Evaluating the Econometric Evaluations of Training Programs with Experiment Data

Article *in* American Economic Review · February 1986

Source: RePEc

CITATIONS
1,520

READS
2,668

1 author:

Robert J. Lalonde
University of Chicago
51 PUBLICATIONS **9,331** CITATIONS

SEE PROFILE

Evaluating the Econometric Evaluations of Training Programs with Experimental Data

Robert J. LaLonde

The American Economic Review, Vol. 76, No. 4 (Sep., 1986), 604-620.

Stable URL:
http://links.jstor.org/sici?sici=0002-8282%28198609%2976%3A4%3C604%3AETEEOT%3E2.0.CO%3B2-P

Evaluating the Econometric Evaluations of Training Programs with Experimental Data

By ROBERT J. LALONDE*

This paper compares the effect on trainee earnings of an employment program that was run as a field experiment where participants were randomly assigned to treatment and control groups with the estimates that would have been produced by an econometrician. This comparison shows that many of the econometric procedures do not replicate the experimentally determined results, and it suggests that researchers should be aware of the potential for specification errors in other nonexperimental evaluations.

Econometricians intend their empirical studies to reproduce the results of experiments that use random assignment without incurring their costs. One way, then, to evaluate econometric methods is to compare them against experimentally determined results.

This paper undertakes such a comparison and suggests the means by which econometric analyses of employment and training programs may be evaluated. The paper compares the results from a field experiment, where individuals were randomly assigned to participate in a training program, against the array of estimates that an econometrician without experimental data might have produced. It examines the results likely to be reported by an econometrician using nonexperimental data and the most modern techniques, and following the recent prescriptions of Edward Leamer (1983) and David Hendry (1980), tests the extent to which the results are sensitive to alternative econometric specifications.[1] The goal is to appraise the likely ability of several econometric methods to accurately assess the economic benefits of employment and training programs.[2]

Section I describes the field experiment and presents simple estimates of the program effect using the experimental data. Sections II and III describe how econometricians evaluate employment and training programs, and compares the nonexperimental estimates using these methods to the experimental results presented in Section I. Section II presents one-step econometric estimates of the program's impact, while more complex two-step econometric estimates are presented in Section III. The re-

*Graduate School of Business, University of Chicago, 1101 East 58th Street, Chicago, IL 60637. This paper uses public data files from the National Supported Work Demonstration. These data were provided by the Inter-University Consortium for Political and Social Research. I have benefited from discussions with Mariam Akin, Orley Ashenfelter, James Brown, David Card, Judith Gueron, John Papandreou, Robert Willig, and the participants of workshops at the universities of Chicago, Cornell, Iowa, Princeton, and MIT.

[1] These papers depict a more general crisis of confidence in empirical research. Leamer (1983) argues that any solution to this crisis must divert applied econometricians from "the traditional task of identifying unique inferences implied by a specific model to the task of determining the range of inferences generated by a range of models." Other examples of this literature are Leamer (1985), Leamer and Herman Leonard (1983), and Michael McAleer, Adrian Pagan, and Paul Volker (1985).

[2] Examples of nonexperimental program evaluations are Orley Ashenfelter (1978), Ashenfelter and David Card (1985), Laurie Bassi (1983a,b; 1984), Thomas Cooley, Thomas McGuire, and Edward Prescott (1979), Katherine Dickinson, Terry Johnson, and Richard West (1984), Nicholas Kiefer (1979a,b), and Charles Mallar (1978).

sults of this study are summarized in the final section.

I. The Experimental Estimates

The National Supported Work Demonstration (NSW) was a temporary employment program designed to help disadvantaged workers lacking basic job skills move into the labor market by giving them work experience and counseling in a sheltered environment. Unlike other federally sponsored employment and training programs, the NSW program assigned qualified applicants to training positions randomly. Those assigned to the treatment group received all the benefits of the NSW program, while those assigned to the control group were left to fend for themselves.[3]

During the mid-1970s, the Manpower Demonstration Research Corporation (MDRC) operated the NSW program in ten sites across the United States. The MDRC admitted into the program AFDC women, ex-drug addicts, ex-criminal offenders, and high school dropouts of both sexes.[4] For those assigned to the treatment group, the program guaranteed a job for 9 to 18 months, depending on the target group and site. The treatment group was divided into crews of three to five participants who worked together and met frequently with an NSW counselor to discuss grievances and performance. The NSW program paid the treatment group members for their work. The wage schedule offered the trainees lower wage rates than they would have received on a regular job, but allowed their earnings to increase for satisfactory performance and attendance. The trainees could stay on their supported work jobs until their terms in the program expired and they were forced to find regular employment.

Although these general guidelines were followed at each site, the agencies that operated the experiment at the local level provided the treatment group members with different work experiences. The type of work even varied within sites. For example, some of the trainees in Hartford worked at a gas station, while others worked at a printing shop.[5] In particular, male and female participants frequently performed different sorts of work. The female participants usually worked in service occupations, whereas the male participants tended to work in construction occupations. Consequently, the program costs varied across the sites and target groups. The program cost $9,100 per AFDC participant and approximately $6,800 for the other target groups' trainees.[6]

The MDRC collected earnings and demographic data from both the treatment and the control group members at the baseline (when MDRC randomly assigned the participants) and every nine months thereafter, conducting up to four post-baseline inter-

[3] Findings from the NSW are summarized in several reports and publications. For a quick summary of the program design and results, see Manpower Demonstration Research Corporation (1983). For more detailed discussions see Dickinson and Rebecca Maynard (1981); Peter Kemper, David Long, and Craig Thornton (1981); Stanley Masters and Maynard (1981); Maynard (1980); and Irving Piliavin and Rosemary Gartner (1981).

[4] The experimental sample included 6,616 treatment and control group members from Atlanta, Chicago, Hartford, Jersey City, Newark, New York, Oakland, Philadelphia, San Francisco, and Wisconsin. Qualified AFDC applicants were women who (*i*) had to be currently unemployed, (*ii*) had spent no more than 3 months in a job in the previous 6 months, (*iii*) had no children less than six years old, and (*iv*) had received AFDC payments for 30 of the previous 36 months. The admission requirements for the other participants differed slightly from those of the AFDC applicants. For a more detailed discussion of these prerequisities, see MDRC.

[5] Kemper and Long present a list of NSW projects and customers (1981, Table IV.4, pp. 65–66). The trainees produced goods and services for organizations in the public (42 percent of program hours), nonprofit (29 percent of program hours), and private sectors.

[6] The cost per training participant is the sum of program input costs, site overhead costs, central administrative costs, and child care costs minus the value of the program's output. These costs are in 1982 dollars. If the trainees' subsidized wages and fringe benefits are viewed as a transfer instead of a cost, the program costs per participant are $3,100 for the AFDC trainees and $2,700 for the other trainees. For a more detailed discussion of program costs and benefits, see Kemper, Long, and Thornton.

