1995

Design of Social Experiments

Christopher J. O'Leary
W.E. Upjohn Institute, oleary@upjohn.org

Robert G. Spiegelman
W.E. Upjohn Institute

Design of Social Experiments
Presentation to Hungarian Delegation
Upjohn Institute, June 14, 1995
by
Christopher O'Leary
Robert Spiegelman

Chris O'Leary and I would like to talk to you about field experiments in the social sciences and the possible application of this evaluative methodology to test innovations in Hungary's UI system. Though complex and relatively expensive, field experiments can provide answers to critical policy questions with a degree of reliability not available from other methods. In addition, since fewer caveats are required in presenting outcomes, experimental results also enjoy greater political credibility.

What constitutes a field experiment in the social sciences? Although the options for experimentation are large, two conditions must be met for an experiment to exist:

1. there must be some imposition in the socio-political environment that can be called a "treatment", and
2. there must be random assignment of eligible participants to the treatment category and to control status.

If these two conditions are met, then the experimental effect is measured as the difference in outcome and/or behavior (depending on the nature of the experiment) between the mean values of the experimental and control groups. This essentially "model free" method of evaluation is politically appealing, but is not always realized in practice, because small samples often result in differences in group composition that effect the results, requiring regression modeling even for experimental evaluations.

It is true that an ordinary least squares regression of the dependent variable of interest (denoted by y) on a constant term

and the treatment variable (T, which equals one for individuals assigned to treatment status and 0 for individuals assigned to control status) will yield unbiased estimate of the effect of the treatment on y. That is,

$$y = b_0 + b_1T + e.$$

In this equation b_1 yields an unbiased estimate of the treatment effect. If the sample sizes are large enough, it will also be a reliable estimate of the population response to the program; however, for smallish samples, differences can arise in the composition of the treatment and control samples, necessitating an equation of the following form in order to arrive at a correct estimate of y for the population:

$$y = b_0 + b_1T + c_1x_1 + \ldots + c_nx_n + e,$$

where the x's represent a set of characteristics of the samples that could, if they differed between the two groups, cause b_1 to be incorrectly estimated for the population.

Now I would like to describe the set of tasks that constitute an experimental design. In this presentation I will reference an experiment in UI work search policy in the State of Washington with which the Institute was involved and that could be relevant to Hungary.

A. Establish experimental goals

The experimental goal is to measure the impact on outcomes and/or behaviors of one or more program initiatives. The control group, against whose outcome the experimental program is being measured, usually represents current practice, and the "treatments" are modifications of current practice. In the Washington Work Search Experiment, three alternatives to current practice were to be investigated. The existing work search program consisted of a requirement that claimants make and document at least three employer contacts per week, and come to the office for an eligibility review interview (ERI) 13-15 weeks

after filing for benefits if still unemployed. This program, termed Treatment Group B, served as the control group. The three new treatment groups were:

1. Group A--Elimination of work search reporting requirements and submission of biweekly claims forms. UI benefit checks would be automatically issued until the claimant informed the office that a job had been obtained. Treatment A was to determine if the reduction in administrative costs would be sufficient to compensate for the costs of any lengthening in the unemployment spell.

2. Group C--Individualized work search requirements. Claimants with different characteristics, from different industries, and at different points in the unemployment spells would face different sets of requirements for numbers of employer contacts and for obtaining ERI's.

3. Group D--More intensive job-search assistance. Members of this groups were called for an ERI after four weeks of unemployment, were required to attend a two-day job search workshop, and after another three weeks of unemployment were required to engage in ten hours of telephone contact with employers.

The overarching goal of the experiment was to determine the cost-effectiveness of alternative work search policies in the Unemployment Insurance (UI) program[1].

B. Determine Eligibility Criteria

---

[1] The results of this experiment showed that elimination of work search reporting (Group A) was a disasterous option, reducing administrative costs only slightly and sharply increasing the weeks of insured unemployment. On the other, more intensive job-search assistance (Group D) decreased the weeks of insured unemployment by about one-half week and was cost-effective. However, the investigators concluded that the decrease was due to less filing for benefits rather than to more rapid reemployment.

