DOCUMENT RESUME

ED 069 724 TM 002 176

AUTHOR	Rubin, Donald
TITLE	Estimating Causal Effects of Treatments in Experimental and Observational Studies.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	ETS-RB-72-39
PUB DATE	Aug 72
NOTE	33p.; Draft
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	*Control Groups; Definitions; Evaluation Methods; Evaluation Techniques; *Experimental Groups; Matched Groups; Measurement Techniques; *Models; Observation; Research Design; *Research Methodology; *Statistical Analysis; Testing

ABSTRACT

Matching, randomization, random sampling, and other methods of controlling extraneous variation are discussed. The purpose was to specify the benefits of randomization in estimating causal effects of treatments. It is concluded that randomization should be employed whenever possible, but the use of carefully controlled nonrandomized data to estimate causal effects is a reasonable and necessary procedure in many cases. (Author)

RB-72-39

RESEARCH BULLETIN

ESTIMATING CAUSAL EFFECTS OF TREATMENTS IN EXPERIMENTAL AND OBSERVATIONAL STUDIES

Donald Rubin

Educational Testing Service
Princeton, New Jersey
August 1972

ESTIMATING CAUSAL EFFECTS OF TREATMENTS IN EXPERIMENTAL AND OBSERVATIONAL STUDIES

Donald Rubin

Abstract

A discussion of matching, randomization, random sampling, and other methods of controlling extraneous variation is presented. The objective is to specify the benefits of randomization in estimating causal effects of treatments. The basic conclusion is that randomization should be employed whenever possible, but the use of carefully controlled nonrandomized data to estimate causal effects is a reasonable and necessary procedure in many cases.

ESTIMATING CAUSAL EFFECTS OF TREATMENTS IN EXPERIMENTAL AND OBSERVATIONAL STUDIES

Donald Rubin

Educational Testing Service

1. Introduction

Recent psychological and educational literature has included extensive criticism of the use of observational studies to estimate causal effects of treatments (see, e.g., Campbell & Erlebacher, 1970). The implication in much of this literature is that only properly randomized experiments can lead to useful estimates of causal treatment effects. If taken as applying to all fields of study, this position is clearly untenable. Since the extensive use of randomized experiments is limited to the last half century,[1] and in fact is not used in much scientific investigation today, one is led to the conclusion that most scientific "truths" have been established without using randomized experiments. In addition, most of us successfully establish the causal effects of many of our everyday actions, even interpersonal behaviors, without the benefit of randomization.

Even if the position that causal effects of treatments can only be well established from randomized experiments is taken as applying only to the social sciences in which there are currently few well-established causal relationships, its implication to ignore existing observational data may be counter-productive. Often the only immediately available data are observational in nature and either (a) the cost of performing the equivalent experiment to test all treatments is prohibitive (e.g., 100 reading programs

[1] Essentially since Fisher (1925).

under study); (b) there are ethical reasons why the treatments cannot be randomly assigned (e.g., estimating the effects of heroin addiction on intellectual functioning); or (c) estimates based on results of experiments would be delayed many years (e.g., effect of childhood intake of cholesterol on longevity). In cases such as these, it seems more reasonable to try to estimate the effects of the treatments from the observational data than to ignore the observational data and dream of the ideal experiment. Concurrently, using the indications in the observational data, one can, if necessary, initiate randomized experiments for those treatments that require better estimates.

The position here is not that randomization is overused. On the contrary, if the choice is between the data from a randomized experiment and an equivalent observational study, one should choose the data from the experiment, especially in disciplines like the social sciences where often much of the variability is unassigned to particular causes. However, by examining the assumptions that are needed in order to believe that the results of an experiment or an observational study yield appropriate answers to questions about the causal effects of treatments, we will develop the position that observational studies as well as randomized experiments can be useful in estimating causal treatment effects.

In order to avoid unnecessary complication, we will restrict discussion to the very simple study consisting of $2N$ units (e.g., subjects), half having been exposed to an experimental treatment E (e.g., a compensatory reading program) and the other half having been exposed to a control treatment C (e.g., a regular reading program). If treatments E and C were

assigned to the 2N units randomly, that is, using some mechanism that assured each unit was equally likely to be exposed to E as to C , then the study is called a randomized experiment or more simply an experiment; otherwise, the study is called a quasi-experiment or an observational study. In either case, the objective is to determine for some general group of units (e.g., underprivileged 6th grade children) the "typical" causal effect of the E vs. C treatment on a dependent variable Y , where Y could be dichotomous (e.g., success-failure) or more continuous (e.g., score on a given reading test).

The central question concerns the benefits of randomization in determining the causal effect of the E vs. C treatment on Y .

