwise differences (70 percent) are statistically significant. About a third of the pairs differ by more than 0.100 (not shown in the table), which indicates a substantial discrepancy, given that the average of the experimental estimates is +0.056. Only a third differ by less than 0.050 (not shown).

Statistical matching improves the accuracy of the cross-state nonexperimental estimates, but not by very much. The mean absolute experimental–nonexperimental difference falls to 0.080 (about an 11-percent drop), but it is still larger than the mean experimental estimate. The percentage of the estimates producing a different inference falls to 38 percent (about a 19-percent drop), and the percentage with a statistically significant difference falls to 67 percent (about a 4-percent drop). Thus, even though the program and comparison groups are more closely matched on their observed baseline personal characteristics, the greater similarity translates into only a modest rather than a substantial improvement in the accuracy of the nonexperimental estimates.

There are several reasons why statistical matching does not more dramatically improve the accuracy of the nonexperimental estimates. First, matching utilizes only observed characteristics, measured at baseline. If there are substantial unobserved differences between the program and comparison groups and if these differences influence employment decisions independently of observed characteristics, then matching is likely to have little effect on the comparability of the two groups. Such unobserved differences may be environment-related (e.g., differences in economic conditions or local institutional structure) or they may be individual-related (e.g., differences in community preferences for work). Second, the particular matching procedure used here may not have given enough weight to the most important baseline characteristics. Some other matching procedure might prove superior. Third, the matching variables are also included as regressors in the models of program effects for both the matched and unmatched samples. In the unmatched samples, these regressors perform much the same function as matching (i.e., they adjust the outcome variable for differences in the observed characteristics). In the matched samples, these characteristics are used twice: first in creating the matched sample, and then again in adjusting the outcome variable in the regression model. In fact, when the nonexperimental models are estimated without using these characteristics as regressors, the matched-sample results hardly change at all, whereas the unmatched-sample results change more and differ more from those of their experimental counterparts. Finally, it should be noted that matching on preprogram characteristics does not logically imply that the model relating those characteristics to subsequent behavior must necessarily be the same for the matched comparison group and the program group in the absence of the program. In fact, the crux of the nonexperimental evaluation problem is that individuals with identical measured baseline characteristics can exhibit divergent follow-up outcomes.

Looking next at the last two columns of Table 2, it is evident that the within-state comparisons perform considerably better than the cross-state comparisons, but inaccuracies still remain. Mean absolute differences are much smaller and are now less than the mean experimental estimates. The decrease in mean absolute differences is particularly notable because the smaller sample sizes would tend to produce increases in this measure (see footnote 26). Mean absolute differences, however, are still large enough to be worrisome (64 percent of the mean experimental estimate in the cross-site specification and 76 percent of the mean experimental estimate in the cross-cohort specification). Note that only 13 percent of the cross-site pairs and only 29 percent of the cross-cohort pairs produce different inferences. Also note that only 31 percent of the cross-site pairs and 4 percent of the cross-cohort pairs produce statistically significant differences. These latter two criteria, however, might tend to indicate

fewer significant differences anyway as the cross-state samples are divided into the smaller within-state subsamples (see footnote 26).

In the middle and bottom panels of Table 2, the pairs are grouped separately for nonexperimental comparisons that pass and fail the specification test, respectively. If the specification test is useful, most of the results in the middle panel should be accurate. It would also be desirable, but not strictly necessary, if the bottom panel contained mostly inaccurate estimates, since that would mean that the test would waste fewer accurate estimates.

The results indicate that the specification test does improve the accuracy of the nonexperimental estimates. In most columns of the table, the mean absolute difference is lower among the group of pairs that pass the specification test than it is among all pairs. However, the ability of the specification test to discriminate between accurate and inaccurate estimates is not great in these samples, and the improvement from using the test is often marginal.

For the cross-state/unmatched comparisons, two-thirds of the nonexperimental estimates fail the specification test. Such a high failure rate is not necessarily bad. In fact, we would hope for a large proportion of test failures, since so many of the results are so far off the mark. As it turns out, weeding out the test failures does improve results somewhat. The mean absolute difference drops by a third (from 0.090 to 0.059), and there are smaller proportions with different inferences or statistically significant differences.

The failed cross-state/unmatched estimates are mostly for pairs with statistically significant experimental–nonexperimental differences (77 percent) and, to a lesser extent, different statistical inference (50 percent). About half the failed pairs differ by more than 0.100 (not shown in the table). Indeed, the specification test seems most effective in eliminating those "outlier" estimates (i.e., nonexperimental estimates differing by more than 0.100 from their corresponding experimental estimates). In all, there are 54 cross-state/unmatched pairs with differences of more than 0.100, and 47 (more than 85 percent) were identified by failing the specification test.

