Chapter 3

Statistical Approaches to the Evaluation Problem: Randomized Social Experiments

3.1 Introduction

Recent academic debates pit two alternative solutions to the evaluation problem. The first one is the *non-experimental* or *econometric* approach that uses a variety of microdata sources, statistical methods and behavioral models to compare the outcomes of participants in social programs with those of non-participants. The second is the *experimental* solution.

A social experiment emphasizes the researcher's control on the variables under investigation and over the environment in which those variables are observed. In a typical scientific experiment, the investigator simply introduces a change in a controlled environment and observes the effect of the change on the material or organism under study. Of course, reliable measurement of the effect requires some basis for comparison. This basis is not always a simple matter.

To simplify, let us limit the discussion to the case of a training program. The analyst needs to measure the effect of training to participants, that is to compare the gains of participants with those of non-participants. To proceed, he has to construct an appropriate comparison group, that is a group consisted with non-participants with characteristics (attributes) comparable to those of participants in order to attain a meaningful comparison procedure. The only fully satisfactory method of achieving such comparison equivalence is to assign subjects to the two groups "completely at random". This kind of observational study is commonly referred to as a *randomized social experiment*.

In this chapter the “experimental” method is going to be described analytically. We will describe how social experiments are implemented in practice and the assumptions that have to be satisfied. The advantages and limitations of that method are critically reviewed at the end of this chapter.

3.2 Historical Review

Social experimentation dates back more than half a century. The original case for social experimentation took as its point of departure the Havelmo’s (1944) *social planning paradigm* that suggested causal inference in statistics based on randomized experiments. Greenberg and Shroder (1991) mention that in a recent catalogue of social experiments found more than 90 separate fields trials involving a wide variety of distinctive research areas, including health insurance, prisoner rehabilitation, labor supply, worker training and housing subsidies. Table 1 presents some of the major recent experiments (taken from Burtless, 1995).

Great real resources were devoted to experimentation in the 1970s and early 1980s. The large-scale social experiments begun in the 1960s and 1970s and were ambitious and costly attempts to estimate basic behavioral parameters –the income and price elasticities of labor supply and housing demand functions and the elasticity of demand for health care in response to alternative insurance arrangements. These lavish experiments generated hundreds of research reports and many articles in leading scholarly journals. *Although recent experiments have been much more numerous, they have also been narrower in focus, less ambitious and less likely to yield major scholarly contributions* (Heckman, LaLonde and Smith, 1999).

Recent years have witnessed an increasing use of experimental designs, basically on evaluation of employment and training programs in North America as well as in Britain, Norway and Sweden. However, some prominent economists have grown disenchanted with this research tool and have challenged the value of experiments in answering central questions about human behavior and policy effectiveness. Criticisms of such experiments by social scientists, if loud and persistent enough, can affect the willingness of policymakers to support this kind of study. Politicians are naturally suspicious of the research method involving experimentation. It is important for them to understand the strengths as well as the limitations of this unique research tool.

Table 3.1: Major Recent Experiments

Experiment	Target Population	Tested Treatment	Publications
Negative Income Tax (NIT) Experiments (1968-1978)	Low- and moderate Income families headed By non-aged adults	NIT plans with Alternative income Guarantees and tax rates	Keeley et.al. (1978); Burtless and Hausman (1978)
Housing Allowance Demand Experiment (1973-1977)	Low- and moderate Income families	Alternative income supplement plans designed to help low-income households pay for housing costs	
RAND Health Insurance Experiment (1974-1982)	Non-aged low- and Moderate income Persons and families Living outside of Institutions	Health insurance plans that varied over two dimensions: upper limit on out-of-pocket Medical expenses and Co-payment rates Ranging from 0% (free Care up to 95%	Manning et. al. (1987)
Electricity Time-of-Use Pricing Experiments (1975-1981)	Residential consumers Of electricity	Alternative pricing Schedules for electricity in which prices vary by time-of-day or season of year	Caves and Christensen (1980)
National Supported Work Demonstration (1975-1980)	Long-term AFDC Recipients; former drug Addicts; ex-offenders; Young school dropouts	12-18 months of structured work experience and on-the-job training, using peer-group support and sympathetic supervision	
MDRC Work-Welfare Experiments (1982-1988)	AFDC applicants and Recipients	A variety of voluntary and mandatory work-oriented programs, Including job search, Skills training and Unpaid public Employment	
National Job Training Partnership Act (JTPA) Study (1986-1994)	Disadvantage adults and out-of-school youth who enroll in programs funded under Title IIA of JTPA	Job search assistance, classroom training, on-the-job training and other forms of training financed under JTPA	Heckman and Smith (1993)

3.3 Experimental Designs

The critical element that distinguishes social experiments from all other methods of research is the random assignment of meaningfully different treatments to the observational units of study. In the context of social science, an experiment takes place outside a laboratory setting, in the usual environment where social and economic interactions occur.

In the simplest case, which is going to be described in this chapter, a single treatment is assigned to a randomly selected sample (treatment group) and withheld from the remainder of the experimental sample (control group). Experiments may, however, include many different treatments and need not include a control group.

In addition with the number of treatments, experiments can include projects that differ markedly in nature. In particular, there are the "black-box" experiments where each treatment is a unique intervention or else each treatment is discriminated in a natural way from any other. It is not essential for an experiment to include a pure control group. Instead the investigators can concentrate on measuring the differences in effect of a number of distinctive new treatments. The definition of a "black-box" experiment can include tests of innovative new policies as well as studies that are indented to measure the effect of current policies relative to a null treatment.

Nevertheless, there are also experiments where the treatments are defined as points within a continuous policy parameter space and the experimental objective is to estimate a smooth response surface. Experiments like these occur when the dummy variable D represents different lengths of program participation, rather than a simple dichotomous indication of participation. The analysis here is only slightly different from the "black-box" case.

Finally, experiments can be conducted to evaluate the effects of a new social program, e.g. a training program, or to evaluate an ongoing program. In the first case, the analyst defines the target population to which the training program is referred. Then a sample of this population is enrolled to participate in the program. The latter case is a rather new kind of social experiment that takes place in an existing program. The only difference from the "new program" case is that the whole target population is asked to participate. Examples of such experiments are referred in Table 3.1. Heckman and Smith

(1993), Heckman, Smith and Clements (1997) and Heckman, LaLonde and Smith (1999) provide an analytic description of the experimental method along with its practical advantages. Some applications are mentioned in LaLonde (1986), Fraker and Maynard (1987), Manning et al (1987) and Dubin and Rivers (1993).