2. Defining the Causal Effect of the E vs. C Treatment

A first step in investigating the benefits of randomization for determining the causal effect of treatments is to define exactly what is meant by the causal effect of a treatment. Intuitively, the causal effect of one treatment, E , over another, C , for a particular unit and an interval of time from t_1 to t_2 is the difference between what would have happened at time t_2 if the unit had been exposed to E at time t_1 and what would have happened at t_2 if the unit had been exposed to C at t_1 : "If an hour ago I had taken two aspirins instead of just a glass of water, my headache would now be gone," or "Because an hour ago I took two aspirins instead of just a glass of water, my headache is now gone." Our definition of the causal effect of the E vs. C treatment will reflect this intuitive meaning.

First define a trial to be a unit (e.g., a subject) and an associated pair of times, t_1 and t_2 , where t_1 denotes the time of initiation of a

treatment and t_2 denotes the time of measurement of a dependent variable, Y, where $t_1 < t_2$.

We restrict our attention to treatments E and C that could be randomly assigned; thus, we assume (a) a time of initiation of treatment can be ascertained for each unit exposed to E or C, and (b) E and C are exclusive of each other in the sense that a trial cannot simultaneously be an E trial and a C trial (i.e., if E is defined to be C plus some action, the initiation of both is the initiation of E; if E and C are alternative actions, the initiation of both E and C is the initiation of neither but rather a third treatment $E + C$).

We also assume that the measured value of Y stated with reference to time t_2 is the "true" value of Y at t_2. This position can be justified by defining Y by a measuring instrument that always yields the measured Y (e.g., Y is the score on a particular IQ test as recorded by the subject's teacher). Since an "error" in the measured Y can only be detected by a "better" measuring instrument (e.g., a machine-produced score on that same IQ test) the values of a "truer" score can be viewed as the values of a different dependent variable. Clearly, any study is more meaningful to the investigator if the dependent variable better reflects underlying concepts he feels are important (e.g., is more accurate) but that does not imply we must consider errors about some unmeasurable "true score."[2]

---

[2]For the reader who prefers the concept of such errors of measurement he may consider the following discussion to assume negligible "technical errors" so that Y is essentially the "true" Y.

Now define the causal effect of the E vs. C treatment on Y for a particular trial (i.e., a particular unit and associated times t_1 , t_2) as follows:

> Let y(E) be the value of Y measured at t_2 on the unit given that the unit received the experimental treatment E initiated at t_1 ;
>
> Let y(C) be the value of Y measured at t_2 on the unit given that the unit received the control treatment C initiated at t_1 ;
>
> Then y(E) - y(C) is the causal effect of the E vs. C treatment on Y for that trial--i.e., for that particular unit and the times t_1 , t_2 .

For example, assume the unit is a particular rat and that the experimental treatment is a high protein diet and the control treatment is a regular diet. Suppose that if the rat were given the high protein diet initiated at time t_1 , 10 days later, at time t_2 , he would weigh 7 oz. and suppose that if the rat instead were given the regular diet initiated at time t_1 , at time t_2 he would weigh 6 oz. Then the causal effect for that trial (that rat and times t_1 , t_2) of the high protein diet vs. the regular diet on weight is 7 - 6 = 1 oz.

The problem in measuring y(E) - y(C) is of course that we can never observe both y(E) and y(C) since we cannot return to time t_1 to give the other treatment. We may have the same unit measured on both treatments in two trials (a repeated measure design), but since there may exist carry-over effects (e.g., the effect of the first treatment wears off slowly) or general time-trends (e.g., as the rat ages, his reactions become slower) we cannot be certain that the unit's responses will be identical at both times.

Assume now that there are M trials for which we want the "typical" causal effect. For simplicity of exposition we assume that each trial is

associated with a different unit and expand the above notation by adding the subscript j to denote the j^{th} trial ($j = 1,2,\ldots,M$); thus $y_j(E) - y_j(C)$ is the causal effect of the E vs. C treatment for the j^{th} trial, i.e., the j^{th} unit and the associated times of initiation of treatment, t_{1j}, and measurement of Y, t_{2j}.

An obvious definition of the "typical" causal effect of the E vs. C treatment for the M trials is the average causal effect for the M trials:

$$\frac{1}{M}\sum_{j=1}^{M}[y_j(E) - y_j(C)].$$

However, notice that if all but one of the individual causal effects are small and that one is very large, the average causal effect may be substantially larger than all but one of the individual causal effects and thus not very "typical." Other possible definitions of the typical causal effect for the M trials are the median causal effect (the median of the individual causal effects) or the midmean causal effect (the average of the middle half of the individual causal effects). If the individual causal effects, $y_j(E) - y_j(C)$, are approximately symmetrically distributed about a central value, sensible definitions of "typical" will yield similar values.

