# THE ROLE OF RANDOMIZED EXPERIMENTS IN EMPLOYMENT AND TRAINING EVALUATIONS

Rebecca A. Maynard, Mathematica Policy Research, Inc.
P.O. Box 2393, Princeton, New Jersey 08543-2393

This paper discusses the current evidence on whether randomized experimental designs are critical to successful, comprehensive evaluations of employment and training programs.[1] Randomized experimental designs were first introduced into social evaluations in the 1960s, and they remained popular through the mid-1970s. Beginning in the mid-1970s, social policy experimentation seemed to fall out of favor, having been replaced with evaluation methods that relied on the use of comparison groups. In the past few years, we have seen a shift back to randomized experiments, due in large part to some empirical research that has highlighted the limitations of and risks associated with comparison-group methodologies.

The first section of the paper presents a brief overview of the range of analytic methods used in evaluations of employment and training (or "manpower") programs and discusses the motivation for empirical research to assess the reliability of the various nonexperimental/comparison-group approaches. The second section of the paper describes an empirical test of alternative methods in which we used a data set from a randomized experiment--the Supported Work demonstration--to examine whether comparison-group methodologies would replicate the policy findings from the original Supported Work Evaluation (see Fraker and Maynard, 1987). The third section reviews the findings from some more recent research that focuses on whether randomized experiments should be used. The final section presents our conclusions and recommendations about the choice between experimental versus nonexperimental methods.

## I. BACKGROUND

Manpower programs have been and are evaluated based on a variety of analytic methods. Indicators of program outcomes--such as placement rates and placement wage rates--are commonly used, but are also commonly acknowledged as unreliable indicators of the overall effectiveness of manpower programs. A twist on the use of program outcome measures is to examine "before and after" measures of employment success, often adjusting somewhat for "normal growth trends" and environmental shifts. However, this before-after method has also not been widely accepted, because of the strong age-earnings trends in employment and earnings outcomes. A third approach that has gained widespread acceptance and has been used in the majority of evaluations of major manpower programs conducted to date is the comparison-group evaluation design. This approach entails selecting either _a priori_ or ex-post a sample of nonparticipants whose experience is used as a benchmark measure to determine the experience of program participants in the absence of the intervention. Comparison-group methods differ according to two factors: (1) the method used to select the comparison sample (e.g., random samples, cell matching, or statistical matching) and (2) the nature of the statistical controls used to account for non-program-induced differences between the participant and comparison samples (e.g., no controls, fixed-effects models, standard regression models, and biased selection models). Finally, a fourth approach entails randomized experimental designs. These experimental evaluation approaches vary primarily according to the point at which the randomization occurs and by what it means to be in a control group--for example, the no-treatment control groups such as are used in clinical trials, the control groups who receive the status-quo treatment, and the control groups who are offered alternative treatments.

The motivation for our study stemmed from four factors. First, the importance of having reliable program evaluations has widely been recognized. Second, using comparison-group methods of evaluation, if they work, offer several advantages, pertaining to the following: (1) ethical considerations associated with denying services to a randomly selected control group; (2) cost considerations in using existing data bases for drawing comparison groups; (3) the minimal burden imposed on program operations; and (4) the minimization of threats of treatment-group contamination, such as may occur in "saturation" experiments. A third and very important factor is the equivocal results from two major sets of manpower evaluations that have relied on comparison-group methods--those of the Comprehensive Employment and Training Act (CETA) and those of youth programs. For example, Barnow (1987) reviewed all of the major CETA impact evaluations and reported an unacceptably wide range of impact estimates which showed earnings gains that ranged from $400 to $2,000 per year for women, from -$3,000 to $2,000 for men, and from -$1,900 to $1,000 for youths. The National Academy study on youth employment programs (Betsey et al., 1985) concluded that only one study out of the hundreds undertaken on the basis of comparison-group methods yielded credible results. The fourth factor that led to our empirical analysis of the reliability of comparison-group methods is that we really did not have an objective method for assessing the reliability of comparison-group methods in advance; for that matter, the ability to assess them ex-post was also limited.

[1] This paper was presented at the ASSA Meetings, San Francisco, CA, August 17, 1987. The paper borrows heavily from co-authored work with Dr. Thomas Fraker, Mathematica Policy Research, Inc., and with Professor Robert LaLonde, University of Chicago and Council of Economic Advisors.