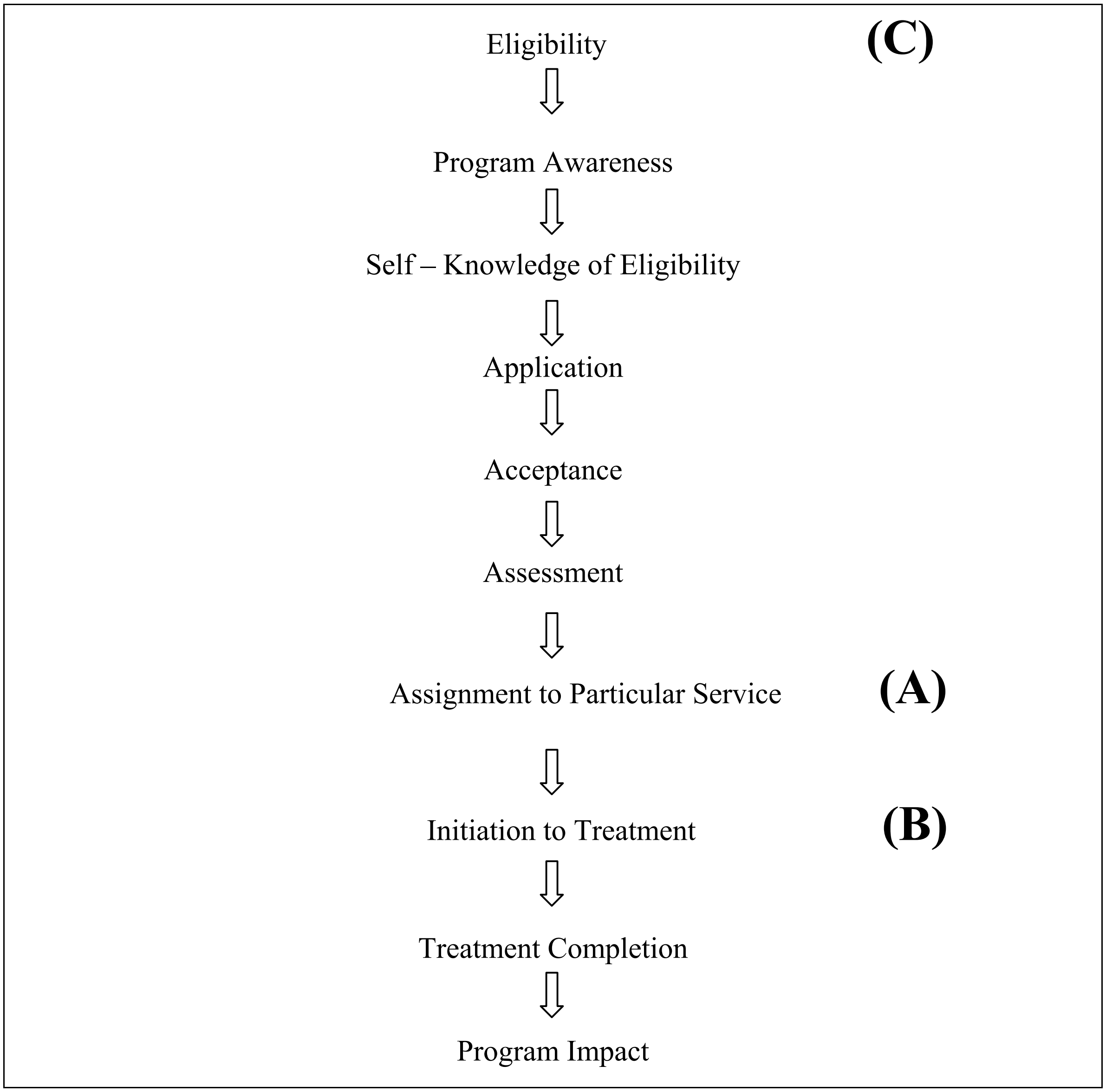
3.4 A Description of the Experimental Procedure in Evaluation Studies

This paragraph describes the standard experimental procedure followed to evaluate a social program. We begin with some relevant notation. Participants *(D_i = 1)* is a pre-defined population, e.g. unskilled workers like ex-drug addicts, ex- criminal offenders and economically disadvantage youths or simply a specific group of individuals from the population. LaLonde (1986) describes the CETA ongoing training program that is focused to such unskilled workers. After the announcement of the program from the authorities, the enrollment usually begins with referrals by welfare agencies, drug rehabilitation agencies or prisoners' assistant societies. In this way the individuals that compose the target population of this program can easily participate into it.

After the participation group is composed, a random subsample of it is asked to *enroll training* and constitutes the *treatment group*. The rest participants are assigned to *control group*. Only the formers receive training and therefore are the *actual participants*. Controls serve for comparison reasons in a way that we will describe later.

In a more schematic way, the participation process of a classical experiment is as follows (see Heckman and Smith, 1993):

Figure 3.1: Description of the Participation Process



* From the Job Training Partnership Act (JTPA) program participation process (1986 – 1994).

The above figure breaks the participation process into a sequence of steps. The process begins with program eligibility (certain characteristics that allow persons to participate in the program), continues through application, acceptance, assessment, assignment to services, the initiation of training and ends with the completion of training and consequently with outcomes reports.

3.5 The Evaluation Problem in Randomized Experiments

The case for randomized experiments is almost always stated within the context of problem (E-1), the "causal problem" as defined by statisticians (see Holland, 1986). As mentioned already, calculation of $\Delta_i = Y_{1i} - Y_{0i}$ for each person is impossible due to the missing data problem. In experiments one has access to participants' outcomes Y_{1i} and seeks to evaluate non-participants outcomes Y_{0i}. This problem, known as experimental Evaluation Problem, is solved in a population level by using mean estimates. By design, the crucial feature is to define the appropriate comparison group to approximate the gains of non-participants.

The idea of constructing a comparison group is actually an important step to the solution of the evaluation problem but it cannot deal with it by itself. An important assumption that has to be satisfied in order to produce consistent estimates is the elimination of any selection bias. Experiments cope with the problem of selection bias in a unique way that is going to be discussed below.

3.5.1 Solution of the Evaluation Problem

Let us define for each person *i* (*i* = 1, ..., N) the outcomes $\left(Y_{1i}^r, Y_{0i}^r, D_i^r\right)$ under randomization (where *r* denotes the existence of randomization), and $\left(Y_{1i}, Y_{0i}, D_i\right)$ under normal operation of the program without randomization. Let $D_i^r = 1$ for persons who participate in a program in the presence of random assignment and $D_i^r = 0$ for everyone else. Randomization is applied to the population for whom $D_i^r = 1$.

Let us denote with $r_i = 1$ if a participant ($D_i^r = 1$) is randomized into the experimental treatment group (*trainee*) and with $r_i = 0$ if a participant is randomized out and as a result belongs to the experimental control group (*control*). To simplify the notation we denote as X$_i$ the vector of observed individual characteristics for both participants and non-participants, that is $X_i = \left(X_{0ai}, X_{1ai}\right)$.