Even though definitions of typical other than the average may seem more reasonable, they lead to more complications when discussing properties of estimates under randomization. Hence we will assume the average causal effect to be the desired typical causal effect for the M trials and proceed to the problem of its estimation given the obvious constraint that we can never actually measure both $y_j(E)$ and $y_j(C)$ for any trial.

3. Randomization, Matching, and Estimating the Typical Causal Effect in the 2N Trial Study

Having defined the typical causal effect of the E vs. C treatment, we proceed to discuss the benefits of randomization and matching in a $2N$ trial study, N trials being with units exposed to E and the other N trials being with units exposed to C.

For now we assume that the immediate objective is to estimate the typical causal effect only for the $2N$ trials in the study. Of course, in order for the results of a study to be of much interest, we must be able to generalize to units and associated times other than those in the study. However, the issue of generalizing results to other trials will be discussed separately from the issue of estimating the typical causal effect for the trials under study. Also, for now we will consider only the simple and standard estimate of the typical causal effect of E vs. C : the average Y difference between those units who received E and those units who received C.

We begin by considering this estimate when only two trials are in the study and then when there are $2N$, $N > 1$, trials in the study. This informal discussion will then lead to the presentation of two formal benefits of randomization.

3.1 The Two Trial Study

Assume two trials under study, one trial with a unit exposed to E and the other with a unit exposed to C. The typical causal effect for the two trials is

$$\frac{1}{2}\ [y_1(E) - y_1(C) + y_2(E) - y_2(C)] \quad . \qquad (1)$$

Note that for only two trials almost any reasonable definition of typical leads to this expression. The estimate of this quantity from the study, the (average) difference between the measured Y for the unit who received E and the measured Y for the unit who received C, is either

$$y_1(E) - y_2(C) \qquad (2)$$

or

$$y_2(E) - y_1(C) \qquad (3)$$

depending upon which unit was assigned E. Neither (2) nor (3) will necessarily be close to (1) or to the causal effect for either unit

$$y_1(E) - y_1(C) \qquad (4)$$

or

$$y_2(E) - y_2(C) \qquad (5)$$

even if these individual causal effects are equal. If the treatments E and C were randomly assigned to units, we are equally likely to have observed the difference (2) as (3) so that the average or "expected" difference in Y between experimental and control units is the average of (2) and (3), $\frac{1}{2}\ [y_1(E) - y_2(C)] + \frac{1}{2}\ [y_2(E) - y_1(C)]$ which equals (1), the typical causal effect for the two trials. For this reason, if the treatments are randomly assigned, the difference in Y between the experimental and control units is called an "unbiased" estimate of the desired typical causal effect.

Now assume that the two units are very similar in the way they respond to the E and C treatments at the times of their trials. By this we mean that on the basis of "extra information" we know $y_1(E)$ is about equal to $y_2(E)$ and

$y_1(C)$ is about equal to $y_2(C)$; that is, the two trials are closely "matched" with respect to the effects of the two treatments. It then follows that (2) is about equal to (3) and both are about equal to the desired typical causal effect (1). In fact, if the two units react identically in their trials, (5) = (4) = (3) = (2) = (1), and randomization is absolutely irrelevant. Clearly, having closely "matched" trials increases the closeness of the calculated-experimental minus control difference to the typical causal effect for the two trials, while random assignment of treatments does not improve that estimate.

Although two trial studies are almost unheard of in the behavioral sciences they are not uncommon in the physical sciences, as the reader might recall from high school physics or chemistry laboratories. For example, when comparing the heat expansion rates (per hour) of a metal alloy in oxygen and nitrogen an investigator might use two one-foot lengths of the alloy. Because the lengths of alloy are so closely matched before being exposed to the treatment (almost identical compositions and dimensions) the units should respond almost identically to the treatments even when initiated at different times, and thus the calculated experimental (oxygen) minus control (nitrogen) difference should be an excellent estimate of the typical causal effect, (1).

Notice, however, that a skeptical observer could always claim that the experimental minus control difference is not a good estimate of the typical causal effect of the E vs. C treatment because the two units were not absolutely identical prior to the application of the treatments. For example, he could claim that the length of alloy molded first would expand more rapidly. Hence, he might argue that what was measured was really the effect of the difference in order of manufacture, not the causal effect of the oxygen vs.

nitrogen treatment. Since units are never absolutely identical before the application of treatments, this kind of argument, whether "sensible" or not, can always be made. However, if the two trials are closely matched with respect to the expected effects of the treatments, that is, if (a) the two units are matched prior to the initiation of treatments on all variables thought to be important in the sense that they causally affect Y, and (b) the possible effect of different times of initiation of treatment and measurement of Y are controlled, then the investigator can be confident that he is in fact measuring the causal effect of the E vs. C treatment for those two trials. This kind of confidence is much easier to generate in the physical sciences where there are models that successfully assign most variability to specific causes than in the social sciences where often we do not know what the important causal variables are.