Notwithstanding, many of the cross-state/unmatched nonexperimental estimates that pass the specification test still differ substantially from their respective experimental estimates: 41 percent of the "passes" yield different statistical inferences, and 56 percent of the differences are statistically significant.

The cross-state/matched estimates do not show improvement from the specification test, as did the unmatched estimates. Only one-quarter of the matched estimates fail the test. The passed estimates have the same mean absolute difference as the set of all estimates in the column; there is an absence of the kind of improvement observed for the unmatched estimates. More of the passed estimates have a different inference, and more have a statistically significant difference.

The within-state estimates show only marginally greater benefit from the specification test than the cross-state/matched estimates. The great majority of the within-state estimates pass the specification test. There is some improvement for cross-site and cross-cohort comparisons, but the set of passes is not markedly better than the set of all untested estimates.

V. Conclusions

Many social programs are evaluated using nonexperimental econometric methods that compare the behavior of individuals exposed to the program with the behavior of individuals who are not. In this paper we have followed previous studies by using experimental data to assess the relative efficacy of certain types of nonexperimental procedures that are frequently used to evaluate social programs. All of these procedures face the common problem of selecting an appropriate comparison group. We also examined two statistical techniques for improving accuracy: statistical matching and specification testing. The procedures we have utilized in this paper were not meant to represent the full catalogue of applicable

nonexperimental comparison methods, matching techniques, or specification tests.

Our first set of nonexperimental estimates utilized comparison samples drawn in different states from the program samples. Nonexperimental evaluations often adopt this method. The resulting nonexperimental estimates were usually quite different from the experimental estimates derived from the same data. This is not surprising, since program and comparison samples for the cross-state procedures were far apart in space and time, and environmental factors could differ considerably. More importantly, however, we found that specification testing and statistical matching procedures did not appear to make major improvements in the cross-state estimates. The specification test did reject many of the nonexperimental estimates, but those that remained were only somewhat better, not markedly better, than the original untested set of estimates. Statistical matching also produced only a modest improvement in the accuracy of the nonexperimental estimates.

Our second set of nonexperimental estimates utilized comparison samples drawn in the same state as the program sample. This approach produced better results than the cross-state procedures. The average discrepancy between experimental and nonexperimental estimates was smaller, although important differences still remained. Furthermore, applying a specification test did not improve these within-state estimates further.[27] Overall, the specification test was more effective in eliminating wildly inaccurate "outlier" estimates than in pinpointing the most accurate nonexperimental estimates. Making the comparison samples closer in time and space by using cross-site and cross-cohort comparisons produced more improvement than specification testing or statistical matching.

The results of our study illustrate the risks involved in comparing the behavior of individuals residing in two different geographic areas. Comparisons across state lines are particularly problematic. Our findings illustrate that estimates of program effects from cross-state comparisons can be quite far from the true effects, even when samples are drawn (as ours were) with the same sample intake procedures and from target populations defined with the same objective characteristics. This is an important lesson, given that studies often rely on cross-state comparisons. Our results suggest that statistical matching or a specification test alone will be unable to reduce markedly the uncertainty surrounding that kind of nonexperimental estimate. Our research indicates that, at a minimum, such studies must demonstrate the similarity of local conditions as a prerequisite to establishing the validity of the comparison. Whether similarity of observable local conditions is sufficient for a valid comparison of two areas remains an open question, however. When we switched the comparison from across states to within a state we did note some improvement, but inaccuracies still remained. Additional research should ask whether these inaccuracies can be reduced further by larger sample sizes, by using particular local variables to identify a suitable comparison area (and by verifying later that the area remained suitable throughout the study period), or by pooling estimates of program effects from a number of independent comparisons or comparison studies.

[27]Additional analysis not reported in this paper indicates that the findings are not sensitive to the choice of dependent variables, to the subgroup analyzed, or to the size of the critical region of the specification test.

REFERENCES

Ashenfelter, Orley. "Estimating the Effect of Training Programs on Earnings." *Review of Economics and Statistics*, February 1978, *60*(1), pp. 47–57.

Ashenfelter, Orley and Card, David. "Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs." *Review of Economics and Statistics*, November 1985, *67*(4), pp. 648–60.

Brown, Randall; Burghardt, John; Cavin, Edward; Long, David; Mallar, Charles; Maynard, Rebecca; Metcalf, Charles; Thornton, Craig and Whitebread, Christine. *Final report: Employment opportunity pilot project:*

Analysis of program impacts. Princeton, NJ: Mathematica Policy Research, February 1983.

Decker, Paul T. *Estimating the effects of the REACH program on AFDC receipt*. Princeton, NJ: Mathematica Policy Research, August 1991.