The randomization process generates the experimental data for the two distinct groups that give us:

$$E\left(Y_i^r \middle| D_i^r = 1, r_i = 1, X_i^r\right) = E\left(Y_{1i}^r \middle| D_i^r = 1, X_i^r\right) \qquad (3.1)$$

$$E\left(Y_i^r \middle| D_i^r = 1, r_i = 0, X_i^r\right) = E\left(Y_{0i}^r \middle| D_i^r = 1, X_i^r\right) \qquad (3.2)$$

Simple application of the effect of Treatment on the Treated estimator, conditional on X_i^r, identifies:

$$\begin{aligned} E\left(\Delta_i^r \middle| D_i^r = 1, X_i^r\right) &= E\left(Y_{1i}^r - Y_{0i}^r \middle| D_i^r = 1, X_i^r\right) \\ &= E\left(Y_{1i}^r \middle| D_i^r = 1, X_i^r\right) - E\left(Y_{0i}^r \middle| D_i^r = 1, X_i^r\right) \qquad (3.3) \end{aligned}$$

The mean effect of equation (3.1) can be calculated easily from the dataset since it is just the mean outcome of the persons participated and then randomly assigned to the treatment group. Nevertheless, estimation of equation (3.2) is more demanding.

After randomization, the participants that belong to the treatment group start to receive treatment. The completion of the treatment is followed by the collection of data for the gains (e.g. payments) of the trainees. Meanwhile, the non-treated do not seem to be benefited from their assignment to the control group. Neither they receive treatment nor they expect higher payments after the completion of the treatment. Therefore, phenomena like attrition or dropout from the program occur very often to that group of people. Moreover, some controls may gain access to close substitutes of that program where they would have an opportunity to be assigned in the treatment group. In that case they belong to both a control group from the first program and the treatment group from its substitute. Under these circumstances, their "actual" gains Y_{0i} are overstated due to receive of "substitution" gains. This problem is usually called *substitution bias*.

All these problems have to be overcome in order to measure adequately the "actual" gains of the controls. A possible solution is the offer of incentives, like payments, to non-treated individuals in order to persuade them to participate in the program and only in it (see Burtless and Orr, 1986). In addition, operators could occupy them with activities as a form of training different from the one offered by the training program. More often, the operators need to supervise the non-treated individuals and do not allow them to participate in another program. However, this strategy, termed in the literature as screening process, entails another source of expenses that raise the cost of the experiments (screening cost).

3.5.2 Elimination of Selection Bias

Someone could argue that since the whole target population joins the program, at least theoretically, the analysis is not affected by selection bias. However, this is not true. The population comes from various social agencies. It is possible that these agencies include persons with only certain characteristics and therefore the target population is not representative of the actual population (sample selectivity). In addition, for different reasons the program operators may not allow to the applicants that "would be disruptive to their program" to participate (choice-based selectivity). Moreover, because of the voluntary character of the program, many applicants (possibly a specific type) may not wish to participate. When the refusal rates are high, this phenomenon also produces another kind of bias, called as *attrition bias*. However, attrition from a program can be confronted by offering certain incentives to individuals in order to persuade them to participate.

Obviously, the intervention of sample or self-selection bias is a practical reality in an experimental evaluation study. Thus, along with the evaluation problem, selectivity has to be considered in determining the benefits from participation.

Random assignment of persons into the treatment group makes the treatment status (r_i) statistically independent of $\left(Y_{0i}^r, Y_{1i}^r, X_i^r\right)$. The pool of persons in the treatment and the control state is the same due to the perceived similarity in $\left(r_i = 1\right)$ and $\left(r_i = 0\right)$ individuals' attributes gained from the randomization procedure. Thus, the control group consists of *actual non-participants* that could approximate the outcomes of participants had they not participated.

To see mathematically how the selection bias is eliminated within an experimental framework let us take a model that supposes a common coefficient β for trainees ($r_i = 1$) and controls ($r_i = 0$):

$$Y_i^r = a + \beta \times r + u_i^r$$

In this model, Y_i^r is the outcome of interest, α is the mean outcome when no one participates ($r_i = 0$), β is the common effect of participation and u_i^r represents the random shock observed by the individual but not by the analyst. Mean earnings in the experimental, treatment and control, groups respectively are:

$$E\left(Y_i^r \middle| D_i^r = 1, r = 1, X_i^r\right) = a + \beta + E\left(u_i^r \middle| D_i^r = 1, r = 1, X_i^r\right) \tag{3.4}$$

$$E\left(Y_i^r \middle| D_i^r = 1, r = 0, X_i^r\right) = a + E\left(u_i^r \middle| D_i^r = 1, r = 0, X_i^r\right) \tag{3.5}$$

Subtracting the two potential means yields $E\left(\Delta_i^r \middle| D_i^r = 1, X_i^r\right) = \beta$, the TT(X) parameter. Selection bias, expressed from $E\left(u_i^r \middle| D_i^r = 1, r, X_i^r\right) \neq 0$, has been canceled out in the calculation of the effect of treatment on the treated. However, nothing about randomization guarantees that $E\left(u_i^r \middle| D_i^r = 1, r, X_i^r\right) = 0$ and selection bias is not eliminated from each mean effect separately. *Rather randomization balances the bias in the two samples, so that it cancels them out when calculating the mean impact estimate*. (Heckman and Smith, 1993).

Intuitively, the above argument has a practical base. Since only participants compose the target population, people with similar observable and unobservable attributes are found in both statuses ($r_i = 0$ and $r_i = 1$). Overestimation (or underestimation) in the mean outcomes in one group is followed by a similar amount of overestimation (or underestimation) in the corresponding second group’s outcomes. Of course the crucial assumption made here is that training biases mean outcomes to the same degree as non-training, that is $E\left(u_i^r \middle| D_i^r = 1, r = 1, X_i^r\right) = E\left(u_i^r \middle| D_i^r = 1, r = 0, X_i^r\right)$. If this is not the case, because, for example, controls are disappointed from the result of random selection and do not obtain the earnings that they would obtain as participants, biases will remain in the analysis. Such a specific case occurs in the presence of substitution bias.

Concentrating on this issue we comment that although randomization is an excellent tool for the solution of the evaluation and the selection bias problems, it must not be implemented without restrictions. There are some crucial assumptions that only when satisfied, the estimator (3.3) is consistent with the data and produces unbiased estimates. These assumptions are outlined in the subsequent paragraph.

3.5.3 Identification Assumptions

The most common estimate in evaluation of social programs is the mean effect of Treatment on the Treated. The essential assumption required to consistently estimate this parameter is:

- $E\left(Y_{1i}^{r} - Y_{0i}^{r} \middle| D_{i}^{r} = 1, X_{i}^{r}\right) = E\left(Y_{1i} - Y_{0i} \middle| D_{i} = 1, X_{i}\right)$

In other words, randomization should not alter the process of selection into the program, so that those who participate in terms of an experiment should not differ from those who would have participated in the absence of an experiment. Put simply, *randomization bias* must not occur.

