DISCUSSION PAPER SERIES

IZA DP No. 8118

The Early History of Program Evaluation and the U.S. Department of Labor

Orley C. Ashenfelter

April 2014

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

The Early History of Program Evaluation and the U.S. Department of Labor

Orley C. Ashenfelter
Princeton University
and IZA

Discussion Paper No. 8118
April 2014

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
E-mail: iza@iza.org

IZA Discussion Paper No. 8118
April 2014

ABSTRACT

The Early History of Program Evaluation and the U.S. Department of Labor*

This paper contains a review of the early history of program evaluation research at the US Department of Labor. Some broad lessons for successful evaluation research are summarized.

JEL Classification: B4, C21, J8

Keywords: program evaluation, training programs, active labor market programs

Corresponding author:

Orley C. Ashenfelter
Industrial Relations Section
Firestone Library
Princeton University
Princeton, NJ 08544-2098
USA
E-mail: c6789@princeton.edu

* Orley C. Ashenfelter is the Joseph Douglas Green 1895 Professor of Economics at Princeton University and a former Director of the Office of Evaluation of the U.S. Department of Labor. These comments were prepared on the occasion of the 100th anniversary of the U.S. Department of Labor.

My contact with the Department of Labor began more than 40 years ago, so it is a little alarming to consider that the Department is merely 100 years old! These remarks are addressed to those early days in the history of program evaluation 40 years ago, days that were both challenging and, in retrospect, more influential than anyone in the Department imagined.

The Early Evaluation of Training Programs

In 1971 and 1972, Washington, D.C., was a hotbed of discussion on the effectiveness of government programs that had been implemented in the "war on poverty" and in response to riots in Washington and elsewhere in the late 1960s. One of the most controversial programs was called the Manpower Development and Training Act (MDTA), and its evaluation provided my introduction to the extraordinarily difficult problems of inference in the evaluation of social programs. The MDTA, like dozens of programs in the United States and Europe, was intended to reduce structural unemployment and, in doing so, to increase the incomes of those who participated. The question many people asked: Did the program do this?

To my astonishment, in early 1972 I was offered a civil service position in the U.S. Department of Labor in which I was to direct an Office of Evaluation whose sole purpose was to ask and answer this and some related questions. The experience was quite exhilarating. Much to my surprise, the office was left to do its work without political interference, and it continued to survive for another 10 years, although in a much-reduced capacity as time passed.

There are three reasons why program evaluation is so difficult, and as shorthand I will refer to these as problems of 1) data collection (data), 2) exogeneity of treatment (exogeneity), and 3) existence of treatment (existence). The appraisal of the MDTA program raised them all, but what made this particular program evaluation of interest was that the "data" problem had, in part, been solved. It is difficult today to appreciate the enormity of this breakthrough, and maybe

only those who lived with the social sciences in this early period can appreciate it. What had been created, and could be used for the evaluation of the program, was a full-scale longitudinal data set on each individual who was part of it.

Let me explain just how we coped with each of the problems of evaluation I have noted in our struggle to understand the effect of this program on its participants and the labor market.

Data

One of the key problems in labor economics is that we cannot explain much of the difference in individual outcomes in the labor market. This heterogeneity is extremely well documented in labor markets, but it is now widely understood to be the case even in financial and product markets. We may know that the average person with a university degree earns more than the average person without such a degree, but much variability remains unexplained within each group. The result is that to test the effect of any program on earnings or unemployment we must have large samples of data, and typically because of the problem of "exogeneity" we also need data that cover the program members before and after they entered the program. These are called "longitudinal" data.

There are two ways to obtain data. You can collect it yourself (I have done this, it is certainly the hard way to go!), or you can find a way to take advantage of data produced by others, perhaps even data produced for another purpose. In this case we actually obtained data from two separate governmental sources and linked them together. One source included the program records on those people who had entered the training program that were maintained by the Department of Labor, and the other source was the federal Social Security data collected for all workers on a quarterly basis. It was this remarkable data set that put in motion an extremely sophisticated effort to solve the other problems I noted above, an effort that continues today.

Exogeneity

Of course, knowing the employment and earnings history of the program participants does not solve the key problem of inference. To what are we to compare this experience? If the program had been operated with random assignment (in subsequent years some programs were operated in this way because of what we learned), we could simply compare those assigned to treatment with those not assigned. But this was not possible. Instead we used a comparison with a random sample of the overall population of workers.

