Working Paper No. 203
Industrial Relations Section
Princeton University
January 1986

The Case for Evaluating Training Programs with Randomized Trials*

by

Orley Ashenfelter
Princeton University

January 8, 1986
Corrected January 9, 1986

*Prepared for presentation to the Education Sector of the World Bank Conference, "Investing in People: New Directions for Education and Training," January 7-10, 1986, Hunt Valley, Maryland.

ABSTRACT

This brief paper presents the reasons that I have come to conclude that the evaluation of the economic benefits of training programs will be greatly enhanced by the use of classical experimental methods. In particular, I am convinced that some of these training programs should be operated so that control and experimental groups are selected by random assignment (randomized trials). It follows that a simple comparison of earnings, employment, and other outcomes as between control and experimental groups subsequent to participation in the experimental program will provide a simple and credible estimate of program success (or failure).

The principal reason why randomized trials should be used in this field is that too much of the non-experimental estimation of the effects of training programs seems dependent on elements of model specification that cannot be subjected to powerful statistical tests. Moreover, these specification tests are merely necessary and not sufficient for the acceptability of a particular non-experimental estimation method, as an extensive example due to LaLonde demonstrates.

This brief paper presents the reasons that I have come to conclude that the evaluation of the economic benefits of training programs will be greatly enhanced by the use of classical experimental methods. In particular, I am convinced that some of these training programs should be operated so that control and experimental groups are selected by random assignment (randomized trials). It follows that a simple comparison of earnings, employment, and other outcomes as between control and experimental groups subsequent to participation in the experimental program will provide a simple and credible estimate of program success (or failure).

The principal reason why randomized trials should be used in this field is that too much of the non-experimental estimation of the effects of training programs seems dependent on elements of model specification that cannot be subjected to powerful statistical tests. As most of us who do econometric work know, it is rarely classical sampling error that makes us uncertain about our inferences. More typically, it is the possibility that some alternative, reasonable model specification will lend us to a different conclusion that makes us uncertain about our inferences. In the field of training program evaluation I believe this model uncertainty can now be documented.

A second reason why randomized trials may be useful is that the enhanced credibility of such studies allows for a far more dramatic effect of evidence in changing our perceptions about useful public policies. Frankly, the experimental studies of employment and training programs that have been completed are demonstrating some surprising results.

Consider, for example, the Dayton study of job placement rates indicated in Table 1. This experiment assigned economically disadvantaged workers randomly to announce their employer would receive (1) Tax Credit Vouchers that entitled an employer to a tax credit for hiring the indicated worker, (2) Direct Rebate Vouchers that provided an equivalent direct payment to an employer for hiring the indicated worker, and (3) no treatment. The results are simple and striking. Job placement is significantly lower for the workers receiving the treatment! As with any analysis, one small study should not be enough to induce a consensus view, but this evidence is hard to square up with a sanguine view of employment vouchers as a policy tool for disadvantaged workers.

On the other hand, consider the results in Table 2, which indicates the experimental findings for the National Supported Work (NSW) sheltered workshop training program. The numbers in columns (4) through (7) of the first row of Table 2 are all simple estimates of the effect of this training program on the earnings of economically disadvantaged groups of women quite similar to those in the Dayton experiment. These results demonstrate equally conclusively that the NSW training program did raise the earnings of disadvantaged workers. Apparently some types of employment and training programs are better than others.

The above results are provocative and, when presented to most groups of economists, it is my experience that the discussion turns almost immediately to substantive, as opposed to methodological, matters. The reason substantive matters get discussed is because the methods used to obtain the results are not only simple, they are cre-

dible. This is clearly the benefit of the experimental method.

On the other hand, there is a cost. These experiments were considerably more costly to monitor and evaluate than any simple non-experimental analysis would be. This raises a fundamental question: Is it possible to obtain the benefits of the credibility of the results of studies that use randomized trials by the use of non-experimental (that is, econometric) methods?

I.

The promise, of course, of econometric methods is the delivery of the information benefits associated with experimentation without the heavy costs. In the normal course of events it is impossible to determine whether the econometric results have produced these benefits because to do so would require a comparison of the results of the non-experimental econometric analysis with the results of an analysis of data from randomized trials. The former would generally never be completed if the latter were feasible. In the evaluation of training programs, however, we now have a remarkably careful study by Robert LaLonde (of the University of Chicago) that carries through analyses of both experimental and non-experimental data so just this comparison can be made.

LaLonde starts with the results of the NSW experiment, which are contained in the first row of Table 2 (for females) and Table 3 (for males). Columns (2) and (3) of the first rows of these tables indicate

that the differences between the earnings of the treatment and control groups in the year prior to training are very small, as should be the case with randomized trials. Although I do not present the results here, LaLonde also shows how remarkably similar the characteristics of the experimental and control groups are, demonstrating the success of the randomization.