There are many reasons to suspect the validity of this assumption. In the context of an ongoing program, if individuals who might have enrolled in a nonrandomized regime make plans anticipating enrollment in training, adding uncertainty at the acceptance stage may alter their decision to apply or to undertake activities complementary to training. Risk-averse persons will tend to be eliminated from the program.

A stronger set of conditions, not strictly required, is:

$$E\left(Y_{1i}^{r} \middle| D_{i}^{r} = 1, X_{i}^{r}\right) = E\left(Y_{1i} \middle| D_{i} = 1, X_{i}\right) \quad (3.6)$$

$$E\left(Y_{0i}^{r} \middle| D_{i}^{r} = 1, X_{i}^{r}\right) = E\left(Y_{0i} \middle| D_{i} = 1, X_{i}\right) \quad (3.7)$$

These assumptions state that the means from the treatment and control groups generated by random assignment produce the desired population parameters. Apparently here, the non-existence of “randomization bias” is assumed for each average separately.

Even when the above assumptions fail to hold, there are two important special cases, under which experimental data still provide unbiased estimates of the effect of treatment on the treated.

1. First we mention the “fixed treatment effect for all units model”, which refers to a situation where everyone has the same gain (or loss) from a program, so that

$$\Delta_i^r = Y_{1i}^r - Y_{0i}^r = a \tag{3.8}$$

the same for everyone, as implied in either unit homogeneity or constant effect assumption.

This model can be written as:

$$Y_{1i}^r = g_1\left(X_{1i}^r\right) = a + g_0\left(X_{0i}^r\right) = a + Y_{0i}^r \tag{3.9}$$

which is an immediate result by simply solving (3.8) for Y_{1i}^r.

In term of a linear regression model, equation (3.9) can be also written as:

$$X_{1i}^r \times \beta_1 = a + X_{0i}^r \times \beta_0 \tag{3.10}$$

Changing the composition of participants by randomization has no effect because the parameter of interest is the same for all possible participants populations. To see this mathematically:

$$\begin{aligned}
& E(Y_{1i}^r \middle| D_i^r = 1, X_i^r) - E(Y_{0i}^r \middle| D_i^r = 0, X_i^r) \\
&= E(\alpha + Y_{0i}^r \middle| D_i^r = 1, X_i^r) - E(Y_{0i}^r \middle| D_i^r = 0, X_i^r) \\
&= \alpha + E(Y_{0i}^r \middle| D_i^r, X_i^r) - E(Y_{0i}^r \middle| D_i^r, X_i^r) \\
&= a \\
&= E\left(\Delta_i^r \middle| D_i^r = 1, X_i^r\right) \\
&= E\left(\Delta_i^r \middle| X_i^r\right) \\
&= \Delta_i^r
\end{aligned}$$

This model is used in applied work. Reliance on it strengthens the popular case for randomization. Questions 1 and 2 (paragraph 2.4) have the same answer under this formulation and randomization provides a convincing way to answer both. However, as Heckman, Smith and Clements (1997) mention, this model is rejected in most practical situations because of its simplicity.

2. The second case arises under somewhat weaker conditions. Let us introduce the "random effect model", which has the form:

$$Y_{1i}^{r} = a + Y_{0i}^{r} + \xi_i^{r} \tag{3.11}$$

where $E(\xi_i^r \mid D_i^r = 1, X_i^r) = 0$. Suppose that potential trainees know the mean impact of training, $E\left(\Delta_i^r \middle| X_i^r\right)$, but not their personal gain (or loss), Δ_i^r, from it at the time participation decisions are made. Under these circumstances where they naturally use population means α in place of personal gains $a + \xi_i^r$ in making participation decisions, the decision about training is not affected by the realized gain from the participation to the program, but rather is random. Therefore there is not selection bias in participation decision and the subsequent randomization do not results in a pool of participants that would differ from the one without randomization. Hence, even in heterogeneous responses to treatment, the simple cross – section mean-difference estimator obtained from experimental data may have desirable properties.

- *A second assumption* is that members of the control group cannot obtain close substitutes for the treatment elsewhere. In the opposite case substitution bias exists. In the presence of substitution bias, the experimental control group no longer corresponds to the desired counterfactual of persons who wanted to receive treatment but did not because their outcomes are not representative non-participant outcomes. As a result the mean difference in outcomes between treatment and control groups no longer provides an estimate of the mean impact of treatment on the treated. Heckman (1991b) discusses the need for absence of substitutes in a program being evaluated and the failure of the randomization bias assumption.

- *A third assumption*, posed by Lewis (1963), supports that controls outcomes within a given policy regime closely approximates the outcomes of non-participants. This assumption allows the analyst to ignore indirect effects from the analysis. Thus, only direct outcomes, e.g. wages, may be taken into account in the evaluation process. However, as Heckman, LaLonde and Smith (1999) mentions *in the context of evaluating large-scale employment and training programs at a national level, it is natural to ask whether this assumption is valid and the consequences of the evaluation if it is not. To answer these questions in a convincing fashion requires constructing a model of the labor market, a task that is rarely performed in conventional evaluation studies*.

In any case, indirect benefits seem to be an important factor to be analyzed in an evaluation study. Such an analysis indicates how the social economy is affected from the conduction of the social program. The problem of indirect effects poses a major challenge to conventional micro-methods that focus only on direct impacts, and demonstrates the need for program evaluations to utilize market-wide data.

3.6 Stages for Randomization

Until now, we have described social experiments as the procedure where the observational units of study are randomly assigned to meaningfully different treatments. The treatment and control groups, constructed in this way, help analyst to evaluate a social program. However, we have not yet determined the stage where the randomization should be implemented. This is going to be discussed now.

Remember Table 3.1 that illustrates the steps of the participation process. In principle, randomization can be performed to evaluate outcomes at each stage. Practically though, cost and ethical reasons allow only one randomization to be performed. The question is at what stage it should be placed. One obvious answer is at the stage where it is least disruptive, in the sense that neither randomization bias nor substitution bias nor dropouts from the treatment group will occur. Nevertheless, the determination of this stage is not an easy matter in the absence of considerable information about the process being studied.

Heckman (1991b) and Heckman and Smith (1993) examine the effect of the location of random assignment on the frequency with which participants drop out of the program between random assignment and the initiation of training. Adopting the previous notation, the effect of Treatment on the Treated is defined as:

$$E\left(\Delta_i^r \,\middle|\, \mathrm{D}_\mathrm{i}^\mathrm{r} = 1, \mathrm{X}_i^r\right) = E\left(\mathrm{Y}_{1i}^r - \mathrm{Y}_{0i}^r \,\middle|\, \mathrm{D}_\mathrm{i}^\mathrm{r} = 1, \mathrm{X}_i^r\right)$$

$$= E(Y_{1i}^r | D_i^r = 1, X_i^r) - E(Y_{0i}^r | D_i^r = 1, X_i^r) \qquad (3.12)$$

Locating random assignment as close as possible to the actual initiation of training helps to reduce the opportunity for dropping out of the program in the period between random assignment and the receipt of training. This strategy makes $r_i \perp \left(Y_{0i}^r, Y_{1i}^r, X_i^r\right)$ relationship true in practice, since everyone in the treatment group receives treatment and the pool of persons in the two “in comparison” groups is not altered. The above reasoning suggests that point **B** in Table 3.1 is the optimal location of random assignment given the stated evaluation question E-1.