TABLE 1—THE SAMPLE MEANS AND STANDARD DEVIATIONS OF PRE-TRAINING EARNINGS AND OTHER CHARACTERISTICS FOR THE NSW AFDC AND MALE PARTICIPANTS

	Full National Supported Work Sample			
	AFDC Participants		Male Participants	
Variable	Treatments	Controls	Treatments	Controls
Age	33.37	33.63	24.49	23.99
	(7.43)	(7.18)	(6.58)	(6.54)
Years of School	10.30	10.27	10.17	10.17
	(1.92)	(2.00)	(1.75)	(1.76)
Proportion	.70	.69	.79	.80
High School Dropouts	(.46)	(.46)	(.41)	(.40)
Proportion Married	.02	.04	.14	.13
	(.15)	(.20)	(.35)	(.35)
Proportion Black	.84	.82	.76	.75
	(.37)	(.39)	(.43)	(.43)
Proportion Hispanic	.12	.13	.12	.14
	(.32)	(.33)	(.33)	(.35)
Real Earnings	$393	$395	1472	1558
1 year Before	(1,203)	(1,149)	(2656)	(2961)
Training	[43]	[41]	[58]	[63]
Real Earnings	$854	$894	2860	3030
2 years Before	(2,087)	(2,240)	(4729)	(5293)
Training	[74]	[79]	[104]	[113]
Hours Worked	90	92	278	274
1 year Before	(251)	(253)	(466)	(458)
Training	[9]	[9]	[10]	[10]
Hours Worked	186	188	458	469
2 years Before	(434)	(450)	(654)	(689)
Training	[15]	[16]	[14]	[15]
Month of Assignment	−12.26	−12.30	−16.08	−15.91
(Jan. 78 = 0)	(4.30)	(4.23)	(5.97)	(5.89)
Number of Observations	800	802	2083	2193

Note: The numbers shown in parentheses are the standard deviations and those in the square brackets are the standard errors.

views. Many participants failed to complete these interviews, and this sample attrition potentially biases the experimental results. Fortunately the largest source of attrition does not affect the integrity of the experimental design. Largely due to limited resources, the NSW administrators scheduled a 27th-month interview for only 65 percent of the participants and a 36th-month interview for only 24 percent of the non-AFDC participants. None of the AFDC participants were scheduled for a 36th-month interview, but the AFDC resurvey during the fall of 1979 interviewed 75 percent of these women anywhere from 27 to 44 months after the baseline. Since the trainee and control group members were randomly scheduled for all of these interviews, this source of attrition did not bias the experimental evaluation of the NSW program.

Naturally, the program administrators did not locate all of the participants scheduled for these interviews. The proportion of participants who failed to complete scheduled interviews varied across experimental group, time, and target group. While the response rates were statistically significantly higher for the treatment as opposed to the control group members, the differences in response rates were usually only a few percentage points. For the 27th-month interview, 72 percent of the treatments and 68 percent of the control group members completed interviews. The differences in response rates were

TABLE 2—ANNUAL EARNINGS OF NSW TREATMENTS, CONTROLS, AND EIGHT CANDIDATE COMPARISON GROUPS FROM THE *PSID* AND THE *CPS-SSA*

Year	Treatments	Controls	Comparison Group[a,b]							
			PSID-1	*PSID*-2	*PSID*-3	*PSID*-4	*CPS-SSA*-1	*CPS-SSA*-2	*CPS-SSA*-3	*CPS-SSA*-4
1975	$895	$877	7,303	2,327	937	6,654	7,788	3,748	4,575	2,049
	(81)	(90)	(317)	(286)	(189)	(428)	(63)	(250)	(135)	(333)
1976	$1,794	$646	7,442	2,697	665	6,770	8,547	4,774	3,800	2,036
	(99)	(63)	(327)	(317)	(157)	(463)	(65)	(302)	(128)	(337)
1977	$6,143	$1,518	7,983	3,219	891	7,213	8,562	4,851	5,277	2,844
	(140)	(112)	(335)	(376)	(229)	(484)	(68)	(317)	(153)	(450)
1978	$4,526	$2,885	8,146	3,636	1,631	7,564	8,518	5,343	5,665	3,700
	(270)	(244)	(339)	(421)	(381)	(480)	(72)	(365)	(166)	(593)
1979	$4,670	$3,819	8,016	3,569	1,602	7,482	8,023	5,343	5,782	3,733
	(226)	(208)	(334)	(381)	(334)	(462)	(73)	(371)	(170)	(543)
Number of Observations	600	585	595	173	118	255	11,132	241	1,594	87

[a] The Comparison Groups are defined as follows: *PSID*-1: All female household heads continuously from 1975 through 1979, who were between 20 and 55-years-old and did not classify themselves as retired in 1975; *PSID*-2: Selects from the *PSID*-1 group all women who received AFDC in 1975; *PSID*-3: Selects from the *PSID*-2 all women who were not working when surveyed in 1976; *PSID*-4: Selects from the *PSID*-1 group all women with children, none of whom are less than 5-years-old; *CPS-SSA*−1: All females from Westat *CPS-SSA* sample; *CPS-SSA*-2: Selects from *CPS-SSA*-1 all females who received AFDC in 1975; *CPS-SSA*-3: Selects from *CPS-SSA*-1 all females who were not working in the spring of 1976; *CPS-SSA*-4: Selects from *CPS-SSA*-2 all females who were not working in the spring of 1976.

[b] All earnings are expressed in 1982 dollars. The numbers in parentheses are the standard errors. For the NSW treatments and controls, the number of observations refer only to 1975 and 1979. In the other years there are fewer observations, especially in 1978. At the time of the resurvey in 1979, treatments had been out of Supported Work for an average of 20 months.

larger across time and target group. For example, 79 percent of the scheduled participants completed the 9th-month interview, while 70 percent completed the 27th-month interview. The AFDC participants responded at consistently higher rates than the other target groups; 89 percent of the AFDC participants completed the 9th-month interview as opposed to 76 percent of the other participants. While these response rates indicate that the experimental results may be biased, especially for the non-AFDC participants, comparisons between the baseline characteristics of participants who did and did not complete a 27th-month interview suggest that whatever bias exists may be small.[7]

Table 1 presents some sample statistics describing the baseline characteristics of the AFDC treatment and control groups as well as those of the male NSW participants in the other three target groups.[8] As would be expected from random assignment, the

[7] This study evaluates the AFDC females separately from the non-AFDC males. This distinction is common in the literature, but it is also motivated by the differences between the response rates for the two groups. The Supported Work Evaluation Study (*Public Use Files User's Guide*, Documentation Series No. 1, pp. 18–27) presents a more detailed discussion of sample attrition. My working paper (1984, tables 1.1 and 2.3), compares the characteristics and employment history of the full NSW sample to the sample with pre- and postprogram earnings data. Randall Brown (1979) reports that there is no evidence that the response rates affect the experimental estimates for the AFDC women or ex-addicts, while the evidence for the ex-offenders and high school dropouts is less conclusive.

[8] The female participants from the non-AFDC target groups were not surveyed during the AFDC resurvey in the fall of 1979 and consequently do not report 1979 earnings and are not included with the AFDC sample. Excluding these women from the analysis does not affect the integrity of the experimental design.