Columns (4)-(6) of Tables 2 and 3 present a variety of different estimates of the effect of training on trainee earnings, all of which are obtained from estimators that are unbiased so long as the data are obtained from randomized trials. Column (4) is the simple post-training earnings difference, column (5) is the post-training earnings difference after a regression adjustment for various demographic variables, column (6) is the earnings change from pre-program to post-program for the trainees less the same change for the control groups, and column (7) adjusts the difference-in-differences estimator in column (6) for a linear term in age (the other demographic variables do not change and so do not enter the regression). As expected, all of these estimates of the training effect on earnings are essentially identical. Moreover, this training effect is positive, although it is only on the borderline of statistical significance for the male group.

Now suppose an econometrician were asked to evaluate the earnings effect of this training program, but that no randomized trials had been performed. Naturally, the econometrician would look for a non-experimental comparison group to use for the analysis and would use econometric methods to further adjust the data for comparability. This

is precisely what LaLonde has done by using the Panel Survey of Income Dynamics (PSID), a special component of which is a longitudinal sample of low income families. The results of a part of LaLonde's analysis of these non-experimental data are contained in the remaining rows of Tables 2 and 3. The way the non-experimental comparison samples was selected is indicated in the stubs of the tables.

The first important conclusion I am lead to draw from these tables is that virtually as many estimates of training effects can be obtained as there are data sets and methods. At a minimum, therefore, the econometrician should draw the conclusion that the estimated training effects are very sensitive to sample specification and to the econometric method used.

In practice, however, would the econometrician ever draw this conclusion? I doubt it, because virtually every econometrician is likely to insist that not all the estimation schemes used in Tables 2 and 3 are equally useful. For example, the difference-in-differences estimator corresponds to the appropriate estimator for a model of earnings that contains person-specific fixed effects, which is commonly thought to be an appropriate description of individual worker's earnings. In Table 2, however, the non-experimental results in columns 6 and 7 that make use of this estimator are completely insensitive to the sample selected for the comparison group. The econometrician is likely, therefore, to conclude that the results are not sensitive to sample selection if the appropriate estimator is selected. Unfortunately, however, the econometrician would then select as

appropriate an estimate of the training effect that is 3.5 times as large as the effect based on the data using randomized trials.

There seem to be two remaining alternative approaches left for selecting an estimate of the training effect. One emphasizes the criterion of a priori reasonableness in the selection of a comparison group sample. This criterion is, to some extent, arbitrary, but it seems likely that some agreement could be reached. For example, in Table 2 I would expect that the PSID-3 comparison group would get the vote as the data set most likely to imitate the behavior of a randomly selected control group, although a regression adjustment for demographic characteristics might seem appropriate to many econometricians if the estimates were based on a comparison of post-training earnings only. As Table 2 indicates, selecting this comparison group leads to an estimate of the training effect that is invariant to the estimation method and that is nearly identical to the common set of estimates produced by the difference-in-differences estimator using all of the comparison groups in Table 2. As before, however, the estimate of a training effect that would be selected on this criterion is three and one-half times the effect produced from the data derived by randomization.

A second approach is to allow the data, by themselves, to tell us which is the appropriate estimate to select. This is accomplished by applying a series of tests of model specification. The most natural test to apply here would be a test of whether the regression-adjusted difference in earnings between the trainees and proposed comparison group is negligible in the pre-training period. This is a backcasting

test, much like the forecasting tests proposed by many econometricians. After all, "if the regression model were entirely successful we would expect the estimated pre-training differences to be negligible."[1] Applying this approach would clearly lead the econometrician to again select the PSID-3 comparison group which leads, as before, to an estimated training effect that is three and one-half times the effect estimated from the data using randomized trials.

It appears that this example is a telling practical demonstration of the principle that specification tests are merely necessary and not sufficient for purely empirical reasoning. A particular telling example of this point is contained in Table 2. After all, the estimators based on randomized trials in the first row of this table pass every specification test we might wish to apply (adjusted and unadjusted pre-training earnings differences are negligible, all four estimates of the training effect are nearly identical) but the estimators based on the PSID-3 comparison group data do too! Yet the estimated training effect from the non-experimental data analysis is three and one-half times the effect estimated from the experimental data analysis.

The results for males in Table 3 generally indicate negative training effects when the preferred econometric procedures are used, which is opposite to the sign of the result obtained from the randomized

---

[1] I wrote this sentence originally in Proceedings of the Twenty-Seventh Annual Winter Meeting, Industrial Relations Research Association, December 28-29, 1974. This specification test is a fundamental advantage of the use of longitudinal data that I pointed out at that time.

trials data. Here, then, the econometrician would underestimate the training effect. Sampling errors in Table 3 are much larger, however, and perhaps these are large enough to provide some warning that not much should be made of any of three results.

In sum, it appears that in the area of the analysis of training programs the econometric methods available may not be able to deliver the benefits that randomized trials offer. At a minimum, therefore, it appears that some experimental evaluation of these programs using randomized trials is desirable, if only to provide the occasional check on how well the non-experimental methods are working.

II.