While point **B** seems to be the optimal one, institutional and political factors made it impracticable to be chosen (see Heckman and Smith, 1995). Instead random assignment is usually located at point **A**, just after assignment to a particular service type. While assignment to a particular service and initiation of treatment are consecutive steps in the participation process, in practice they may be separated in time by weeks or even months. As many occupational training classes are offered on an academic schedule, a trainee assigned to a particular training course must often wait until the beginning of the next academic quarter or semester to begin training. During these waiting periods, the possibility that a trainee will become disinterest or take training from another source is increasing. Thus, there is substantial attrition in the experimental treatment group and the problem of randomization bias occurs.

Randomization in eligibility

Eligibility (point **C**) has been proposed by Heckman (1991b) as a less disruptive point for randomization in the participation process. This point reduces the selection bias in the estimation of the mean benefits from participation that is produced due to institutional limitations. Moreover, it avoids the application and screening costs that are incurred when accepted individuals are randomized out of a program. Since the randomization is performed outside of the training center, it prevents the training center from bearing the political costs of denying eligible persons the right to participate in the program.

Suppose that eligibility, denoted by e, is randomly assigned in the population with probability q, $P\left(e=1\middle|X_i^r\right)=q$, and such assignment does not affect the decision to participate in the program among the eligibles. By denoting as p the probability of participation in the program, $p^r=P\left(D_i^r=1\middle|X_i^r\right)$, we have:

$$E(Y_{1i}^r|D_i^r=1,e=1,X_i^r)-E(Y_{0i}^r|e=0,X_i^r)$$

$$=E(Y_{1i}^r|D_i^r=1,X_i^r)-\left\{E(Y_{0i}^r|D_i^r=1,X_i^r)\times\frac{p^r(1-q)}{p^r(1-q)+1-p^r}+E(Y_{0i}^r|D_i^r=0,X_i^r)\times\frac{(1-p^r)}{p^r(1-q)+1-p^r}\right\}$$

$$=E(Y_{1i}^r-Y_{0i}^r|D_i^r=1,X_i^r)+\frac{p^r(1-q)}{p^r(1-q)+1-p^r}\times\left[E(Y_{0i}^r|D_i^r=1,X_i^r)-E(Y_{0i}^r|D_i^r=0,X_i^r)\right]$$

$$=E(\Delta_i^r|D_i^r=1,X_i^r)+\frac{p^r(1-q)}{p^r(1-q)+1-p^r}\times\left[E(Y_{0i}^r|D_i^r=1,X_i^r)-E(Y_{0i}^r|D_i^r=0,X_i^r)\right]\quad(3.13)$$

Thus, the bias is smaller in absolute value than would be from a mean comparison between treated and untreated samples without randomization on eligibility, as long as *0 < q < 1* and $0\prec p^r\prec 1$, since

$$E(Y_{1i}^r|D_i^r=1,X_i^r)-E(Y_{0i}^r|D_i^r=0,X_i^r)$$

$$=E(\Delta_i^r|D_i^r=1,X_i^r)+\left\{E(Y_{0i}^r|D_i^r=1,X_i^r)-E(Y_{0i}^r|D_i^r=0,X_i^r)\right\}\qquad(3.14)$$

A straight comparison of equations (3.13) and (3.14) proves the argument.

The intuition is clear:

> *By making some potential participants ineligible, the non – participant population now includes some persons whose mean outcomes are the same as what participant outcomes would have been if they did not participate.*

Using data on those who are eligible and do not participate, Heckman (1996b) manipulates further equation (3.13) to show that:

$$E(\Delta_i^r \big| D_i^r = 1, X_i^r) = \frac{E(Y_{1i}^r \big| e = 1, X_i^r) - E\left(Y_{0i}^r \big| e = 0, X_i^r\right)}{P(D_i^r = 1 \big| X_i^r)}$$

provided that $p^r = P(D_i^r = 1 \big| X_i^r) \neq 0$ and $\left(Y_{0i}^r, Y_{1i}^r, D_i^r, X_i^r\right) \perp e$.

In practice, this estimator is likely to be useless because p^r is often small so that the sampling variability of the estimator is likely to be large. Very large samples would be required to reliably evaluate low probability outcomes, which imply a proportional increase in financial costs.

3.7 Identification of the Impact Distribution

In order to identify the joint distribution of outcomes, $F\left(Y_{0i}^r, Y_{1i}^r, D_i^r \big| X_i^r\right)$, or the distribution of the benefits, $F\left(\Delta_i^r, D_i^r \big| X_i^r\right)$, directly, one have to determine first the conditional distributions:

$$F_1\left(Y_{1i}^r \big| D_i^r = 1, X_i^r\right)$$
$$F_0\left(Y_{0i}^r \big| D_i^r = 0, X_i^r\right)$$

and

$$F_1\left(Y_{1i}^r \big| D_i^r = 0, X_i^r\right)$$
$$F_0\left(Y_{0i}^r \big| D_i^r = 1, X_i^r\right)$$

Although an analyst can identify the first two from ordinary experimental data, he cannot infer anything but the mean outcomes for the last two conditional distributions, $E\left(Y_{1i}^{r}\middle| D_{i}^{r}=0,X_{i}^{r}\right)$ and $E\left(Y_{0i}^{r}\middle| D_{i}^{r}=1,X_{i}^{r}\right)$ by conducting the randomization process as discussed. Hence, neither the joint distribution of impacts nor features of this distribution, e.g. median, percentiles etc, can be obtained from ordinary data unless specific assumptions are adopted. A discussion in these assumptions for different response patterns is found below.

3.7.1 Distribution Identification Assumptions – Response Patterns

1. *Temporal Stability and Causal Transience*

Under Temporal Stability and Causal Transience it is a simple matter to calculate Y_{0i}^{r} and Y_{1i}^{r} for each person and in extend determine the joint distribution of impacts. However, neither this assumption is valid in most practical situations, nor the implied procedure is usually performed due to the great financial and time costs that involves.

2. *The common effect model*

As mentioned before, the common effect or "fixed treatment effect for all units" model has the form:

$$Y_{1i}^{r}=a+Y_{0i}^{r}$$

In this model, everyone has the same gain from a program, so that:

$$E\left(\Delta_{i}^{r}\middle|D_{i}^{r},X_{i}^{r}\right)=E\left(\Delta_{i}^{r}\middle|D_{i}^{r}\right)=\Delta_{i}^{r}=a$$

This particular model favors social experiments for two reasons. The first is the exclusion of randomization bias. The second and the most important is that the link between outcomes in the two states is known for each individual regardless of their observed state. Under this assumption, experimental data reveal the full joint distribution of outcomes.