TABLE 3—ANNUAL EARNINGS OF NSW MALE TREATMENTS, CONTROLS, AND SIX CANDIDATE COMPARISON GROUPS FROM THE *PSID* AND *CPS-SSA*

			Comparison Group[a,b]					
Year	Treatments	Controls	*PSID*-1	*PSID*-2	*PSID*-3	*CPS-SSA*-1	*CPS-SSA*-2	*CPS-SSA*-3
1975	$3,066	$3,027	19,056[a]	7,569	2,611	13,650	7,387	2,729
	(283)	(252)	(272)	(568)	(492)	(73)	(206)	(197)
1976	$4,035	$2,121	20,267	6,152	3,191	14,579	6,390	3,863
	(215)	(163)	(296)	(601)	(609)	(75)	(187)	(267)
1977	$6,335	$3,403	20,898	7,985	3,981	15,046	9,305	6,399
	(376)	(228)	(296)	(621)	(594)	(76)	(225)	(398)
1978	$5,976	$5,090	21,542	9,996	5,279	14,846	10,071	7,277
	(402)	(227)	(311)	(703)	(686)	(76)	(241)	(431)
Number of Observations	297	425	2,493	253	128	15,992	1,283	305

[a]The Comparison Groups are defined as follows: *PSID*-1: All male household heads continuously from 1975 through 1978, who were less than 55-years-old and did not classify themselves as retired in 1975; *PSID*-2: Selects from the *PSID*-1 group all men who were not working when surveyed in the spring of 1976; *PSID*-3: Selects from the *PSID*-1 group all men who were not working when surveyed in either spring of 1975 or 1976; *CPS-SSA*-1: All males based on Westat's criteria, except those over 55-years-old; *CPS-SSA*-2: Selects from *CPS-SSA*-1 all males who were not working when surveyed in March 1976; *CPS-SSA*-3: Selects from the *CPS-SSA*-1 unemployed males in 1976 whose income in 1975 was below the poverty level.

[b]All earnings are expressed in 1982 dollars. The numbers in parentheses are the standard errors. The number of observations refer only to 1975 and 1978. In the other years there are fewer observations. The sample of treatments is smaller than the sample of controls because treatments still in Supported Work as of January 1978 are excluded from the sample, and in the young high school target group there were by design more controls than treatments.

means of the characteristics and pretraining hours and earnings of the experimental groups are nearly the same. For example, the mean earnings of the AFDC treatments and the AFDC controls in the year before training differ by $2, the mean age of the two groups differ by 3 months, and the mean years of schooling are identical. None of the differences between the treatment's and control's characteristics, hours, and earnings are statistically significant.

The first two columns of Tables 2 and 3 present the annual earnings of the treatment and control group members.[9] The earnings of the experimental groups were the same in the pre-training year 1975, diverged during the employment program, and converged to some extent after the program ended. The post-training year was 1979 for the AFDC females and 1978 for the males.[10]

Columns 2 and 3 in the first row of Tables 4 and 5 show that both the unadjusted and regression-adjusted pre-training earnings of the two sets of treatment and control group members are essentially identical. Therefore, because of the NSW program's experimental design, the difference between the post-training earnings of the experimental groups is an unbiased estimator of the training effect, and the other estimators described in columns 5–10(11) are unbiased estimators as well. The estimates in column 4 indicate that the

[9]All earnings presented in this paper are in 1982 dollars. The NSW Public Use Files report earnings in experimental time, months from the baseline, and not calendar time. However, my working paper describes how to convert the experimental earnings data to the annual data reported in Tables 2 and 3.

[10]The number of NSW male treatment group members with complete pre- and postprogram earnings is much smaller than the full sample of treatments or the partial sample of control group members. This difference is largely explained by the two forms of sample attrition discussed earlier. In addition, however, (*i*) this paper excludes all males who were in Supported Work in January 1978, or entered the program before January 1976; (*ii*) in one of the sites, the administrators randomly assigned .4 instead of one-half of the qualified high school dropouts into the treatment group.

TABLE 4—EARNINGS COMPARISONS AND ESTIMATED TRAINING EFFECTS FOR THE NSW AFDC PARTICIPANTS USING COMPARISON GROUPS FROM THE *PSID* AND THE *CPS-SSA*[a,b]

Name of Comparison Group[d]	Comparison Group Earnings Growth 1975–79 (1)	NSW Treatment Earnings Less Comparison Group Earnings: Pre-Training Year, 1975 Unadjusted (2)	Pre-Training Year, 1975 Adjusted[c] (3)	Post-Training Year, 1979 Unadjusted (4)	Post-Training Year, 1979 Adjusted[c] (5)	Difference in Differences: Difference in Earnings Growth 1975–79 Treatments Less Comparisons Without Age (6)	With Age (7)	Unrestricted Difference in Differences: Quasi Difference in Earnings Growth 1975–79 Unadjusted (8)	Adjusted[c] (9)	Controlling for All Observed Variables and Pre-Training Earnings: Without AFDC (10)	With AFDC (11)
	2,942	−17	−22	851	861	833	883	843	864	854	–
Controls	(220)	(122)	(122)	(307)	(306)	(323)	(323)	(308)	(306)	(312)	
PSID-1	713	−6,443	−4,882	−3,357	−2,143	3,097	2,657	1746	1,354	1664	2,097
	(210)	(326)	(336)	(403)	(425)	(317)	(333)	(357)	(380)	(409)	(491)
PSID-2	1,242	−1,467	−1,515	1,090	870	2,568	2,392	1,764	1,535	1,826	–
	(314)	(216)	(224)	(468)	(484)	(473)	(481)	(472)	(487)	(537)	
PSID-3	665	−77	−100	3,057	2,915	3,145	3,020	3,070	2,930	2,919	–
	(351)	(202)	(208)	(532)	(543)	(557)	(563)	(531)	(543)	(592)	
PSID-4	928	−5,694	−4,976	−2,822	−2,268	2,883	2,655	1,184	950	1,406	2,146
	(311)	(306)	(323)	(460)	(491)	(417)	(434)	(483)	(503)	(542)	(652)
CPS-SSA-1	233	−6,928	−5,813	−3,363	−2,650	3,578	3,501	1,214	1,127	536	1,041
	(64)	(272)	(309)	(320)	(365)	(280)	(282)	(272)	(309)	(349)	(503)
CPS-SSA-2	1,595	−2,888	−2,332	−683	−240	2,215	2,068	447	620	665	–
	(360)	(204)	(256)	(428)	(536)	(438)	(446)	(468)	(554)	(651)	
CPS-SSA-3	1,207	−3,715	−3,150	−1,122	−812	2,603	2,615	814	784	−99	1,246
	(166)	(226)	(325)	(311)	(452)	(307)	(328)	(305)	(429)	(481)	(720)
CPS-SSA-4	1,684	−1,189	−780	926	756	2,126	1,833	1,222	952	827	–
	(524)	(249)	(283)	(630)	(716)	(654)	(663)	(637)	(717)	(814)	

[a] The columns above present the estimated training effect for each econometric model and comparison group. The dependent variable is earnings in 1979. Based on the experimental data, an unbiased estimate of the impact of training presented in col. 4 is $851. The first three columns present the difference between each comparison group's 1975 and 1979 earnings and the difference between the pre-training earnings of each comparison group and the NSW treatments.

[b] Estimates are in 1982 dollars. The numbers in parentheses are the standard errors.

[c] The exogenous variables used in the regression adjusted equations are age, age squared, years of schooling, high school dropout status, and race.

[d] See Table 2 for definitions of the comparison groups.

earnings of the AFDC females were $851 higher than they would have been without the NSW program, while the earnings of the male participants were $886 higher.[11] Moreover, the other columns show that the econometric procedure does not affect these estimates.

[11] It is commonly believed that the NSW program had little impact on the earnings of the male participants (see MDRC; A. P. Bernstein et al., 1985). My working paper discusses why this estimated impact differs from the results discussed elsewhere. The 1978 earnings data were largely collected during the 36th-month interview, where the difference between the male treatment and control group members' earnings averaged $175 per quarter.