> "Manpower training programs have not been run as if they were experiments in the past, and there is a great deal of difficulty and resistance in changing past practices. The argument against experimentation is usually shrouded in the rhetoric of morality, implying that deliberate exclusion of any groups from entrance into a training program to which they were entitled is unethical. The fact that this argument is entirely fallacious is demonstrated by two simple points. First, virtually all MDTA programs have offered substantial stipends to participants, so that there has normally been an excess supply of applicants. This means that some eligible trainees have always been excluded from training, and must have been excluded by some criteria. In response to this point it might be argued that those applicants who are accepted into training have been chosen by program operators on the basis of the likelihood that they would benefit from the program. There is, of course, no reason why this policy could not be continued in an experimental environment so long as the basis of the selection criteria were known and measureable, in which case the selection criteria could be controlled statistically. No doubt there would be resistance to requiring the practice of using explicit and objective criteria for entrance to the programs, but such a practice might be desirable on grounds of fairness even in the

absence of an experimental environment."[2]

I wrote the above words over ten years ago and, since that time, some progress have been made as LaLonde's work shows. Further progress will have been made if randomized trials are used again in the evaluation of training programs.

---

[2]Proceedings of the Twenty-Seventy Annual Winter Meeting of the Industrial Relations Research Association, December 28-29, 1974, pp. 252-260.

Table 1. Job Placement Rates in Dayton Treatment Groups.*

Group	*Sample Size Enrolled*	*Number Placed in Jobs*	*Percentage Placed in Jobs*
Tax Credit Voucher	247	32	13.0
Direct Rebate Voucher	299	38	12.7
Control	262	54	20.6
Total	808	124	15.3

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

Table 2

Earnings Comparisons and Estimated Training Effects for the National Supported Work AFDC Participants Using Comparison Groups From the PSID and the CPS-SSA[1]

Name of Comparison Group	Comparison Group Earnings Growth 1975-1979	NSW Treatment Earnings Less Comparison Group Earnings Pre-training year, 1975		NSW Treatment Earnings Less Comparison Group Earnings Post-training year, 1979		Difference in Differences: Difference in Earnings Growth 1975-1979 Treatments less Comparisons	
		Unadjusted	Adjusted[2]	Unadjusted	Adjusted[2]	Without Age	With Age
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
NSW Controls	2,942[4]	-17	-22	851	861	833	883
	(220)	(122)	(122)	(307)	(306)	(323)	(323)
PSID-1[3]	713	-6,443	-4,882	-3,357	-2,143	3,097	2,657
	(210)	(326)	(336)	(403)	(425)	(317)	(333)
PSID-2	1,242	-1,467	-1,515	1,090	870	2,568	2,392
	(314)	(216)	(224)	(468)	(484)	(473)	(481)
PSID-3	665	-77	-100	3,057	2,915	3,145	3,020
	(351)	(202)	(208)	(532)	(543)	(557)	(563)
PSID-4	928	-5,694	-4,976	-2,822	-2,268	2,883	2,655
	(311)	(306)	(323)	(460)	(491)	(417)	(434)

Notes: 1. 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 Column 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.

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

3. See Table 1.4 for definitions of the comparison groups.

4. Estimates are in 1982 Dollars. The numbers in parentheses are the standard errors.

5. The Comparison Groups are defined as follows:

(i) 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.

(ii) PSID-2: Selects from the PSID-1 group all women who received AFDC in 1975.

(iii) PSID-3: Selects from the PSID-2 all women who were not working when surveyed in 1976.

(iv) PSID-4: Selects from the PSID-1 group all women with children, none of whom are less than 5 years old.

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

Table 3

Earnings Comparisons and Estimated Training Effects for the National Supported Work Male Participants Using Comparison Groups From the PSID and the CPS-SSA[1]

Name of Comparison Group	Comparison Group Earnings Growth 1975-1979	NSW Treatment Earnings Less Comparison Group Earnings Pre-training year, 1975		NSW Treatment Earnings Less Comparison Group Earnings Post-training year, 1979		Difference in Differences: Differences in Earnings Growth 1975-1979 Treatments Less Comparisons	
		Unadjusted	Adjusted[2]	Unadjusted	Adjusted[2]	Without Age	With Age
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Controls	$2,063[4]	$39	$-21	$886	$798	$847	$856
	(325)	(383)	(378)	(476)	(472)	(560)	(558)
PSID-1[3]	$2,043	-$15,997	-$7,624	-$15,578	-$8,067	$425	-$749
	(237)	(795)	(851)	(913)	(990)	(650)	(692)
PSID-2	$6,071	-$4,503	-$3,669	-$4,020	-$3,482	$484	-$650
	(637)	(608)	(757)	(781)	(935)	(738)	(850)
PSID-3	$3,322	$455	$455	$697	-$509	$242	-$1,325
	(780)	(539)	(704)	(760)	(967)	(884)	(1078)

Notes:

1. 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 column 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.
2. The exogenous variables used in the regression adjusted equations are age, age squared, years of schooling, high school dropout status, and race.
3. See Table 2.6 for definitions of the comparison groups.
4. Estimates are in 1982 Dollars. The numbers in parentheses are the standard errors.
5. The Comparison Groups are defined as follows:
   (i) 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.

   (ii) PSID-2: Selects from the PSID-1 group all men who were not working when surveyed in the spring of 1976.

   (iii) PSID-3: Selects from the PSID-1 group all men who were not working when surveyed in either spring of 1975 or 1976.

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