As showed by Heckman (1991b), since $Y_{1i}^{r} - Y_{0i}^{r} = a$, a constant, knowledge of either Y_{0i}^{r} or Y_{1i}^{r} determines the other. Graphically, the distribution of Y_{1i}^{r} equals the distribution of Y_{0i}^{r} shifted over by α. More formally, if *F₁* is the cumulative distribution function of Y_{1i}^{r} and *F₀* the cumulative of Y_{0i}^{r}, then $F_1\left(Y_{0i}^{r} + a \middle| X_i^{r}\right) = F_0\left(Y_{0i}^{r} \middle| X_i^{r}\right)$. As a result, estimation of features of the joint distribution, other than the mean effect, can be obtained.

The common effect model simplifies greatly the evaluation problem but it is rejected in most practical situations. In order to obtain a more plausible model, analysts suggest that heterogeneity in responses should be allowed, so that Δ_i^{r} be unequal across participants *i*.

3. Heterogeneity in Responses

Heterogeneity in responses can be expressed easily in a model form:

$$Y_{1i}^{r} = a + Y_{0i}^{r} + \xi_i^{r}$$

By assuming the above model, one can extract the joint distribution of outcomes in a rather interesting way.

Under the assumptions described in paragraph 3.5.3, suppose access to data of 2 × N participants from a social experiment[1]. Half of the participants have been randomly assigned to the treatment state ($r_i = 1$) while the others are randomized out and compose the experimental control group ($r_i = 0$). The outcomes, which are continuously distributed, can be represented in two N × 1 vectors. Ignoring ties, we rank individual outcomes from the highest to the lowest as follows:

[1] This example is illustrated in Heckman and Smith (1995).

Treatment outcome distribution $F_1\left(Y^r_{1i}\middle|D^r_i=1,X^r_i\right)$	**Control outcome distribution** $F_0\left(Y^r_{0i}\middle|D^r_i=1,X^r_i\right)$
$Y^r_{1i}=\begin{pmatrix} y^r_{11} \\ y^r_{12} \\ . \\ . \\ y^r_{1N} \end{pmatrix}$	$Y^r_{0i}=\begin{pmatrix} y^r_{01} \\ y^r_{02} \\ . \\ . \\ y^r_{0N} \end{pmatrix}$

From the data we can identify the marginal distributions $F_1\left(Y^r_{1i}\middle|D^r_i=1,X^r_i\right)$ and $F_0\left(Y^r_{0i}\middle|D^r_i=0,X^r_i\right)$, but we do not know where person *i* in the treatment distribution would appear in the non-treatment distribution. By considering all possible permutations we obtain the N! possible sorting of treatment Y^r_{1i} and control Y^r_{0i} outcomes using realized values from one distribution as a counterfactual for the other. In other words, we can form a collection of N! possible impact distributions, i.e. alternative distributions of:

$$\Delta^r_N = Y^r_{1i} - \prod\nolimits_N Y^r_{0i}$$

where Π_N is a particular N × N permutation matrix of Y^r_{0i} in the set of all N! permutations, associating the ranks in the Y^r_{1i} distribution with the ranks in the Y^r_{0i} distribution. Δ^r_N, Y^r_{0i} and Y^r_{1i} are the N × 1 vectors of impacts, of controls and treatment outcomes, respectively.

By considering all possible links of Y^r_{0i} and Y^r_{1i}, one can bound the full impact distribution and thus obtain a bounded joint distribution of outcomes. However, due to variability of the impact estimates, additional assumptions are required to obtain informative bounds.

Based on experimental data, Heckman and Smith (1995) found that departures from high levels of positive dependence between Y^r_{0i} and Y^r_{1i} produce absurd ranges of impacts on gross outcomes. Consequently, an important assumption to limit the range of the bounds is $Corr\left(Y_{0i}^{r}, Y_{1i}^{r}\right) \approx 1$.

In addition, rationality on the program participation decision may contribute to recover the joint distribution of outcomes from experimental data. Defining $U\left(Y_{0i}^{r}\right)$ and $U\left(Y_{1i}^{r}\right)$ as the utility expected by each person prior to participation decision in state 0 and 1 respectively, the participation rule can be defined as:

$$\int U\left(Y_{1i}^{r}\right) dF_{1}\left(Y_{1i}^{r} \middle| D_{i}^{r}=1\right) \succ \int U\left(Y_{0i}^{r}\right) dF_{0}\left(Y_{0i}^{r} \middle| D_{i}^{r}=1\right)$$

This relationship imposes a restriction on the nature of dependence between Y_{0i}^{r} and Y_{1i}^{r} given $D_{i}^{r}=1$. With enough variability in the values of the corresponding outcomes, the full joint distribution can be recovered.

Although this process seems to be quite simple, in most practical situations two kinds of problems occur:

- In most data sets there are unequal numbers of treated and control persons to calculate the two empirical distributions. That is $N_{1_1} \neq N_{0}$.
- Even if $N_{1}=N_{0}=N$, N is usually very large and is computationally demanding to consider all possible permutations of the data distribution.

To circumvent these problems one may work with quantiles of the two distributions and permute them instead of the outcomes (a detailed description of this process is referred to Heckman, Smith and Clements, 1997).

3.8 Ethical Issues

Experimental evaluation studies have been criticized sharply for the ethical costs that their implementation entails. The ethical issues that are often discussed in the presence of a social experiment have to do with the protection of privacy and confidentiality of the data obtained from participants as well as with the rights of the targeted individuals to refuse to participate or later to withdraw from the experiment. The perceived importance of those issues is proved by the methods developed to protect both participants' right to privacy (transformation of data in aggregate forms, laws for privacy of personal data) and withdrawal from the experiment (incentives to maximize the probability of participation).

However, as Burtless and Orr (1986) mention, the most important ethical issue that occurs in classical experiments is stated upon whether is right to withhold potentially beneficial services from the control group while the treatment group is free to enjoy the supply of these services. In experiments, the regarded fundamental ethical principle is that the expected net benefits to experimental subjects, trainees and controls, should not be less than those they could expect in the absence of the experiment. In experiments providing a beneficial treatment that would not otherwise have been available, this condition is met. The treatment group is unambiguously better off by participating and controls are no worse off because of the experiment.

However, for programs that do not include exclusive services this is not the case. For example on ongoing programs the controls of one program can be trained elsewhere. Since the research objective requires that controls be denied services that they might otherwise receive because of the substitution bias problem, it can be said that the control group is made worse off. In this case, the issue whether randomization is ethical or not is completely justified. However, when budgetary or other constraints imply that services are not available to all participants, random assignment is the only ethical way to ration available services.