Before discussing the specific eligibility requirements for the work search experiment, I'd like to suggest some general considerations. First, to obtain the most robust statistical results from a sample of given size, the sample population should be relatively homogeneous and should be selected from the population categories most likely to respond to the experimental treatment. To assure that the experiment can provide estimates of the effect of an actual program, the characteristics of the experimental groups must be the same as those of the entire population that might paricipate in the program.

A second consideration is more subtle, dealing with an issue of external reliability of the experiment. The conditions under which the experimental sample is enrolled must replicate the conditions that will face the eligible population in a full program. For example, a UI bonus experiment run in New Jersey was fatally flawed, because it enrolled UI claimants into the experiment and informed them of the bonus offer only after seven weeks of unemployment. Since unlike the experiment, the existence of the program will be widely known, this condition cannot be replicated in a program, and the population of claimants with seven weeks of unemployment will differ because of this knowledge. This external validity problem could have been avoided by letting claimants know at filing that after seven weeks of unemployment they will become eligible for a bonus offer and additional services.

In the effort to be cost-effective, some experiments focus on sub-groups of the population of interest. This decision can be risky, because the selected sub-group may not always be one of policy interest. As an example, one of the bonus experiments in the US was limited to "displaced workers", defined as those laid off after three years of continued employment with the same employer. The problem is that the definition of "displaced worker" in the US changed, making the results of this experiment

of little interest.

Alternatively, in the Washington Work Search Experiment, all individuals filing a new initial claim for benefits and eligible to receive benefits were assigned randomly to one of the four treatment groups. This large group continues to remain of policy interest.

C. Treatment Design

Treatment design, the heart of the experiment, may be categorized as entailing three decisions: (1) The number of program options to be tested; (2) the architecture of each treatment; and (3) the range over which to test each option.

1. Number of program options:

For the Washington Work Search Experiment, three new program options were tested. This was an unusual situation in which options could be tested that were both more and less stringent than current practice. Either legally or practically, this is not always possible. For instance, in the US it is not possible to test a UI option in which benefits are less than those available in the regular program.

In the New Jersey UI experiment, bonus offers and work search assistance programs were tested jointly. In the Washington bonus experiment, only variation in the bonus offer was tested. For Canada, we designed an experiment in which bonus offers and wage supplements would be tested together, with the goal of determining for a given cost, which option would be more effective in reducing covered unemployment.

In testing combinations of programs, such as work search assistance and bonus offers, it is important that the structure of the experimental treatments permit estimation of separate as well as combined effects. It is not necessary that all possible combinations be tested. Lets consider an example. If there were

three options being tested, namely a low bonus offer, a high bonus offer, and work search assistance, there are five potential treatment cells, as follows:

low bonus only
high bonus only
work search assistance only
low bonus/work search assistance
high bonus/work search assistance.

The experiment could comprise only the last three treatment options. However, this implies the extrapolation of results to combinations not tested, which will provide correct estimates of the treatment effect only if the effects are linear and additive. A four cell option adds further information about the treatment effects of a bonus offer with and without work search assistance.

2. The architecture of the treatment:

The bonus offer has three parts:

-the amount of the bonus offer, which may be a fixed dollar amount or a percent of the Weekly Benefit Amount;

-the qualification period, i.e., the elapsed time to return to work, which may be in fixed number of weeks, or percent of entitled duration;

-the reemployment period, i.e., the time after qualifying the claimant must remain fully employed, usually a fixed number of weeks.

Wage supplement has four parts:

-the amount of the supplement, which may be a fixed number of dollars per unit of time, or a percent of the earnings/wage gap;

-the base for supplementation, which may be total hours worked, or weekly earnings, or some sub-set of earnings or hours;

-qualification period, the same as for the bonus;

-duration of supplementation, may be a specific number

of weeks or may be until a pool of supplementary funds are exhausted.