## II. ANALYTIC APPROACH

The analytic approach used in this empirical study entailed taking an experimental data set and determining whether we could identify comparison-group methods that would replicate the experimental results. We chose the Supported Work data set for our study for two reasons: (1) it included samples of youths and AFDC recipients, both of which are targeted by current employment programs; and (2) the data set included Social Security earnings data, thus enabling us to obtain comparable outcome measures from Current Population Survey (CPS) data files (see Hollister et al., 1984).

We explored different methods of constructing the sample, examined different analytic models, and explored several model-specification tests in an attempt to identify some good comparison-group methods. We generated comparison groups that included:

- Random CPS samples of cases that met basic target group eligibility criteria (e.g., youths and AFDC recipients)
- Cell-matched samples, where cells were defined by such factors as gender (youths only), preprogram earnings, changes in preprogram earnings, race/ethnicity, education, and age
- Statistically matched samples drawn on the basis of the closeness of fit on a predicted preprogram earnings measure

We then examined the comparison samples to determine how well they matched the control group (the group whom they were expected to replicate), compared program impact estimates generated with the comparison groups and the randomly selected control groups, and tested the sensitivity of the results to the specifications of the model.

## III. EMPIRICAL RESULTS

As shown in Figure 1, a reasonable correspondence earnings trends existed between

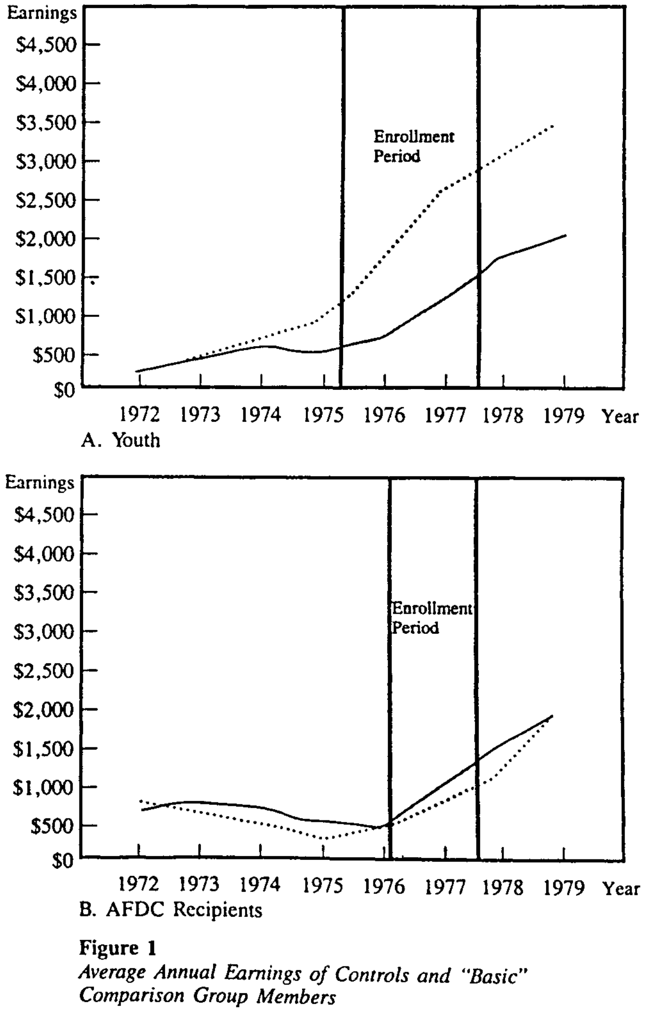

**Figure 1**
*Average Annual Earnings of Controls and "Basic" Comparison Group Members*

——— Controls ······ Comparison Group

the AFDC control and comparison groups, but not between the two youth groups. Comparison-group youths had much steeper age-earnings profiles than did control-group youths. Chow tests of preprogram earnings models showed no clear pattern in terms of the best comparison group for either sample. In fact, the test failed more often for AFDC recipients than for youths. Thus, the question that arises pertains to the comparability of the impact estimates generated with the control group and the various comparison groups, given these differences in the characteristics and employment trends of the groups.