An important requirement in experimental studies is that the participants have to be informed about the random assignment of persons to the treatment group and give their consent to participation. By consenting to the conditions of the experiment, the participant in effect certifies that, in his or her view, the experiment is expected to yield positive net benefits. For informed consent to afford the assurance of positive net benefits to each individual participant, individuals who refuse to consent must receive the same services they would have received in the absence of the experiment. If the experiment entails nothing more than random assignment to program services or the control group without any services, few applicants will voluntary choose to participate; participation would simply reduce their chances of receiving services since they can find the same services in another program. This outcome would obviously destroy the research value of the experiment. To see this, we place at this point the results of a specific training program, reported by Doolittle and Traeger (1990).

Report from JTPA training program

The U.S. Department of Labor financed a large-scale experimental evaluation of the ongoing large scale Job Training Partnership Act (JTPA), which is the main vehicle for providing government training in the U.S. Randomization evaluation was implemented in a variety of sites. The organization implementing this experiment was the Manpower Demonstration Research Corporation (MDRC).

Job training in the U.S. is organized through geographically decentralized centers. These centers receive incentive payments for placing unemployed persons and persons on welfare in "high paying" jobs. The participation of centers in the experiment was not compulsory.

In attempting to enroll geographically dispersed sites MDRC experienced a training refusal rate in excess of 90%. The refusal reasons are presented in the following table (the reasons are not mutually exclusive):

Table 3.2: Percent of Training Centers Cited Specific Concerns About Participating In The JTPA Experiment

Concern	Percent of Training Centers Citing the Concern
1. Ethical and Public Relations Implications of:	
a) Random Assignment in Social Programs	61.8
b) Denial of Services to Controls	54.4
2. Potential Negative Effect of Creation of Control Group on Achievement of Client Recruitment Goals	47.8
3. Potential Negative Impact on Performance Standards	25.4
4. Implementation of the Study When Service Providers Do Intake	21.1
5. Objections of Service Providers to the Study	17.5
6. Potential Staff Administrative Burden	16.2
7. Possible Lack of Support by Elected Officials	15.8
8. Legality of Random Assignment and Possible Grievances	14.5
9. Procedures for Providing Controls With Referrals to Other Services	14.0
10. Special Recruitment Problems for Out-Of-School Youths	10.5
Sample Size	**228**

Leading the list are ethical and public relations objections to randomization. Major fears (items 2 and 3) were expressed about the effects of randomization on the quality of the applicant pool, which would impede the profitability of the training centers. In attempting to persuade centers to participate, MDRC had to reduce the randomized rejection probability from $1/2$ to $1/6$ for certain centers, widening in this way the available pool of persons deemed eligible. The resulting reduction in the size of the control sample impairs the power of statistical tests designed to test the null hypothesis of no program effect. Doolotle and Traeger (1990) also report:

> *"Implementing a complex random assignment research design in an ongoing program providing a variety of services does inevitably change its operation in some ways…The most likely difference arising from a random assignment field study of program impacts … is a change in the mix of client served. Expanded recruitment efforts needed to generate the control group, draw in additional applicants who are not identical to the people previously served. A second likely change is that the treatment categories may somewhat restrict program staff's flexibility to change service recommendations…*
>
> *Some training centers because of severe recruitment problems or up-front services cannot implement the type of random assignment model needed to answer the various impact questions without major changes in the procedure".*

The above evidence witnesses the major concerns in producing credible estimates of the JTPA program impacts. The selective participation occurred by the problems in the implementation of the randomization procedure as well as the change in the mix of treated and non-treated individuals are major sources of randomization bias that destroy the results of the experiments.

3.9 Evaluation Under Dropouts

A rather important issue that frequently occurs in the presence of an experimental study is that persons randomly assigned to the experimental treatment group often drop out from the program under study. In the presence of dropouts, the usual experimental mean difference estimator provides an estimate of the mean impact of the assignment to treatment rather than the mean impact of the treatment itself. In other words, the mean effect of the treatment on the treated is actually an estimate of the availability of the treatment. This parameter is usually referred as the “intention to treat” and is not a useful economic measure. In the presence of dropouts from the program alternative estimators have to be considered.

Heckman, Smith and Taber (1998) consider two patterns of dropouts. In the first, the experimental group members may drop out of the program prior to receiving treatment. In the second, the dropouts receive a partial dose of the program treatment prior dropping out. The authors give a complete representation of these situations and develop a theory in order to estimate adequately the mean effect of treatment on the treated in each case.

3.10 The Case For and Against Social Experimentation

Experimentation is the pure statistical method to deal with the evaluation problem. In the simplest case, by randomizing participants into a treatment and a control group, the analyst approximates the counterfactual parameter of what participants would have earned had they not participated. Thus, the mean effect of treatment on the treated can be estimated easily by subtracting the mean earnings of controls (counterfactual mean) from them of trainees. Naturally, social experiments are reported to have their advantages and disadvantages. Several authors have studied the experimental procedures and provide various comments on them. A relative discussion is provided below.

3.10.1 Advantages of experiments

Compared to other methods, experiments are easy to be described. Since experimental subjects are randomly assigned to a particular group, the mean effects of outcomes can be estimated with high reliability. Random assignment removes any systematic correlation between treatment status and both observed and unobserved participant characteristics leading in comparable treatment and control groups and elimination of self-selection bias. Provided a mindful experimental planning, randomization on different stages can also produce mean estimates other than the effect of treatment on the treated.

Another advantage of experimental methods is that they do not make any distributional assumptions. The control group is generated through randomization in participants' group. Further assumptions are not required to obtain a comparable counterfactual and estimate its mean. Due to this simplicity in the procedure, experiments are easily described to policy-makers and the results are also easily explained.

A controversial argument in support of experimental methods is that experiments produce a consensus, which is one result only. There are not a variety of estimators that under different assumptions produces different results. For example, randomization at training initiation stage produces just a simple estimator of the effect of treatment on the treated. Many experimenters, e.g. Burtless and Orr (1986), LaLonde (1986) and Holland (1989), have considered it as a great advantage of experimentation. The opposite arguments are going to be described later at the non-experimental approaches to the evaluation problem.

Heckman (1996b) recognizes that experiments balance the distribution of X_i^r values in the treatment and control groups and thus lead in comparable group of participants. He comments that even when in the population is valid that

$$Support\left(X_i^r \middle| D_i^r = 1\right) \neq Support\left(X_i^r \middle| D_i^r = 0\right)$$

which means that there are participants with specific characteristics X_i^r that cannot be found in the group of non-participants, experiments lead to a solution. Randomization enriches the support of X_i^r in a way that:

$$Support\left(X_i^r \middle| D_i^r = 1, r = 1\right) = Support\left(X_i^r \middle| D_i^r = 1, r = 0\right)$$

Thus, incomparability in the population stage results in comparability under randomization.

3.10.2 Limitations of experiments

Despite the obvious advantages of experimentation, there are also great limitations that in several situations make its use impractical. Many limitations concern the various forms of bias inserted at the "acceptance" to "treatment completion" stages, as well as the high financial costs that characterize any social experiment.