II. Nonexperimental Estimates

In addition to providing researchers with a simple estimate of the impact of an employment program, MDRC's experimental data can also be used to evaluate several nonexperimental methods of program evaluation. This section puts aside the NSW control group and evaluates the NSW program using some of the econometric procedures found in studies of the employment and training programs administered under the MDTA, CETA, and JTPA.[12]

[12] These acronyms refer to the Manpower Development and Training Act–1962, the Comprehensive Em-

TABLE 5—EARNINGS COMPARISONS AND ESTIMATED TRAINING EFFECTS FOR THE NSW MALE PARTICIPANTS USING COMPARISON GROUPS FROM THE *PSID* AND THE *CPS-SSA*[a,b]

Name of Comparison Group[d]	Comparison Group Earnings Growth 1975–78 (1)	NSW Treatment Earnings Less Comparison Group Earnings: Pre-Training Year, 1975 Unadjusted (2)	Pre-Training Year, 1975 Adjusted[c] (3)	Post-Training Year, 1978 Unadjusted (4)	Post-Training Year, 1978 Adjusted[c] (5)	Difference in Differences: Difference in Earnings Growth 1975–78 Treatments Less Comparisons Without Age (6)	Difference in Differences: Difference in Earnings Growth 1975–78 Treatments Less Comparisons With Age (7)	Unrestricted Difference in Differences: Quasi Difference in Earnings Growth 1975–78 Unadjusted (8)	Unrestricted Difference in Differences: Quasi Difference in Earnings Growth 1975–78 Adjusted[c] (9)	Controlling for All Observed Variables and Pre-Training Earnings (10)
Controls	$2,063	$39	$−21	$886	$798	$847	$856	$897	$802	$662
	(325)	(383)	(378)	(476)	(472)	(560)	(558)	(467)	(467)	(506)
PSID-1	$2,043	−$15,997	−$7,624	−$15,578	−$8,067	$425	−$749	−$2,380	−$2,119	−$1,228
	(237)	(795)	(851)	(913)	(990)	(650)	(692)	(680)	(746)	(896)
PSID-2	$6,071	−$4,503	−$3,669	−$4,020	−$3,482	$484	−$650	−$1,364	−$1,694	−$792
	(637)	(608)	(757)	(781)	(935)	(738)	(850)	(729)	(878)	(1024)
PSID-3	($3,322	($455	$455	$697	−$509	$242	−$1,325	$629	−$552	$397
	(780)	(539)	(704)	(760)	(967)	(884)	(1078)	(757)	(967)	(1103)
CPS-SSA-1	$1,196	−$10,585	−$4,654	−$8,870	−$4,416	$1,714	$195	−$1,543	−$1,102	−$805
	(61)	(539)	(509)	(562)	(557)	(452)	(441)	(426)	(450)	(484)
CPS-SSA-2	$2,684	−$4,321	−$1,824	−$4,095	−$1,675	$226	−$488	−$1,850	−$782	−$319
	(229)	(450)	(535)	(537)	(672)	(539)	(530)	(497)	(621)	(761)
CPS-SSA-3	$4,548	$337	$878	−$1,300	$224	−$1,637	−$1,388	−$1,396	$17	$1,466
	(409)	(343)	(447)	(590)	(766)	(631)	(655)	(582)	(761)	(984)

[a] The columns above present the estimated training effect for each econometric model and comparison group. The dependent variable is earnings in 1978. Based on the experimental data an unbiased estimate of the impact of training presented in col. 4 is $886. The first three columns present the difference between each comparison group's 1975 and 1978 earnings and the difference between the pre-training earnings of each comparison group and the NSW treatments.

[b] Estimates are in 1982 dollars. The numbers in parentheses are the standard errors.

[c] The exogenous variables used in the regression adjusted equations are age, age squared, years of schooling, high school dropout status, and race.

[d] See Table 3 for definitions of the comparison groups.

The researchers who evaluated these federally sponsored programs devised both experimental and nonexperimental procedures to estimate the training effect, because they recognized that the difference between the trainees' pre- and post-training earnings was a poor estimate of the training effect. In a dynamic economy, the trainees' earnings may grow even without an effective program. The goal of these program evaluations is to estimate the earnings of the trainees had they not participated in the program. Researchers using experimental data take the earnings of the control group members to be an estimate of the trainees' earnings without the program. Without experimental data, researchers estimate the earnings of the trainees by using the regression-adjusted earnings of a comparison group drawn from the population. This adjustment takes into account that the observable characteristics of the trainees and the comparison group members differ, and their unobservable characteristics may differ as well.

Any nonexperimental evaluation of a training program must explicitly account for these differences in a model describing the observable determinants of earnings and the process by which the trainees are selected into the program. However, unlike in an experimental evaluation, the nonexperimental estimates of the training effect depend crucially on the way that the earnings and participation equations are specified. If the econometric model is specified correctly, the nonexperimental estimates should be the same (within sampling error) as the training effect generated from the experimental data, but if there is a significant difference between the nonexperimental and the experi-

ployment and Training Act–1973, and the Job Training Partnership Act–1982.

mental estimates, the econometric model is misspecified.[13]

The first step in a nonexperimental evaluation is to select a comparison group whose earnings can be compared to the earnings of the trainees. Tables 2 and 3 present the mean annual earnings of female and male comparison groups drawn from the *Panel Study of Income Dynamics* (*PSID*) and Westat's Matched *Current Population Survey–Social Security Administration File* (*CPS-SSA*). These groups are characteristic of two types of comparison groups frequently used in the program evaluation literature. The *PSID*-1 and the *CPS-SSA*-1 groups are large, stratified random samples from populations of household heads and households, respectively.[14] The other, smaller, comparison groups are composed of individuals whose characteristics are consistent with some of the eligibility criteria used to admit applicants into the NSW program. For example, the *PSID*-3 and *CPS-SSA*-4 comparison groups in Table 2 include females from the *PSID* and the *CPS-SSA* who received AFDC payments in 1975, and were not employed in the spring of 1976. Tables 2 and 3 show that the NSW trainees and controls have earnings histories that are more similar to those of the smaller comparison groups, whose characteristics are similar to theirs, than those of the larger comparison groups.[15]

The second step in a nonexperimental evaluation is to specify a model of earnings and program participation to adjust for differences between the trainees and comparison group members. Equations (1) through (4) describe a conventional model of earnings and program participation that is typical of the kind econometric researchers use for this problem:

$$(1) \quad y_{it} = \delta D_i + \beta X_{it} + b_i + n_t + \varepsilon_{it}$$

$$(2) \quad \varepsilon_{it} - \rho \varepsilon_{it-1} = \nu_{it}$$

$$(3) \quad d_{is} = y_{is} + \gamma Z_{is} + \eta_{is}$$

$$(4) \quad D_i = 1 \quad \text{if } d_{is} > 0; \quad D_i = 0 \quad \text{if } d_{is} < 0.$$

In equation (1), earnings in each period are a function of a vector of individual characteristics, X_{it}, such as age, schooling, and race for individual i in time t; a dummy variable indicating whether the individual participated in training in period $s+1$, D_i; and an error with individual- and time-specific components and a serially correlated transitory disturbance. The transitory disturbance follows the first-order serial corre-

[13] Thomas Fraker, Maynard, and Lyle Nelson (1984) describe a similar study using the NSW AFDC and Young High School Dropouts. Instead of focusing the study on models of earnings and program participation, their study evaluates several strategies for choosing matched comparison groups. They use grouped Social Security earnings data when comparing the annual earnings of the NSW treatments to the earnings of each of the comparison groups.