When we examined our ability to replicate the experimental findings, the results were quite discouraging. As shown in Table 1, using both a basic earnings equation that controlled for the background and demographic characteristics of the sample at the time of their enrollment and a cell-matched comparison group, we estimated a large and significant negative impact on earnings for youths (-$700 to -$1,200 per year); the corresponding experimental estimates for youths ranged from significant positive ($313 per year) to zero impacts. For AFDC recipients, the comparison-group results were in fact quite similar to the experimental estimates (an approximately $1,400 annual earnings gain during the program period and a $350 to $500 annual gain in the postprogram years). However, the comparison-group estimates have much larger standard errors than do the control-group estimates, for both the youth and the AFDC samples.

In comparing the results generated with comparison groups that were constructed through different methods, we observed qualitatively similar results for the AFDC sample. All of the AFDC sample impact estimates generated with the comparison groups were positive, and they ranged from about 70 percent smaller to 130 percent larger than the experimental results. For the youth sample, however, the impact estimates were highly sensitive to the particular comparison group selected. For example, the experimental estimate of 1977 earnings gains among youths is a statistically significant $313 annual earnings gain. In contrast, the estimates based on six different comparison groups ranged from -$774 to

TABLE 1

EXPERIMENTAL VERSUS NONEXPERIMENTAL ESTIMATES OF PROGRAM-INDUCED ANNUAL EARNINGS EFFECTS:
"BASIC" COMPARISON GROUP AND ANALYTIC MODEL
(Standard Errors are in Parentheses)

| | Youth | | AFDC Recipients | |
|---|---|---|---|---|
| Year | Control Group | Comparison Group | Control Group | Comparison Group |
| 1977 | 313* | -668* | 1,423** | 1,560* |
| | (134) | (310) | (162) | (400) |
| 1978 | -28 | -1,191** | 505** | 537 |
| | (135) | (373) | (137) | (335) |
| 1979 | -18 | -1,179** | 351* | 257 |
| | (166) | (375) | (174) | (465) |
| Number of Individual Observations | | | | |
| Experimentals | 566 | 566 | 800 | 800 |
| Controls/comparisons | 678 | 2,368 | 802 | 909 |
| Number of Grouped Observations | | | | |
| Experimentals | 69 | 69 | 110 | 110 |
| Controls/comparisons | 87 | 112 | 107 | 73 |

NOTE: These results were estimated on the grouped observations using weighted least squares. See further details on the specification of the model in Fraker and Maynard (1987).

*Statistically significant at the 5 percent level.
**Statistically significant at the 1 percent level.

$166, of which only the three very large negative estimates were statistically significant (see Fraker and Maynard, 1987, Table 3, for details on these results).

In contrast to the results of our examination of alternative comparison-group construction procedures, we found that the results were much less sensitive to the specification of the model. The one exception was some evidence of sensitivity to using a fixed-effects model specification in contrast with others (see Fraker and Maynard, 1987, Table 4). In addition to varying the specification of the analytic model, we pursued the importance of several other analytic factors, including the use of grouped as opposed to individual data, the use of weighted versus unweighted data, and the variables included in the analytic models. The results indicated that these other factors did not affect the conclusions about the efficacy of using comparison groups.

## IV. COMPLEMENTARY RESEARCH

Ours was one study. It seems prudent to offer somewhat reserved conclusions until there is corroborating evidence. There are two such studies to note in this regard.

In an independent analysis, LaLonde (1986) examined the quality of impact estimates generated from comparison-group methodologies, also based on the Supported Work demonstration data. LaLonde defined comparison groups for two subsets of the Supported Work sample--the AFDC target group and males who enrolled in the youth, ex-addict, or ex-offender target groups--by drawing comparison groups from the 1976 CPS sample. Using these comparison groups and the Supported Work control group, LaLonde estimated program impacts on annual earnings based on several analytic models. LaLonde's results corroborate several of the findings from our study. First, he found that the analytic models significantly affected the impact results when the comparison samples were used. Second, he found that comparison groups worked better for AFDC recipients than they did for males. LaLonde's results also demonstrate two other important points: first, controlling for preprogram earnings differences is very important, and, second, including a nonlinear control for the program participation decision will tend to reduce bias relative to other model specifications.