Dickinson, Katherine P.; Johnson, Terry R. and West, Richard W. "An Analysis of the Impact of CETA Programs on Participants' Earnings." Final report prepared for the U.S. Department of Labor, Employment and Training Administration, SRI International, Menlo Park, CA, November 1984.

_____. "An Analysis of the Sensitivity of Quasi-Experimental Net Impact Estimates of CETA Programs." *Evaluation Review*, August 1987, *11*(4), pp. 452–72.

Farkas, George; Smith, David and Stromsdorfer, Ernst. "The Youth Entitlement Demonstration: Subsidized Employment with a Schooling Requirement." *Journal of Human Resources*, Fall 1983, *18*(4), pp. 557–73.

Farkas, George; Olsen, Randall; Stromsdorfer, Ernst W.; Sharpe, Linda C.; Skidmore, Felicity; Smith, D. Alton and Merrill, Sally. *Post-program impacts of the youth incentive entitlement pilot projects*. New York: Manpower Demonstration Research Corporation, June 1984.

Fraker, Thomas and Maynard, Rebecca. "Evaluating Comparison Group Designs with Employment-Related Programs." *Journal of Human Resources*, Spring 1987, *22*(2), pp. 194–227.

Friedlander, Daniel and Gueron, Judith M. "Are High-Cost Services More Effective than Low-Cost Services?" in Charles F. Manski and Irwin Garfinkel, eds., *Evaluating welfare and training programs*. Cambridge, MA: Harvard University Press, 1992, pp. 143–98.

Friedlander, Daniel; Hoerz, Gregory; Long, David and Quint, Janet. *Maryland: Final report on the employment initiatives evaluation*. New York: Manpower Demonstration Research Corporation, December 1985a.

Friedlander, Daniel; Hoerz, Gregory; Quint, Janet and Riccio, James. *Arkansas: Final report on the work program in two counties*. New York: Manpower Demonstration Research Corporation, September 1985b.

Garfinkel, Irwin; McLanahan, Sara and Robins, Philip K. *Child support assurance: Design issues, expected impacts, and political barriers as seen from Wisconsin*. Washington, DC: Urban Institute Press, 1992.

Greenberg, David H. and Wiseman, Michael. "What Did the OBRA Demonstrations Do?" in Charles F. Manski and Irwin Garfinkel, eds., *Evaluating welfare and training programs*. Cambridge, MA: Harvard University Press, 1992, pp. 25–75.

Gueron, Judith M. "Work and Welfare: Lessons on Employment Programs." *Journal of Economic Perspectives*, January 1990, *4*(1), pp. 79–98.

Gueron, Judith M. and Pauly, Edward. *From welfare to work*. New York: Russell Sage Foundation, 1991.

Hamilton, Gayle and Friedlander, Daniel. *Final report on the saturation work initiative model in San Diego*. New York: Manpower Demonstration Research Corporation, November 1989.

Hand, D. J. *Discrimination and classification*. New York: Wiley, 1981.

Heckman, James J. and Hotz, V. Joseph. "Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training." *Journal of the American Statistical Association*, December 1989, *84*(408), pp. 862–74.

Hollister, Robinson G. and Hill, Jennifer. "Problems in the Evaluation of Community-Wide Initiatives," in James P. Connell, Anne C. Kubisch, Lisbeth B. Schorr, and Carol H. Weiss, eds., *New approaches to evaluating community initiatives: Concepts, methods, and contexts*. Washington, DC: Aspen Institute, 1995, pp. 127–72.

LaLonde, Robert. "Evaluating the Econometric Evaluations of Training Programs with Experimental Data." *American Economic Review*, September 1986, *76*(4), pp. 604–20.

LaLonde, Robert and Maynard, Rebecca. "How Precise Are Evaluations of Employment and Training Programs: Evidence from a Field Experiment." *Evaluation Review*,

August 1987, *11*(4), pp. 428–51.

Long, David A. "Cleveland Analysis Plan." Mimeo, Manpower Demonstration Research Corporation, New York, December 1991.

Long, Sharon K. and Wissoker, Douglas A. *Net impacts of the Washington State Family Independence Program (FIP): The first two years.* Washington, DC: Urban Institute, June 1992.

Moffitt, Robert. "Program Evaluation with Nonexperimental Data." *Evaluation Review*, June 1991, *15*(3), pp. 291–314.

Polit, Denise; Kahn, Janet and Stevens, David. *Final impacts from project redirection.* New York: Manpower Demonstration Research Corporation, 1985.

Riccio, James; Cave, George; Freedman, Stephen and Price, Marilyn. *Final report on the Virginia Employment Services Program.* New York: Manpower Demonstration Research Corporation, August 1986.

Schiller, Bradley R. and Brasher, C. Nielsen. "Effects of Workfare Saturation on AFDC Caseloads." *Contemporary Policy Issues*, April 1993, *11*(2), pp. 39–49.