Non-comparability of the control group

The first criticisms for experimental methods are found in the work of Cochran (1965). He denotes the major difficulties an analyst is faced up with when applying a social experiment to evaluate a program. Specifically, having decided on the type of the essential comparisons, the analyst has often to search for some environment in which it may be possible to collect the appropriate data. As representative examples can be thought the studies for the effect of air-pollution on the health of urban dwellers, the type of protection afforded by seat belts under actual accident conditions or the relative effectiveness of surgery and radiation for the treatment of malignant conditions in which ethical considerations forbid randomized experimentation. To produce results in such experiments, the analyst is compelled to compare groups that are not ideal for his purposes or else to postpone the study, hoping that later a more suitable environment will be found.

Later, Heckman, LaLonde and Smith (1999) consider the non-comparability argument from a different point of view. They denote that although experiments balance the distribution of X_i^r values in the treatment and control groups and thus lead in comparable group of participants, they cannot provide estimates of $\left(\Delta_i^r \middle| D_i^r = 1, X_i^r\right)$ for values of X_i^r

such that $P\left(D_i^r = 1 \middle| X_i^r\right) = 0$. Since experiments by design accept volunteers as participants to the program and these volunteers are randomized into two groups, people with characteristics that give them probability zero to participate will not appear in the program. Thus this group of persons is not represented in the study.

Cost of experiments

Cochran (1965) is the first who indicates the costs of experiments. He mentions that financial cost of experiments is usually very high and, sometimes, random sampling may not even be feasible. Instead of this other, cheaper sampling schemes may be preferred that most of the times yield inconsistent results. Rivlin (1974) agrees with the comment of the great financial costs, but notes that even the costs of major experiments are small compared with the costs of social policies that do not work at all, or at least properly, because its effectiveness have not ever been examined by an experiment.

Burtless and Orr (1986) add another dimension in the costs arguments. They indicate that apart from the great financial costs that very often are over $5 - $10 million dollars even for relative small experiments, there are also great time costs. To properly design and analyze a classical experiment might take several years. For example, the NIT experiment was launched in 1970 and the final report was not issued yet until 1983. If policy-makers need authoritative results within one or two years, experiments are not feasible. The delay between the initial decision to experiment and the final report can have important consequences for the policy usefulness of experimental results. Issues that appear timely at the moment an experiment is launched may fade to insignificance before the findings become available.

In addition, experiments often involve significant political costs. It is more difficult to develop, implement and administer a new treatment than it is simply to analyze information about past economic conditions or collect and analyze new information about economic behavior. Voters and policy-makers are rightly concerned about possible ethical issues raised by experiments.

In favor to experiments, Heckman, LaLonde and Smith (1999) argue that the high financial and time cost of social experiments results not from administering randomization, but mainly from data collection, careful documentation of the implementation of the program, analysis, and dissemination of reports. Yet, as it is discussed later on econometric estimators, these costs are not unique to social experiments, but arise in any careful program evaluation.

Non-response bias

Burtless and Orr (1986) indicate that non-response bias frequently occurs in evaluation studies since some participants either refuse or cannot be located. In this way the sample of participants may suffer from selection bias from the initial stage and then even randomization is not able to cope with it. In addition, when the refusal or no location rates are high in a sample, this will no longer be representative of the entire population. Thus, problems in population inferences occur.

Limited duration bias

Inference problems may also occur when the sampled persons behave differently as experimental units than they would behave in the absence of an experiment. For example, Burtless and Orr (1986) mention that in the NIT experiment, individuals did not declare their actual income and as a result the income effect of the experimental transfers was probably understated and the price effect overstated. This problem, however, occurs only in limited duration programs where individuals may try to mislead the analyst. In regular "long- term" programs it is unlikely that they will manage to preserve this "fake" behavior for a long period. The authors mention that in any case, designing the experiment in such a way to allow detection and measurement of the bias can solve this problem, called Limited Duration Bias problem.

Hawthorne effects

Similar to this problem is also the one named Hawthorne Effect where participants behave differently simply because they know the program is an experiment, or because they do not take an experiment as seriously as they would a "real" program. Although, Burtless and Orr (1986) insist that there is not any convincing evidence that Hawthorn effects is a significant factor in experiment outcomes, others like Heckman and Smith (1998) consider it as a major reason for the existence of biases in social experiments.

Queueing Bias

Burtless and Orr (1986) extend their analysis to another problem termed as queuing bias. To see this let us introduce a specific example. A wage subsidy restricted to a number of disadvantage workers might give these workers an advantage in obtaining the limited number of low-wage jobs at the expense of unsubsidized workers; but if all disadvantage workers in the area were subsidized, the advantage to each subsidized worker would be much smaller. Thus, the effect of treatment on certain outcomes (e.g. employment) observed in the experiment may overstate the effects that would be expected under a universal program.

Randomization and substitution bias

Heckman (1991b) and Heckman and Smith (1993, 1995) refer to these problems in terms of an experimental evaluation study. “Bribing” participants in order to maximize participation rates and in extend to ensuring that participants will receive services only from the original program may eliminate both kinds of biases. However, money payments and screening procedures always increase the financial cost of the study. A common way to avoid screening costs is, as mentioned, randomization on eligibility.

Parameter estimation problem

Heckman (1991b) argues that although no distributional assumptions need to be invoked, experiments suffer from a major problem that restricts their use for evaluation purposes. Despite of the simplicity in estimating the mean effect of treatment on the treated, several alternative parameters, such as ATE or the effect of certain attributes to post-program incomes cannot be evaluated unless randomization is implemented in more than one stage. However, as Heckman and Smith (1995) indicate multiple randomization designs are not commonly encountered in experimental evaluation studies since “practical difficulties would make it impossible in most cases”.

Institutional problems

Very often certain institutional factors may cause severe problems in an experimental evaluation study. One of them is the problem of attrition, specifically when random assignment takes place several months after the initiation of training, although random assignment and receive of treatment are consecutive steps!

A second institutional problem is the difficulties in generating separate experimental estimates of the impact of different service types. Heckman and Smith (1995) provide an illustrative example from the JTPA program. That program offers a number of different employments and training services. Some participants receive a single service type while others receive specially designed sequences of services. Obtaining estimates of individual services from this program was a primary objective. However, the structure of the JTPA made it impossible, at least without multistage randomization that are costly and rarely occur in a program.

Finally, the authors denote that voluntary participation in a program can cause severe attrition bias in it since it may participate only a certain group of people. On the other hand, forced participation does not seem to preserve as an appealing solution since it hides high screening costs and ethical problems.

Generally, experimentation is a useful tool in evaluation studies since it can easily estimate an economically interesting parameter, namely the effect of treatment on the treated. Several other impacts can be estimated by implementing randomization at different stages of the study. However, the disadvantages of high financial and time costs are the main inhibitory factors. The various forms of bias that occur independently of the stage of randomization are also considered as great limitations of experimental designs. All these problems lead to the development of alternative approaches to the evaluation of social programs.