[14] The *PSID* file including the poverty subsample selects only women and men who were household heads continuously from 1975 to 1979, and 1978, respectively. The *CPS-SSA* file matches the March 1976 *Current Population Survey* with Social Security earnings. Only individuals in the labor force in March 1976 with nominal income less than $20,000 and household income less than $30,000 are in this sample. In 1976, 2 percent of the females and 21 percent of the males had earnings at the Social Security maximum. In this paper, females younger than 20 or older than 55 and males older than 55 are excluded from the comparison groups.

[15] Not only are the pre-training earnings of the *PSID*-3 comparison group in Table 2 similar to the earnings of the NSW experimental groups, but the characteristics of these groups are similar as well. The mean age for the *PSID*-3 women is 40.95; the mean years of schooling is 10.31; the proportion of high school dropouts is 0.63; the proportion married is 0.01; the proportion black is 0.85; and the proportion Hispanic is 0.03. I experimented with matching the comparison groups even more closely to the pre-training characteristics of the experimental sample. However, these closely matched comparison groups are extremely small. For example there were 57 women from the *PSID* who received welfare payments in 1975, were not employed at the time of the survey in 1976, resided in a metropolitan area, and had only school-age children. The mean earnings of this group were $1,137 in 1975; $673 in 1976; $743 in 1977; $1,222 in 1978; and $1,697 in 1979.

lation process described in equation (2). Equations (3) and (4) specify the participation decision: an individual participates in training and is admitted into the program in period $s+1$ if the latent variable d_{is} rises above zero. The participation equation is typically rationalized by the notion that the supply of individuals who decide to participate in training depends on the net benefit they expect to receive from participation and on the demand of the program administrators for training participants. The participation latent variable is typically a function of a vector of characteristics Z_{is}, current earnings y_{is}, and an error.

The estimators described in the column headings in Tables 4 and 5 (as well as many others in the literature) are based on econometric specifications that place different restrictions on the training model represented by equations (1)–(4) (although one common restriction assumes that the unobservables in the earnings and participation equations are uncorrelated). These estimates are consistent only insofar as their restrictions are consistent with the data. The restrictions can be tested provided the nonexperimental data base has sufficient information on the pre-training earnings and demographic characteristics of the trainees and comparison group members. An econometrician is unlikely to take seriously an estimate based on a model that failed one of these specification tests. Therefore, the results of such tests can often aid the researcher in choosing among alternative estimates. It follows, then, that simply checking whether the nonexperimental estimates replicate the experimental results and whether these estimates vary across different econometric procedures is not the only motivation for comparing experimental to nonexperimental methods. By making this comparison, we can also discover whether the nonexperimental data alone reliably indicate when an econometric model is misspecified and whether specification tests, which are supposed to ensure that the econometric model is consistent with the data, lead researchers to choose the "right" estimator.

In practice, the available data affect the composition of the comparison groups and the flexibility of the econometric specifications. For example, since there is only one year of pre-training earnings data, we cannot evaluate all of the econometric procedures that have been used in the literature, nor can we test all of the econometric specifications analyzed in this paper with the nonexperimental data alone.[16]

Nevertheless, several one-step estimators are evaluated in Tables 4 and 5, starting with the simple difference between the treatment and comparison group members' post-training earnings in column 4. Column 5 presents this earnings difference controlling for age, schooling, and race. This cross-sectional estimator is based on a model where these demographic variables are assumed to adequately control for differences between the earnings of the trainees and comparison group members. Column 6 presents the difference between the two nonexperimental groups' pre- and post-training earnings growth. This estimator allows for an unobserved individual fixed effect in the earnings equation and for the possibility that individuals with low values of this unobservable are more likely to participate in training. The cross-sectional estimator described in column 5 is now biased since the training dummy variable is correlated with the error in the earnings equation. Differencing the earnings equation removes the fixed effect, leaving[17]

$$(5)\quad y_{it} - y_{is} = \delta D_i + \beta \cdot AGE_i + (\eta_t - \eta_s) + \varepsilon_{it} - \varepsilon_{is}.$$

[16] One limitation of the NSW Public Use File is that there is only one year of pre-experimental data available in calendar time as opposed to experimental time. Consequently, there are several nonexperimental procedures which require more than a year of pre-training earnings data that are not evaluated in this paper. If additional data were available, it is possible that these procedures would adequately control for differences between the NSW treatments and comparison group members and that the results of the specification tests would correctly guide an econometrician away from some of the estimates presented in this paper to the estimates based on these other procedures. See John Abowd (1983), Ashenfelter, Ashenfelter and Card, Bassi (1983b, 1984), and James Heckman and Richard Robb (1985).

[17] The other demographic variables, schooling and race, are constant over time.

The comparison group's earnings growth represents the earnings growth that the trainees would have experienced without the program. However, since the trainees may experience larger earnings growth than the comparison group members simply because they are usually younger, column 7 presents the difference between the earnings growth of the two groups controlling for age.

Column 8 presents the difference between the post-training earnings of the treatment and comparison group members, holding constant the level of pre-training earnings, while the estimator in column 9 controls both for pre-training earnings and the demographic variables. These estimators are consistent when the model of program participation stipulates that the trainees' pre-program earnings fell (see Table 1) because some of the training participants experienced some bad luck in the years prior to training. In this case, we would expect the trainees' earnings to grow even without the program.[18] The difference in differences estimator in columns 6 and 7 is now biased, since the training dummy variable is correlated with the transitory component of pre-training earnings in equation (5).[19] Finally, columns 10 and 11 report the estimates of the training effects controlling for all observed variables. Besides the variables described earlier, the additional regressors are employment status in 1976, AFDC status in 1975, marital status, residency in a metropolitan area with more than 100,000 persons, and number of children.

Unlike the experimental estimates, the nonexperimental estimates are sensitive both to the composition of the comparison group and to the econometric procedure. For example, many of the estimates in column 9 of Table 4 replicate the experimental results, while other estimates are more than $1,000 larger than the experimental results. More specifically, the results for the female participants (Table 4) tend to be positive and larger than the experimental estimate, while for the male participants (Table 5), the estimates tend to be negative and smaller than the experimental impact.[20] Additionally, the nonexperimental procedures replicate the experimental results more closely when the nonexperimental data include pre-training earnings rather than cross-sectional data alone or when evaluating female rather than male participants.

The sensitivity of the nonexperimental estimates to different specifications of the econometric model is not in itself a cause for alarm. After all, few econometricians expect estimators based on misspecified models to replicate the results of experiments. Hence the considerable range of estimates is understandable given that inconsistent estimators are likely to yield inaccurate estimates. Before taking some of these estimates too seriously, many econometricians at a minimum would require that their estimators be based on econometric models that are consistent with the pre-training earnings data. Thus, if the regression-adjusted difference between the post-training earnings of the two groups is going to be a consistent estimator of the training effect, the regression-adjusted pre-training earnings of the two groups should be the same.