Another source of confidence that the experimental minus control difference is a good estimate of the causal effect of E vs. C is replication: are similar results obtained under similar conditions. One type of replication is the inclusion of more than two trials in the study. Hence we now turn to the discussion of the study with $2N$ trials.

3.2 The 2N Trial Study

Now assume there are $2N$ trials ($N > 1$) in the study, half with N units having received the E treatment and the other half with N other units having received the C treatment. The immediate objective is to find the typical causal effect of the E vs. C treatment on Y for the $2N$ trials, say τ:

$$\tau = \frac{1}{2N} \sum_{j=1}^{2N} [y_j(E) - y_j(C)] \quad .$$

Let S_E denote the set of indices of the N E trials and S_C denote the set of indices of the N C trials ($S_E \cup S_C = \{i = 1,2,\ldots,2N\}$). Then the difference between the average observed Y in the E trials and the average observed Y in the C trials can be expressed as

$$\bar{y}_d = \frac{1}{N} \sum_{j \in S_E} y_j(E) - \frac{1}{N} \sum_{j \in S_C} y_j(C) \quad ,$$

where $\sum_{j \in S_E}$ and $\sum_{j \in S_C}$ indicate, respectively, summation over all indices in S_E (i.e., all E trials) and over all indices in S_C (i.e., all C trials). There are $\binom{2N}{N}$ different possible index sets S_E corresponding to the distinct ways of choosing N different numbers from a total of 2N different numbers, where the binomial coefficient $\binom{2N}{N} = [2N]! \,/\, [N!]^2$. Depending upon which N of the 2N units received E we observe one and only one of those $\binom{2N}{N}$ possible allocations. We now consider how close this estimate $\bar{y}_d$ is to the typical causal effect τ and what advantage there might be if we knew the treatments were randomly assigned.

First assume that for each unit receiving E there is a unit receiving C who reacts identically at the times of their trials; that is, the 2N trials are actually N perfectly matched pairs. It is a simple extension of the discussion in Section 3.1 to see that the estimate $\bar{y}_d$ in this case equals τ . $\bar{y}_d$ can be expressed as the average experimental minus control (E - C) difference across the N matched trials. Since the E - C difference in each matched pair of trials is the typical causal effect for both trials of that pair, the average of those differences is the typical causal effect for

all N pairs and thus all $2N$ trials.[3] Again this result holds whether the treatments were randomly assigned or not. In fact, if one had N identically matched pairs a "thoughtless" random assignment could be worse than a nonrandom assignment of E to one member of the pair and C to the other. By "thoughtless" we mean some random assignment that does not assure that the members of each matched pair get different treatments--picking the N indices to receive E "from a hat" containing the numbers 1 through $2N$, rather than tossing a fair coin for each matched pair to see which unit is to receive E.[4]

In practice of course we never have exactly matched trials. However, if matched pairs of trials are very similar in the sense that the investigator has controlled those variables prior to the initiation of treatments that might appreciably affect Y, $\bar{y}_d$ should be close to τ. If in addition the estimated causal effect is replicable in the sense that the N individual estimated causal effects for each matched pair are very similar, the investigator might feel even more confident that he is in fact estimating the typical causal effect for the $2N$ trials. For example, given $2N$ rats from the same litter matched by sex and initial weight into N pairs, assume that we observe the same $E - C$ difference in final weight in each matched pair. Similarly, if the trials are not pair-matched but all are similar (e.g., all rats are mature males from the same litter with similar weights),

---

[3] Incidentally, notice that in this case we can calculate the typical causal effect for any definition of typical, e.g., the median of the observed matched pair differences is the median causal effect for the $2N$ trials.

[4] The difference between these two methods of randomization is the difference between the "completely randomized" experiment and the "randomized blocks" experiments (see Cochran and Cox, 1957).

and we observe that all $y_j(E)$ $j \varepsilon S_E$ are about equal and all $y_j(C)$ $j \varepsilon S_C$ are about equal, the investigator would also feel confident that he is in fact estimating the typical causal effect for the $2N$ trials.

Nevertheless, it is obvious that if treatments were systematically assigned to units, the addition of replication evidence cannot dissuade the critic who believes the effect being measured is due to a variable used to assign treatments (e.g., in the weight-gain study if the more active rat always received the special diet, or in the heat-expansion study if the first molded alloy always was measured in oxygen). If treatments were randomly assigned, all systematic sources of bias would be made random, and thus it would be unlikely, especially if N is large, that almost all E trials would be with the more active rat or the first molded alloy, so that any effect of that variable would be at least partially balanced in the sense of systematically favoring neither the E treatment nor the C treatment over the $2N$ trials. In addition, using the replications there could be evidence to refute the skeptic's claim of the importance of that variable (e.g., in each matched trial we get about the same estimate whether the more active rat gets E or C). Of course, if we knew beforehand of the skeptic's claim, a specific control of this additional variable would be more advisable than relying on randomization (e.g., in a random half of the matched trials assign E to the more active rat and in the other half assign C to the more active rat, or include rat's activity as a matching variable).

