MICHAEL H. MOSKOW HONORARY PAPER SERIES

Evidence Based Policy Evaluation and the Design of Natural Experiments: The Origins of Program Evaluation in the Department of Labor and Beyond

Orley C. Ashenfelter, Joseph Douglas Green 1895 Professor of Economics, Princeton University

April 2019

The title above is taken directly from one of the most famous and influential scientific books of the twentieth century, R.A. Fisher's *Design of Experiments*. Fisher struggled with, and analyzed, what is now considered the "gold standard" method for making causal inferences. This method, which had been hinted at in scientific work for centuries, finally reached its full development in the twentieth century. Designed to produce highly credible evidence in complex situations, it is based on the idea that we determine the causal effect of a treatment or intervention by assigning randomly some fraction of the units we wish to influence and reserving the remainder as a control group. In medicine, it is said that we test a drug or procedure by using randomized clinical trials, but Fisher studied primarily agricultural experiments hence the naming "randomized field trials."

The key point is that these are experiments that take place in the real world — not a laboratory — and provide the final, conclusive test of the efficacy of a treatment or intervention. An interesting aspect of Fisher's work is that it was always motivated by actual problems of experimental inference and evolved as a fundamentally practical analysis. Fisher's lasting contribution, now taken for granted by virtually all scientists, is a set of methods for determining when observed differences are unlikely to be due to chance alone.

Much of what I have tried to do in my own research in the last forty years is to find some way to implement highly credible methods for the study of important and controversial problems of inference in economics. These methods, which tend to be opportunistic because they differ with the problem being studied, have come to be called "natural experiments." Natural experiments are sometimes randomized trials (jokingly called "unnatural experiments" by a few of my colleagues), but often they

must depend on some method that falls short of this gold standard. The key point is that these are analyses of what happens in practice, not just in theory, and the emphasis is on the credibility of the results.

Of course, most of the important methodological problems of economics, and especially labor economics where I have often worked, differ from agriculture and medicine. The differences can be categorized into three groups:

1) We often do not have the data necessary to study a problem.

2) We often cannot, and would not wish to, control the division of the units to be studied into control and treatment groups.

3) Finally, even in the best of circumstances we cannot guarantee the integrity of the treatment assignment, so we must make provision for a difference between what we intend to measure and what we do measure.

In these remarks, what I would like to do is to share with you a few personal comments about how I was "converted" to the view that credible, that is, genuinely believable, inference was the key to progress in economics.

The Early Evaluation of Training Programs

In 1971 and 1972, Washington, DC, was a hotbed of discussion of the effectiveness of government programs that had been implemented in the "war on poverty" and in response to riots in the capital 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 was, did the program do this?

To my astonishment, in early 1972 I was offered a civil service position in the US Department of Labor in which I was to direct an Office of Evaluation, whose sole purpose was to ask and to answer this and some related questions.[i] 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.

You may recall that 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 a part of it.

Let me explain just how we coped with each of the problems of evaluation I have noted 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 there is much variability that 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 covers 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.

However, to control for overall changes in the labor market, 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 so-called "difference in differences" that has virtually 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, DC, 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 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 approach produced quite different 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, but your view on this issue may, in fact, be different from mine.[ii]

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 there is a program? 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 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, 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. There is nothing about Fisher's work that suggests he was unaware of the usefulness of scientific theories for suggesting treatments to test in his field experiments, just as those who study natural experiments 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. Just as Fisher recognized that the heterogeneity of treatment effects may have scientific explanations, economists can

also 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 there were very few others who shared these interests or who carefully thought through the key elements in an evaluation design. Today, it has become commonplace to see literally hundreds of studies 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. Today this is called "evidence-based" policy evaluation.

---

[i] Michael Moskow, a former professor himself, and subsequently president of the Chicago Fed, was my immediate superior and, I suspect, protector. Laurence Silberman, undersecretary of the Department of Labor at the time, and subsequently a US Circuit Court judge, played an important role also. It is worth considering that this was really the introduction of "evidence-based evaluation" to a government agency not always used to it.

[ii] Robert Lalonde's work ("Evaluating the Econometric Evaluations of Training Programs with Experimental Data," *American Economic Review*, September 1986) comparing results obtained by using randomization trials with those used by various ingenious comparison groups provides the first careful discussion of this issue.