Work search requirement/assistance:

-the number of parts can vary considerably. In the Washington Work Search Experiments, the following components were varied in one or more of the treatments--number of employer contacts per weeks, timing and frequency of eligibility review interviews, use of telephone banks, and use of job search workshops.

3. The range over which to test the option:

-the range issue essentially arises in connection with financial incentives, such as a bonus offer or a wage supplement. The decisions to be made are: the minimum and maximum values that are of policy interest, and the number of treatment cells within this range that would capture the possible non-linearities and that meet criteria regarding sample size requirements and budget constraints.

D. Determining Sample Size and Site Selection

In arriving at a final decision regarding the size of the sample for the experiment, there are three issues to consider: (1) the sample size per cell, (2) the number of treatment cells; and (3) the number of sites in which to run the experiment.

1. Sample size per cell:

The required sample size per cell depends upon the expected treatment effect and the desired level of statistical significance and power. Statistical significance refers to the probability of accepting an alternative hypothesis that is false (the usual alternative hypotheses is that the effect of the

treatment is not zero). The power of the test is the probability of accepting an alternative hypothesis that is true.

Setting the statistical significance criterion at .05 (two tail test) and the power at .8, the required sizes per treatment cell for various levels of the effect measure are:

Effect Measure	Estimated Effect Weeks of Unempl.	Sample Size per cell
.05	.5-.6	6,300
.075	.75-.9	2,800
.10	1.0-1.2	1,600
.15	1.5-1.8	700

source: Cohen, Jacob (1977), Statistical Power Analysis for the Behavioral Sciences, Revised Edition, Academic Press, pp. 53-59.

The effect measure is the treatment impact normalized by the standard deviation of the distribution of the effected variable in either the treatment or control cell. In the example above, the effect measure is the treatment effect on weeks of insured unemployment divided by the standard deviation of the distribution of weeks of insured unemployment in a cell. The effects in the table represent a range from about one-half week to 1.8 weeks, with a standard deviation of 10 to 12 weeks. This is within the range of all the bonus experiments and the Washington Work Search Experiment.

Of great interest here is the exponential increase in sample size requirements as the expected impact of the treatment falls from one and one-half to one-half week. The latter was the effect of Treatment D in the Washington Work Search Experiment, the former was experienced by one set of claimants in one of the four

bonus experiments.

2. Number of Cells per Site and Number of Sites

These are determined together and reflect the conjunction of experimental requirements and budget constraints. As noted above, not all combinations of treatment options must be offered. This particularly applies in a multi-site experiment, where some basic treatment can be carried out in all sites, but other treatment options are randomly assigned to some of the sites.

The selection of sites is a critical decision which must be consistent with overall experimental goals. If the goal is to determine the average effect of a program in Hungary, the required number of sites would be considerably fewer than if the goal was to determine the effect in each county separately. If economic conditions differ across counties, then more experimental sites are needed to capture the effect of economic conditions on outcomes.

Lastly, there must be a model for selecting a set of sites that optimally meet the criteria. The Upjohn Institute has developed a model based on "nearest neighbor" concepts to select sites (used by the State of Minnesota for a workfare experiment).

E. Constructing the Budget

If there is a budget amount that cannot be exceeded, then it is best to determine if any experiment that meets the information requirements set for the experiment can be designed within that budget. If there is flexibility in the budget, then a range of more and less expensive experiments can be proposed, with the

final decision representing a decision as to how much information you can afford to buy.

In arriving at total cost, the following parameters must be estimated:

1. the cost of service to recipient--for a work search experiment, these are the administrative costs of providing the additional services and/or monitoring the added work search requirements. For a bonus experiment, the costs include the payment of bonuses.
2. the take up rate--the proportion of claimants assigned to treatments who will actually participate. This is very important for a bonus experiment, but much less so for a work search experiment, in which all enrollees may be required to participate. Costs per enrollee are the costs of service per recipient times the take-up rate.
3. the experimental costs--costs of operational materials, such as a procedures manual and forms, costs of installing and operating the experimental data system, costs associated with staff training and costs of evaluation. Many costs vary in proportion to the number of pariticpants, but other costs are fixed or semi-fixed.