It is important to realize, however, that whether treatments are randomly assigned or not, no matter how carefully matched the trials, and no matter how large N, a skeptical observer could always eventually find some

variable that systematically differs in the E trials and C trials (e.g., length of longest hair on the rat) and claim that $\bar{y}_d$ estimates the effect of this variable rather than τ, the causal effect of the E vs. C treatment. Within the given experiment there will be no refutation of his claim; only a logical argument explaining that the variable cannot causally affect the dependent variable or additional data outside the study can be used to counter his position.

4. Two Formal Benefits of Randomization

If randomization can never assure us that we are correctly estimating the causal effect of E vs. C for the 2N trials under study, what are the benefits of randomization besides the intuitive ones that follow from making all systematic sources of bias into random ones? Formally, randomization provides a mechanism to derive probabilistic properties of estimates without making further assumptions. We will consider two such properties which are important:

(1) the average E - C difference is an "unbiased" estimate of τ, the typical causal effect for the 2N trials; and

(2) precise probabilistic statements can be made indicating how unusual the observed E - C difference, $\bar{y}_d$, would be under specific hypothesized causal effects.

More advanced discussion of the formal benefits of randomization may be found in Scheffé (1959) and Kempthorne (1952).

4.1 Unbiased Estimation over the Randomization Set

We begin by defining the "randomization set" to be the set of r allocations that were equally likely to be observed given the randomization

plan. For example, if the treatments were randomly assigned to trials with no restrictions (the completely randomized experiment), each one of the $\binom{2N}{N}$ possible allocations of N trials to E and N trials to C was equally likely to be the one observed allocation. Thus, the collection of all of these $r = \binom{2N}{N}$ allocations is known as the randomization set for this completely randomized experiment. If the treatments were assigned at random within matched pairs (the randomized blocks experiment), any allocation with both members of the pair assigned the same treatment could not be observed; the remaining 2^N allocations with each member of the pair receiving a different treatment was equally likely to be the observed one. Hence, for the experiment with randomization done within matched pairs, the collection of these $r = 2^N$ equally likely allocations is known as the randomization set.[5]

For each of the r allocations in the randomization set there is a corresponding average E - C difference that we would have calculated had that allocation been chosen. If the expectation (i.e., average) of these r average differences equals τ, the average E - C difference is called unbiased over the randomization set for estimating τ. We now show that given randomly assigned treatments, the average E - C difference is an unbiased estimate of τ, the typical causal effect for the 2N trials.

By the definition of random assignment, each trial is equally likely to be an E trial as a C trial. Hence, the contribution of the j^{th} trial $(j = 1,\ldots,2N)$ to the average E - C difference in half of the r allocations in the randomization set is $\frac{1}{N}\, y_j(E)$ and in the other half is $-\frac{1}{N}\, y_j(C)$; thus, the expected contribution of the j^{th} trial to the average E - C

[5] There are of course other methods of randomly assigning two treatments to 2N trials but there is no need to consider them here.

difference is $\frac{1}{2}\ [\frac{1}{N}\ y_j(E)] + \frac{1}{2}\ [-\frac{1}{N}\ y_j(C)]$. Adding over all $2N$ trials we have that the expectation of the average $E - C$ difference over the r allocations in the randomization set is

$$\frac{1}{2N} \sum_{j=1}^{2N} [y_j(E) - y_j(C)] \quad ,$$

which is the typical causal effect for the 2N trials, τ .

Although the unbiasedness of the $E - C$ difference is appealing in the sense that it is an indication that we are tending to estimate τ , its impact is not immediately overwhelming: the one $E - C$ difference we have observed, $\bar{y}_d$, may or may not be close to τ . In a vague sense we may believe $\bar{y}_d$ should be close to τ because the unbiasedness indicates that "on the average" the $E - C$ difference is τ , but this belief may be tempered when other properties of the estimate are revealed; for example, without additional assumptions about the symmetry of effects the average $E - C$ difference is <u>not</u> equally likely to be above as below τ .

In addition, after observing the values of some important unmatched variable we may no longer believe $\bar{y}_d$ tends to estimate τ . For example, suppose in the study of the effect of diet on rats' weight, initial weight is not a matching variable, and after the experiment is complete we observe that the average initial weight of the rats exposed to E was higher than the average initial weight of the rats exposed to C . Clearly we would now believe that $\bar{y}_d$ probably overestimates τ even if treatments were randomly assigned.