A second complementary study has been conducted by Heckman and Hotz (forthcoming). Heckman and Hotz have undertaken theoretical work on specification tests that can be used to judge comparison-group methods, and they have attempted to validate these tests based on the same Supported Work data set used in our analysis. The range of specification tests examined included tests of significant intercept shifts between experimentals and controls; tests of the equality of coefficients in outcome equations estimated for control and comparison groups; and tests of whether the inclusion of additional years of pretraining earnings in a "random growth" model estimated with comparison samples will yield reliable impact estimates. The key findings from this study, summarized in Table 2,

TABLE 2

COMPARISON OF EXPERIMENTAL AND NONEXPERIMENTAL ANNUAL EARNINGS IMPACT ESTIMATES BASED ON MODELS PASSING ALL SPECIFICATION TESTS

| | 1978 | | 1979 | |
|---|---|---|---|---|
| Model | Experimental Estimate | Nonexperimental Estimate | Experimental Estimate | Nonexperimental Estimate |
| **YOUTH** | | | | |
| Random Growth Model A | -11.4 | -23.2 | -30.9 | -85.2 |
| (B1 + B2) | (306.4) | (476.5) | (351.3) | (546.8) |
| Random Growth Model B | -98.9 | -614.4 | -83.7 | -700.9 |
| (B1) | (298.9) | (43.10) | (361.9) | (509.6) |
| Random Growth Model B | -30.4 | -624.4 | 11.1 | -806.5 |
| (B1 + B2) | (298.9) | (497.4) | (359.3) | (586.1) |
| Modified Random Growth | 64.1 | -241.1 | 100.0 | -319.2 |
| Model B (B1) | (323.3) | (456.9) | (351.4) | (506.0) |
| **AFDC RECIPIENTS** | NO MODEL PASSED ALL SPECIFICATION TESTS | | | |

SOURCE: Tables 5-11 of Heckman and Hotz (forthcoming).

NOTE: Standard errors are in parentheses.

reveal that all model specifications are rejected for females, and that four model specifications are supported for youths. However, the estimated impacts based on the "accepted" comparison-group models for youths vary qualitatively (ranging from -$23 to -$806 per year), and many differ substantially from the experimental estimates, although it should be pointed out that all of these estimates have large standard errors.

## V. CONCLUSIONS AND RECOMMENDATIONS

We draw four very important conclusions from our empirical research and from the results of the two complementary studies:

1. Experimental methods are very important for employment and training evaluations.
2. Further work on statistical methods for correcting for biased selection is critical.
3. We must begin to collect better information on the selection process so as to improve the quality of nonexperimental evaluations.
4. Practitioners, policy analysts, and econometricians/statisticians must work more closely on this problem.

Worse than not evaluating our major manpower programs is evaluating them and ending up with equivocal or unreliable findings. In our judgment, comparison-group methods are risky--we certainly cannot judge *a priori* whether the results of such studies will be reliable, and, if we are to believe the experiences of the past, we may not be able to judge the reliability of the results ex-post.

## REFERENCES

Barnow, B. "The Impact of CETA Programs on Earnings: A Review of the Literature." *Journal of Human Resources*, vol. 22, no. 2, 1987.

Betsey, C., R. Hollister, and M. Papageorgio, editors. *Youth Employment and Training Programs: The YEDPA Years*, Washington, D.C.: National Academy Press, 1985.

Fraker, T., and R. Maynard. "Evaluating Comparison Group Designs with Employment-Related Programs." *The Journal of Human Resources*, vol. 22, no. 2, Spring 1987.

Heckman, J., and J. Hotz. "On the Use of Nonexperimental Methods for Estimating the Impact of Manpower Training Programs: Re-Evaluating the Evaluations." Chicago, IL: NORC, 1987 (forthcoming in *Journal of Human Resources*).

Hollister, R., P. Kemper, and R. Maynard. *The National Supported Work Demonstration.* Madison, WI: University of Wisconsin Press, 1984.

LaLonde, R. "Evaluating the Econometric Evaluations of Training Programs with Experimental Data." *American Economic Review*, vol. 76, no. 4, 1986, pp. 604-20.

LaLonde, R., and R. Maynard. "How Precise Are the Evaluations of Employment and Training Programs: Evidence from a Field Experiment." In *Evaluation Review.* Beverly Hills, CA: Sage Publications, 1987.