Based on this specification test, econometricians might reject the nonexperimental estimates in columns 4–7 of Table 4 in favor of the ones in columns 8–11. Few econometricians would report the training effect of $870 in column 5, even though this estimate differs from the experimental result

[18] Researchers have observed this dip in pre-training earnings for successive MDTA and CETA cohorts since 1964. See Ashenfelter (Table 1); Ashenfelter and Card (Table 1); Bassi (1983a, Table 4.1); and Kiefer (1979a, Table 4-1).

[19] This estimator is similar to one devised by Arthur Goldberger (1972) (or see G. S. Maddala, 1983) to evaluate the Head Start Program where participation in the program depended on a child's test score plus a random error. Similarly, participation in a training program can be thought of as a function of pre-training earnings and a random error. My working paper shows that this estimator is consistent as long as the unobservables in the earnings and participation equations are uncorrelated, and all of the observable variables in the model are used as regressors in the earnings equation.

[20] The magnitude of these training effects is similar to the estimates reported in studies of the 1964 MDTA cohort, the 1969–70 MDTA cohort, and the 1976–77 CETA cohort. (See my working paper, Table I.1.)

by only $19. If the cross-sectional estimator properly controlled for differences between the trainees and comparison group members, we would not expect the difference between the regression adjusted pre-training earnings of the two groups to be $1,550, as reported in column 3. Likewise, econometricians might refrain from reporting the difference in differences estimates in columns 6 and 7, even though all these estimates are within two standard errors of $3,000. As noted earlier, this estimator is not consistent with the decline in the trainees' pre-training earnings.

This point can also be made with the estimates for the NSW male participants (Table 5). For example, all but one of the difference in differences estimates in column 6 are within one standard error of the experimental estimate. Yet for two reasons it is unlikely econometricians would report these estimates. First, as the results in column 7 suggest, since the trainees are younger their earnings might be expected to grow faster than the earnings of the comparison group members even without training. Second, as shown in Table 1, the pre-training earnings of the male participants fell in the period before training, suggesting that the trainees' earnings will grow even if the program is ineffective. Here again, econometricians might turn to the considerable range of estimates in columns 8–10.

The results of these specification tests suggest that an econometrician might report one of the estimates in columns 8–11. However, even without the experimental data, a researcher would find that the estimated training effect is still sensitive to the set of variables included in the earnings equation and to the composition of the comparison group. In Table 4, the estimates using the female household heads with school-age children (*PSID*-4) as a comparison group differ by more than $1,000. The largest estimate overstates the experimental result by $1,300, while the smallest estimate is within $100 of the experimental estimate. Likewise in column 11, we find that the same estimator with different comparison groups yields a set of estimates that vary by more than $1,000. The estimates for the male participants exhibit the same sensitivity to the choice of a comparison group and to the set of variables used as regressors in the earnings equation. However, the estimated standard errors associated with these training effects are larger than for the female estimates, making it more difficult to draw many conclusions from these results.

Without additional data it is difficult to see how a researcher would choose a training effect from among estimates. Moreover, the nonexperimental data base alone does not allow the econometrician to test whether these estimates are based on econometric models that adequately control for differences between the earnings of the trainees and comparison group members. In this case, comparisons between the experimental and nonexperimental estimates is the best specification test available.[21]

Specification tests that use pre-training earnings data are an appealing means to choose between alternative estimates, but these tests are not themselves always sufficient to identify unreliable estimators. This point becomes clear when we compare the estimates using the *PSID*-3 comparison group (as defined in Table 2) and those using the NSW control group. The characteristics of these two groups are nearly the same, as are their unadjusted and adjusted pre-training earnings. In each case the cross-sectional estimator in column 5 appears to be an unbiased estimate of the training effect. Moreover, both sets of estimates are unaffected by alternative econometric procedures. Thus both the experimental and nonexperimental estimates pass the same specification tests; nevertheless the nonexperimental estimate is approximately $2,100 larger than the experimental result. If a researcher did not know that one set of estimates was based on an experimental data set, it is hard to see how she or he would

[21]Ashenfelter, Ashenfelter and Card, and Bassi (1984) have noted in their studies using nonexperimental data that their results are sensitive to alternative econometric specifications and that there is evidence for male training participants that the econometric models are misspecified.

choose between two estimates where one training effect is roughly 3.5 times larger than the other.

III. Two-Step Estimates

The unobservables in the earnings equation were uncorrelated with those in the participation equation in all of the econometric models analyzed in the previous section. If, instead, the unobservables are correlated, none of the one-step least squares procedures are consistent estimators of the training effect. Individuals with high unobservables in their participation equation are more likely to participate in training. Yet if the unobservables in the earnings and participation equations are negatively correlated, these individuals are likely to have relatively low earnings, even after controlling for the observable variables in the model. Consequently, least squares underestimates the impact of training.

James Heckman (1978) proposes a two-step estimator that controls for the correlation between the unobservables by using the estimated conditional expectation of the earnings error as a regressor in the earnings equation. If the errors in the earnings and participation equations are jointly normally distributed, this conditional expectation is proportional to the conditional expectation of the error in the participation equation. Using the notation introduced in the last section, this relationship is expressed formally as

$$(6)\quad E(b_i + \varepsilon_{it} | Z_i, D_i) = \rho\sigma_\varepsilon \left[D_i \frac{\phi(\gamma Z_i)}{1 - \Phi(\gamma Z_i)} - (1 - D_i)\frac{\phi(\gamma Z_i)}{\Phi(\gamma Z_i)} \right] = rH_i,$$

where Z_i is a vector of observed variables, ρ is the correlation between the unobservables in the model, σ_ε^2 is the variance of the unobservables in the earnings equation, and $\phi(\cdot)$ and $\Phi(\cdot)$ are the normal density and distribution functions. Therefore the earnings equation can be rewritten as

$$(7)\quad Y_{it} = \delta D_i + \beta X_{it} + rH_i + v_i^*,$$

where v_i^* is an orthogonal error by construction. To estimate the training effect, δ, the researcher first uses the coefficients from a probit estimate of the reduced-form participation equation to calculate the conditional expectation, H_i, for both the trainees and comparison group members,[22] and, second, uses this estimate, $\hat{H}_i$, as a regressor in the earnings equation. The training effect is then estimated by least squares.[23]

Table 6 presents estimates for the female and male training participants using the NSW controls, the *PSID*-1 and *CPS-SSA*-1 as comparison groups.[24] Unless some variables are excluded from the earnings equation, the training effect in this procedure is identified by the nonlinearity of the probit function. Hence, the rows of Table 6 allow us to evaluate the sensitivity of these estimates to different exclusion restrictions. The second column associated with each set of training effects presents the estimated participation coefficient. If the unobservables are uncorrelated, this estimate should not be significantly different from zero. Therefore, these estimates allow us to test whether this restriction on the correlation between the unobservables is consistent with the nonex-

[22] This is a choice-based sampling problem, since the probability of being in the nonexperimental data set is high for the NSW treatment group members and low for the comparison group members. The estimated probability of participation depends not only on the observed variables but on the numbers of trainees and comparison group members. Heckman and Richard Robb (1985) show that this procedure is robust to choice-based sampling. For an example of an application of this estimator in the evaluation literature, see Mallar.

[23] Since the estimated value of this conditional expectation is used as a regressor instead of the true value, the estimated standard errors associated with the least squares estimates are inconsistent and must be corrected. See Heckman (1978,: 1979); William Greene (1981); John Ham (1982); and Ham and Cheng Hsiao (1984).