In sum, then, the unbiasedness of the $E - C$ difference for τ follows from the random assignment of treatments; it is a desirable property in that it indicates "on the average" we tend to estimate the correct quantity but it hardly solves the problem of estimating the typical causal effect. We as

yet have no indication whether to believe $\bar{y}_d$ is close to τ nor any ability to adjust for important information we may possess.

4.2 Probabilistic Statements from the Randomization Set.

A second formal advantage of randomization is that it provides a mechanism for making precise probabilistic statements indicating how unusual the observed E - C difference, $\bar{y}_d$, would be under specific hypotheses. The following discussion of "significance levels" derived from the randomization set will tend to be more technical than the other discussion in this paper but is still basically straightforward.

Assume that the investigator hypothesizes exactly what the individual causal effects are for each of the 2N trials and these hypothesized values are $\tilde{\tau}_j$, $j = 1,\ldots,2N$. The hypothesized typical causal effect for the 2N trials is thus

$$\tilde{\tau} = \frac{1}{2N} \sum_{j=1}^{2N} \tilde{\tau}_j \quad .$$

Assuming the $\tilde{\tau}_j$ are correct and having the observed $y_j(E)$, $j \varepsilon S_E$ and $y_j(C)$, $j \varepsilon S_C$, we can calculate hypothesized values, say $\tilde{y}_j(C)$ and $\tilde{y}_j(E)$, for all of the 2N trials. For $j \varepsilon S_E$, $y_j(E)$ is observed and $y_j(C)$ is unobserved; hence for these trials $\tilde{y}_j(E) = y_j(E)$ and $\tilde{y}_j(C) = y_j(E) - \tilde{\tau}_j$. For $j \varepsilon S_C$, $y_j(C)$ is observed and $y_j(E)$ is unobserved; hence for these trials $\tilde{y}_j(C) = y_j(C)$ and $\tilde{y}_j(E) = y_j(C) + \tilde{\tau}_j$. Thus, assuming the hypothesized $\tilde{\tau}_j$ are correct, we can calculate hypothesized $\tilde{y}_j(E)$ and $\tilde{y}_j(C)$ for all 2N trials. Then using these, we can calculate an hypothesized average E - C difference for each of the r allocations of the 2N trials in the randomization set.

Assume that we calculate all r hypothesized average E - C differences and list them from high to low noting which E - C difference corresponds to the S_E, S_C allocation we have actually observed. This difference, $\bar{y}_d$, is the only one which does not use the hypothesized $\tilde{\tau}_j$. If treatments were assigned completely at random to the trials and the hypothesized $\tilde{\tau}_j$ are correct, any one of the $r = \binom{2N}{N}$ differences was equally likely to be the observed one; similarly, if treatments were randomly assigned within matched pairs, each of the $r = 2^N$ differences with each member of a matched pair getting a different treatment was equally likely to be the observed one. Intuitively, if the hypothesized $\tilde{\tau}_j$ are essentially correct, we would expect the observed difference $\bar{y}_d$ to be rather typical of the $(r - 1)$ other differences that were equally likely to be observed; that is, $\bar{y}_d$ should be near the center of the distribution of the r E - C differences. If, in fact, the observed difference is in the tail of the distribution and so not typical of the r differences we might doubt the correctness of the hypothesized $\tilde{\tau}_j$.

More formally we proceed as follows. The average of the r E - C differences is in fact the hypothesized typical causal effect, $\tilde{\tau} = \frac{1}{2N}\sum\tilde{\tau}_j$. This result follows immediately from the unbiasedness of the E - C difference for the actual typical causal effect τ. Now, using the equal likeliness of the r allocations we can make the following kind of probabilistic statement: "Under the hypothesis that the causal effects are given by the $\tilde{\tau}_j$, $j = 1,\ldots,2N$, the probability that we would observe an average E - C difference as far or farther from $\tilde{\tau}$ than the one we have observed is m/r where m is the number of allocations in the randomization set that yield E - C differences as far or farther from $\tilde{\tau}$ than $\bar{y}_d$." If this

probability, called the "significance level" for the hypothesized $\tilde{\tau}_j$, is very small, that is, if the observed $\bar{y}_d$ is farther from $\tilde{\tau}$ than most of the other differences in the randomization set, we either must admit that the observed value was unusual in the sense of being in the tail of the distribution of the equally likely differences, or we must reject the plausibility of the hypothesized $\tilde{\tau}_j$.