The key thing learned from this comparison was that the program participants had lower earnings, both before and after the program, than the comparison group. This automatically made it clear that the analysis would not meet the highest standards for credibility. This also suggested that the participants should be compared with themselves instead of with the comparison group alone, and with longitudinal data that is precisely what was possible.

To control for overall changes in the labor market, however, it was critical to also have a second benchmark, and the comparison group provided just that. In short, the difference from the pre to the post period in earnings for the treatment group could be compared against the difference from the pre to the post period for the comparison group. This is the origin of the "difference-in-differences" method that has come to dominate discussions in labor economics and, in fact, is found in much of the empirical study of economics more generally.

There are, in fact, two ironic features about the widespread adoption of the difference-in-differences approach to the evaluation of programs in economics more generally. First, a key reason why this procedure was so attractive to a bureaucrat in Washington, D.C., was that it was a transparent method that did not require elaborate explanation and was therefore an extremely credible way to report the results of what, in fact, was a complicated and difficult study. From a

technical point of view, a difference-in-differences study controls for fixed effects for individuals, and thus heterogeneity across people, and for fixed effects for time periods, and thus variability over time. It was meant, in short, not to be a method, but instead a way to display the results of a complex data analysis in a transparent and credible fashion.

Second, as it turned out, things were considerably more complicated than this analysis indicated. Because the participants had a pattern of earnings that tended to decline dramatically prior to program entrance, a simple difference-in-differences produced quite dissimilar results depending on what precisely was called the "pre-treatment" period. My own conclusion was that randomization was the only transparent and credible cure for this problem. An early summary of what we learned, and my plea for randomized trials, appeared in a paper I presented at the Industrial Relations Research Association Meetings in 1974.[1] It is hard to appreciate today just how controversial this proposal was.

Existence

It is often surprising to learn that the mere existence of a program needs to be established empirically. After all, some will ask, surely a law has been passed, or money has been allocated, and doesn't this establish that a program is available? In fact, this problem is much more difficult than it appears at first blush. Consider, for example, training programs. Although the government may subsidize these, and we can surely count up the number of participants, how do we know that the training provided would not have been provided by private employers? When we investigate the effect of a minimum wage on employment, how do we know that the law, in fact,

[1] See Ashenfelter 1975a and 1975b; the paper titled "The effect of manpower training on earnings: Preliminary results" was presented at the 27th Annual Meeting of the Industrial Relations Research Association, 1974. See also Robert LaLonde (1986) in which he expressed his influential support for this position, comparing results obtained by using randomized trials with those used by various ingenious comparison groups.

changed wages? I think one of the most critical lessons learned from the program evaluation literature is the necessity of first showing that a program exists.

A Word about Theory

In this discussion I have said nothing about the role of economic theory in the design of natural experiments. As in all sciences, data analysis has two roles: description and hypothesis testing. The early program evaluation literature was aware of the usefulness of scientific theories for suggesting treatments to test in field experiments, just as those who study natural experiments today are often motivated by economic theories. It is no doubt harder to provide sharp tests of economic theories in the field than in the laboratory, but field tests are one step closer to inferences that may be externally valid. Moreover, economists can treat differential treatment effects as something to be explained by an economic theory, not merely as a nuisance.

Some Lessons

When I first became interested in the credible, transparent evaluation of social programs, very few others shared these interests or carefully thought through the key elements in an evaluation design. Today, it has become commonplace to see literally hundreds of studies[2] that follow the steps many of us stumbled onto—data collection, an empirical appraisal of whether a program exists, and an attempt to define an exogenous treatment—which is now called "evidence-based policy evaluation." Program evaluation has a long history in the Department of Labor, and its spread to many other program areas and countries can be traced directly to those early days in the Department. I hope that similar credible, transparent evaluations will continue to spread to other areas of government behavior and spending that are so ripe for a quantitative appraisal.

[2] See Card, Kluwe, and Weber 2010 for a summary review of 200 such studies.

References

Ashenfelter, Orley. 1975a. Manpower training and earnings. *Monthly Labor Review* 98(4): 46–48.

Ashenfelter, Orley. 1975b. The effect of manpower training on earnings: Preliminary results. In Proceedings of the 27th Annual Winter Meeting of the Industrial Relations Research Association, pp. 252–60. Madison, WI: Industrial Relations Research Association.

Card, David , Jochen kluwe, and Andrea Weber. Active Labor Market Policy Evaluations: A Meta-Analysis. *The Economics Journal* 120(Novermber): F452-F477.

LaLonde, Robert J. 1986. Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review* 76(4): 604–20.