[24] The two-step estimates using the smaller comparison groups were associated with large estimated standard errors.

TABLE 6—ESTIMATED TRAINING EFFECTS USING TWO-STAGE ESTIMATOR

Variables Excluded from the Earnings Equation, but Included in the Participation Equation	Comparison Group	NSW AFDC Females: Heckman Correction for Program Participation Bias, Using Estimate of Conditional Expectation of Earnings Error as Regressor in Earnings Equation: Estimate of Coefficient for Training Dummy	NSW AFDC Females: Estimate of Expectation	NSW Males: Training Dummy	NSW Males: Estimate of Expectation
Marital Status, Residency in an SMSA, Employment Status in 1976, AFDC Status in 1975, Number of Children	*PSID*-1	1,129	−894	−1,333	−2,357
		(385)	(396)	(820)	(781)
	CPS-SSA-1	1,102	−606	−22	−1,437
		(323)	(480)	(584)	(449)
	NSW Controls	837	−18	899	−835
		(317)	(2376)	(840)	(2601)
Employment Status in 1976, AFDC Status in 1975, Number of Children	*PSID*-1	1,256	−823	–	–
		(405)	(410)		
	CPS-SSA-1	439	−979	–	–
		(333)	(481)		
	NSW Controls	–	–	–	–
Employment Status in 1976, Number of Children	*PSID*-1	1,564	−552	−1,161	−2,655
		(604)	(569)	(864)	(799)
	CPS-SSA-1	552	−902	13	−1,484
		(514)	(551)	(584)	(450)
	NSW Controls	851	147	889	−808
		(318)	(2385)	(841)	(2603)
No Exclusion Restrictions	*PSID*-1	1,747	−526	−667	−2,446
		(620)	(568)	(905)	(806)
	CPS-SSA-1	805	−908	213	−1,364
		(523)	(548)	(588)	(452)
	NSW Controls	861	284	889	−876
		(318)	(2385)	(840)	(2601)

Notes: The estimated training effects are in 1982 dollars. For the females, the experimental estimate of impact of the supported work program was $851 with a standard error of $317. The one-step estimates from col. 11 of Table 4 were $2,097 with a standard error of $491 using the *PSID*-1 as a comparison group, $1,041 with a standard error of $503 using the *CPS-SSA*-1 as a comparison group, and $854 with a standard error of $312 using the NSW controls as a comparison group. Estimates are missing for the case of three exclusions using the NSW controls since AFDC status in 1975 cannot be used as an instrument for the NSW females. For the males, the experimental estimate of impact of the supported work program was $886 with a standard error of $476. The one-step estimates from col. 10 of Table 5 were $−1,228 with a standard error of $896 using the *PSID*-1 as a comparison group, $−805 with a standard error of $484 using the *CPS-SSA*-1 as a comparison group, and $662 with a standard error of $506 using the NSW controls as a comparison group. Estimates are missing for the case of three exclusions for the NSW males as AFDC status is not used as an instrument in the analysis of the male trainees.

perimental data, and to examine whether this specification test leads econometricians to choose the "right" estimator.

The experimental estimates in Table 6 are consistent with MDRC's experimental design. All of these estimates are nearly identical to the experimental results presented in Tables 4 and 5. And furthermore, since the unobservables are uncorrelated by design, the estimated participation coefficients are never significantly different from zero.

Turning to the nonexperimental estimates we find that although the instruments used to identify the earnings equation have some effect on the results, generally these estimates are closer to the experimental estimates than are the one-step estimates (in column 11 of Tables 4 and 5). For the females, the difference between the two-step and one-step estimates are small relative to the estimated standard errors, and the estimates of the participation coefficient are only

marginally significantly different from zero. Interestingly, in one case when the *PSID*-1 sample is used as a comparison group, the estimated participation coefficient is significant (the *t*-statistic is 2.25) and the training effect of $1,129 is $968 closer to the experimental result than the one-step estimate. Additionally, this estimate is identical to the estimate using the *CPS-SSA*-1 comparison group, whereas the one-step estimates differed by $1,056. However, if an econometrician reported this training effect, she or he would have to argue that variables such as place of residence and prior AFDC status do not belong in the earnings equation. Otherwise, the econometrician is left to choose between a set of estimates that vary by as much as $1,308.

The two-step estimates are usually closer than the one-step estimates to the experimental results for the male trainees as well. One estimate, which used the *CPS-SSA*-1 sample as a comparison group, is within $600 of the experimental result, while the one-step estimate falls short by $1,695. The estimates of the participation coefficients are negative, although unlike these estimates for the females, they are always significantly different from zero. This finding is consistent with the example cited earlier in which individuals with high participation unobservables and low earnings unobservables were more likely to be in training. As predicted, the unrestricted estimates are larger than the one-step estimates. However, as with the results for the females, this procedure may leave econometricians with a considerable range ($1,546) of imprecise estimates; although, like the results for the females, there is no evidence that the results of the specification tests would lead econometricians to choose the "wrong" estimator.

IV. Conclusion

This study shows that many of the econometric procedures and comparison groups used to evaluate employment and training programs would not have yielded accurate or precise estimates of the impact of the National Supported Work Program. The econometric estimates often differ significantly from the experimental results. Moreover, even when the econometric estimates pass conventional specification tests, they still fail to replicate the experimentally determined results. Even though I was unable to evaluate all nonexperimental methods, this evidence suggests that policymakers should be aware that the available nonexperimental evaluations of employment and training programs may contain large and unknown biases resulting from specification errors.[25]

This study also yields several other findings that may help researchers evaluate other employment and training programs. First, the nonexperimental procedures produce estimates that are usually positive and larger than the experimental results for the female participants, and are negative and smaller than the experimental estimates for the male participants. Second, these econometric procedures are more likely to replicate the experimental results in the case of female rather than male participants. Third, longitudinal data reduces the potential for specification errors relative to the cross-sectional data. Finally, the two-step procedure certainly does no worse than, and may reduce the potential for specification errors relative to, the one-step procedures discussed in Section II.

More generally, this paper presents an alternative approach to the sensitivity analyses proposed by Leamer (1983, 1985) and others for bounding the specification errors associated with the evaluation of economic hypotheses. This objective is accomplished by comparing econometric estimates with experimentally determined results. The data from an experiment yield simple estimates of the impact of economic treatments that are independent of any model specification. Successful econometric methods are intended to

[25] There is some evidence that this message has been passed on to the appropriate policymakers. See Recommendations of the Job Training Longitudinal Survey Research Advisory Panel to Office of Strategic Planning and Policy Development, U.S. Department of Labor, November 1985. This has led to at least a tentative decision to operate some part of the Job Training Partnership Act program sites using random assignment. (See Ernst Stromsdorfer et al., 1985.)

reproduce these estimates. The only way we will know whether these econometric methods are successful is by making the comparison. This paper takes the first step along this path, but there are other experimental data bases available to econometricians and much work remains to be done. For example, there have been several other employment and training experiments testing the effect of training on earnings, four Negative Income Tax Experiments testing hypotheses about labor supply, a medical insurance experiment testing hypotheses about insurance and medical demand, a housing experiment testing hypotheses about housing demand and supply, and a time-of-day electricity pricing experiment testing hypotheses about electricity demand.[26] There clearly remain many opportunities to use the experimental method to assess the potential for specification bias in the evaluation of social programs, and in other areas of econometric research as well.