The most common hypothesis for which a significance level is calculated is that the E vs. C treatment has no effect on Y whatsoever (i.e., $\tilde{\tau}_j = 0$). Other common hypotheses assume that the effect of the E vs. C treatment on Y is a nonzero constant (i.e., $\tilde{\tau}_j = \tilde{\tau}_o$) for all trials.[6]

The ability to make precise probabilistic statements about the observed $\bar{y}_d$ under various hypotheses without additional assumptions is a tremendous benefit of randomization especially since $\bar{y}_d$ tends to estimate τ. However, one must realize that these simple probabilistic statements refer only to the 2N trials used in the study and do not reflect additional information (i.e., other variables) that we may have measured.

5. Additional Assumptions Often Needed to Present the Results of a Study as Being of General Interest

There are two kinds of issues that have been mentioned that often arise when presenting the results of an experiment as being relevant and which so far have not been handled in our discussion of randomization and matching.

---

[6] These hypotheses for a constant effect can be used to form "confidence limits" for τ. Given that the τ_j are constant, the set of all hypothesized $\tilde{\tau}_o$ such that the associated significance level is greater than or equal to $\alpha = m/r$ form a $(1 - \alpha)$ confidence interval for τ: of the r such $(1 - \alpha)$ confidence intervals one could have constructed (one for each of the r allocations in the randomization set), $r(1 - \alpha) = r - m$ of them include the true value of τ assuming all $\tau_j = \tau$. See Lehman (1959, p. 59) for the proof.

The first concerns additional variables not explicitly controlled in the experiment: the thoughtful investigator must be prepared to consider the effect of variables that may systematically differ in E trials and C trials. The second issue concerns the ability to generalize the results: the investigator must be able to indicate the applicability of his results to a population of trials other than the $2N$ in the study.

5.1 Considering Additional Variables

As has been indicated in our previous discussion, in most studies whether observational or experimental, the investigator should be prepared to consider the possible effect of other variables besides those explicit in the experiment. Often additional variables will be ones that the investigator considers relevant in the sense that he feels they may causally affect Y; therefore, he may want to adjust the estimate $\bar{y}_d$ and significance levels of hypotheses to reflect the values of these variables in his study. At times the variables will be ones which he feels cannot causally affect Y even though in his study they may be correlated with the observed values of Y. An investigator who refuses to consider any additional variables brought to his attention is in fact saying that he does not care if $\bar{y}_d$ is a bad estimate of the typical causal effect of the E vs. C treatment and instead is satisfied with mathematical properties (i.e., unbiasedness) of the process by which he calculated it.

Consider first the case of an obviously important variable. As an example assume in the rat weight gain study, with diets randomly assigned we found that the average $E - C$ difference in final weight was 1 oz. and that under the hypothesis of no effects the significance level was .01; also assume that initial weight was not a matching variable and in fact the difference in

initial weight was also 1 oz. Admittedly, this is probably a rare event given the randomization but rare events do happen rarely. Given that it did happen we would indeed be foolish to believe $\bar{y}_d = 1$ oz. is a good estimate of τ and/or the implausibility of the hypothesis of no treatment effects indicated by the .01 significance level. Rather, it would seem more sensible to believe that $\bar{y}_d$ overestimates τ, and significance levels underestimate the plausibility of hypotheses that suggest no or negative effects for the treatments.

A commonly used and obvious correction is to calculate the average E - C difference in gain score rather than final score. That is, for each trial there is a "pretest" score (e.g., initial weight) which was measured before the initiation of treatments, and the gain score for each trial is the final score minus the pretest score. More generally we will speak of a "prior" score or "prior" variable which would have the same value, x_j, whether the j^{th} unit received E or C.[7] It then follows given random assignment of treatments that the adjusted estimate (e.g., gain score)

$$\frac{1}{N} \sum_{j \epsilon S_E} [y_j(E) - x_j] - \frac{1}{N} \sum_{j \epsilon S_C} [y_j(C) - x_j]$$

remains an unbiased estimate of τ over the randomization set: each prior score appears in half of the equally likely allocations as $\frac{1}{N} x_j$ and the other half as $-\frac{1}{N} x_j$; hence, averaged over all allocations the j^{th} prior

[7] Even though "prior" indicates that the variable attained its value for all trials prior to the initiation of the treatments, a prior variable can be any variable that cannot be causally affected by the treatments and thus would have the same value whether the unit received E or C.

score has no effect.[8] But this result holds for any set of prior scores x_j, $j = 1,\ldots,2N$, whether sensible or not. For example, in an experiment evaluating a compensatory reading program, with Y being the final score on a reading test, the prior score "pretest reading score" or perhaps "IQ" properly scaled makes sense but "height in millimeters" does not. Also why not use the prior score "one-half pretest score"?