[26]See Linda Aiken and Barbara Kehrer (1985), Abt Associates (1984), Gary Burtless (1985), Barbara Goldman (1981), Goldman et al. (1985), Jerry Hausman and David Wise (1985), J. Ohls and G. Carcagno (1978), and SRI International (1983).

REFERENCES

Abowd, John, "Program Evaluation," Working Paper, University of Chicago, 1983.

Aiken, Linda and Kehrer, Barbara, *Evaluation Studies Review Annual*, Vol. 10, Beverly Hills: Sage Publications, 1985.

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

_____ **and Card, David,** "Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs," *Review of Economics and Statistics*, November 1985, *67*, 648–60.

Bernstein, A. P. et al., "The Forgotten Americans," *Business Week*, September 2, 1985, 50–55.

Bassi, Laurie, (1983a) "Estimating the Effect of Training Programs With Non-Random Selection," Princeton University, 1983.

_____, (1983b) "The Effect of CETA on the Post-Program Earnings of Participants," *Journal of Human Resources*, Fall 1983, *18*, 539–556.

_____, "Estimating the Effects of Training Programs with Nonrandom Selection," *Review of Economics and Statistics*, February 1984 *66*, 36–43.

Brown, Randall, "Assessing the Effects of Interview Nonresponse on Estimates of the Impact of Supported Work," Mathematica Policy Research Inc., Princeton, 1979.

Burtless, Gary, "Are Targeted Wage Subsidies Harmful? Evidence from a Wage Voucher Experiment," *Industrial and Labor Relations Review*, October 1985, *39*, 105–114.

Cooley, Thomas, McGuire, Thomas and Prescott, Edward, "Earnings and Employment Dynamics of Manpower Trainees: An Exploratory Econometric Analysis," in Ronald Ehrenberg, ed., *Research in Labor Economics*, Vol. 4, Suppl. 2, 1979, 119–47.

Dickinson, Katherine and Maynard, Rebecca, *The Impact of Supported Work on Ex-Addicts*, New York: Manpower Demonstration Research Corporation, 1981.

_____, **Johnson, Terry and West, Richard,** *An Analysis of the Impact of CETA Programs on Participants' Earnings*, Washington: Department of Labor, Employment and Training Administration, 1984.

Fraker, Thomas, Maynard, Rebecca and Nelson, Lyle, *An Assessment of Alternative Comparison Group Methodologies for Evaluating Employment and Training Programs*, Princeton: Mathematica Policy Research Inc., 1984.

Goldberger, Arthur, "Selection Bias in Evaluating Treatment Effects," Discussion Paper No. 123–72, Institute for Research on Poverty, University of Wisconsin, 1972.

Goldman, Barbara, "The Impacts of the Immediate Job Search Assistance Experiment," Manpower Demonstration Research Corporation, New York, 1981.

_____ **et al.,** "Findings From the San Diego Job Search and Work Experience Demonstration," New York: Manpower Demonstration Research Corporation, 1985.

Greene, William, "Sample Selection Bias as a Specification Error: Comment," *Econometrica*, May 1981, *49*, 795–98.

Ham, John, "Estimation of a Labor Supply Model with Censoring Due to Unemployment and Underemployment," *Review of Economic Studies*, July 1982, *49*, 335–54.

_____ **and Hsiao, Cheng,** "Two-Stage Estimation of Structural Labor Supply Parameters Using Interval Data From the 1971 Canadian Census," *Journal of Econometrics*, January/February 1984, *24*, 133–58.

Hausman, Jerry A. and Wise, David A., *Social Experimentation*, NBER, Chicago: University of Chicago Press, 1985.

Heckman, James, "Dummy Endogenous Variables in a Simultaneous Equations System," *Econometrica*, July 1978, *46*, 931–59.

_____, "Sample Selection Bias as a Specification Error," *Econometrica*, January 1979, *47*, 153–61.

_____ **and Robb, Richard,** "Alternative Methods for Evaluating the Impact of Interventions: An Overview," Working Paper, University of Chicago, 1985.

Hendry, David, "Econometrices: Alchemy or Science?" *Economica*, November 1980, *47*, 387–406.

Kemper, Peter and Long, David, "The Supported Work Evaluation: Technical Report on the Value of In-Program Output Costs," Manpower Demonstration Research Corporation, New York, 1981.

_____, _____, **and Thornton, Craig,** "The Supported Work Evaluation: Final Benefit-Cost Analysis," Manpower Demonstration Research Corporation, New York, 1981.

Kiefer, Nicholas, (1979a) *The Economic Benefits of Four Employment and Training Programs*, New York: Garland Publishing, 1979.

_____, (1979b) "Population Heterogeneity and Inference from Panel Data on the Effects of Vocational Training," *Journal of Political Economy* October 1979, *87*, S213–26.

LaLonde, Robert, "Evaluating the Econometric Evaluations of Training Programs With Experimental Data," Industrial Relations Section, Working Paper No. 183, Princeton University, 1984.

Leamer, Edward, "Let's Take the Con Out of Econometrics," *American Economic Review*, March 1983, *73*, 31–43.

_____, "Sensitivity Analysis Would Help," *American Economic Review*, June 1985, *75*, 308–13.

_____ **and Leonard, Herman,** "Reporting the Fragility of Regression Estimates," *Review of Economics and Statistics*, May 1983, *65*, 306–12.

McAleer, Michael, Pagan, Adrian and Volker, Paul, "What Will Take the Con Out of Econometrics?," *American Economic Review*, June 1985, *75*, 293–306.

Maddala, G. S., *Limited Dependent and Qualitative Variables in Econometrics*, Cambridge: Cambridge University Press, 1983.

Mallar, Charles, "Alternative Econometric Procedures for Program Evaluations: Illustrations From an Evaluation of Job Corps Book," *Proceedings of the American Statistical Association*, 1978, 317–21.

_____, **Kerachsky, Stuart and Thornton, Craig,** *The Short-Term Economic Impact of the Jobs-Corps Program*, Princeton: Mathematica Policy Research Inc., 1978.

Masters, Stanley and Maynard, Rebecca, "The Impact of Supported Work on Long-Term Recipients of AFDC Benefits," Manpower Demonstration Research Corporation, New York 1981.

Maynard, Rebecca, "The Impact of Supported Work on Young School Dropouts," Manpower Demonstration Research Corporation, New York, 1980.

Ohls, J. and Carcagno, G., *Second Evaluation of the Private Employment Agency Job Counsellor Project*, Princeton: Mathematica Policy Research Inc., 1978.

Piliavin, Irving and Gartner, Rosemary, "The Impact of Supported Work on Ex-Offenders," Manpower Demonstration Research Corporation, New York, 1981.

Stromsdorfer, Ernst et al., "Recommendations of the Job Training Longitudinal Survey Research Advisory Panel to the Office of Strategic Planning and Policy Development, U.S. Department of Labor," unpublished report, Washington, November 1985.

Abt Associates, "AFDC Homemaker-Home

Health Aid Demonstration Evaluation," 2nd Annual Report, Washington, 1984.

Manpower Demonstration Research Corporation, *Summary and Findings of the National Supported Work Demonstration*, Cambridge: Ballinger, 1983.

SRI International, *Final Report of the Seattle-Denver Income Maintenance Experiment: Design and Results*, Washington: Department of Health and Human Services, 1983.