Clearly, in order to make an intelligent adjustment for extra information we cannot be guided solely by the concept of unbiasedness over the randomization set. We need some model for the effect of the prior variables in order to use their values in an intelligent manner. The gain score, for example, assumes that the final score typically would equal the initial score if there were no $E - C$ treatment effect and is perfectly reasonable for the length of the alloys in the heat expansion experiment or the weight of mature rats in the diet experiment. In the physical sciences, more complex models which represent generally accepted functional relationships are often used; however, in the social sciences there are rarely such accepted relationships to rely upon. What then does the investigator do who wants to adjust intelligently the final reading scores for the subjects' varying IQ's, grade levels, SES, and so on? Clearly, he must be willing to make some assumptions about the functional form of the causal effect of these other variables on Y. If he assumes, perhaps based on indications in previous data, some "known" function for x_j (e.g., in the compensatory reading program example, suppose x_j equals $[.01 \times IQ]^2 \times$ pretest $\times$ [percentile of family income]), so that x_j is the same whether the j^{th} unit received

---

[8] If the prior score could vary depending on whether the unit received E or C (i.e., it is a variable measured after the initiation of the treatments that may be causally affected by the treatment) we would have no assurance that the adjusted $E - C$ difference is an unbiased estimate over the randomization set.

E or C , from the previous discussion the average E - C difference in adjusted scores remains an unbiased estimate of τ . If the investigator assumes a model whose parameters are unknown and estimates these parameters by some method from the data, in general, the average E - C difference in adjusted scores is no longer unbiased over the randomization set because the adjustment for the j^{th} trial depends on which trials received E and which received C (e.g., in the analysis of covariance, the estimated regression coefficients in general vary over the r allocations in the randomization set).

Clearly, forming an intelligent adjusted estimate may not be simple even in a randomized experiment. However, significance levels for any adjusted estimate can be found by calculating the adjusted estimate rather than the simple E - C difference for each of the r equally likely allocations in the randomization set. Nonetheless, if the adjusted estimate does not tend to estimate τ in a sensible manner, the resulting significance level may not be of much interest.

Now consider a variable that is brought to the investigator's attention but he feels cannot causally affect Y (e.g., in the compensatory reading example, age of oldest living relative). Eventually a skeptic can find such a variable that systematically differs in the E trials and the C trials even in the best of experiments. Considering only that variable it is indeed unlikely given randomization that there would be such a discrepancy between its values in E trials and C trials, but its occurrence cannot be denied. If the skeptic adjusts $\bar{y}_d$ by using a standard model (e.g., covariance), the adjusted estimate and related significance levels may then give misleading results (e.g., zero estimate of τ , hypothesis that all causal effects are zero, $\tilde{\tau}_j \equiv 0$, is very plausible). In fact, using such models one can obtain

any estimated causal effect desired by searching for and finding a prior variable or combination of prior variables that yield the desired result.[9] Such a search will in a sense be more difficult if randomization was performed, but clearly, even with randomized data, the investigator must be prepared to ignore variables that he feels cannot causally affect Y. On the other hand, he may want to adjust for such a variable if he feels it is a surrogate for an unmeasured variable that can causally affect Y (e.g., age of oldest living relative as a surrogate for mental stability of the family in the compensatory reading example).

The point of this discussion is that when trying to estimate the typical causal effect in the 2N trial experiment, handling outside information may not be trivial without a well-developed causal model that will properly adjust for those prior variables that causally affect Y and ignore other variables even if they are highly correlated with the observed values of Y. Without such a model, the investigator must be prepared to ignore some variables he feels cannot causally affect Y and use a possibly arbitrary model to adjust for those variables he feels are important.

5.2 Generalizing Results to Other Trials

In order to believe that the results of an experiment are of interest we generally must believe that the 2N trials in the study are representative of a population of other future trials. For example, if the experimental treatment is a compensatory reading program and the trials are composed of 6th grade school children with treatments initiated in fall 1970 and Y measured in Spring 1971, the results are of little interest unless we believe they

---

[9] Consider for example adjusting for 2N covariates in a 2N trial study.

tell us something about future 6th graders who might be exposed to the compensatory reading program.

For simplicity assume the 2N trials in the study are a simple random sample from a "target population" of M trials to which we want to generalize the results; by simple random sample we mean that each of the M trials is equally likely to be used in the study, or equivalently, each of the $\binom{M}{2N}$ ways of choosing the 2N trials is equally likely. If τ is the typical (average) causal effect for all M trials, it then follows given random assignment of treatments that the average E - C difference for the 2N trials used is an unbiased estimate of τ over the random sampling plan and over the randomization set. In other words, in each of the $r \times \binom{M}{2N}$ ways of choosing 2N trials from M trials and then randomly assigning N trials to E and N trials to C there is a calculated average E - C difference, and the average of these $r \times \binom{M}{2N}$ differences is τ : because of the randomization and random sampling each trial is equally likely to be an E trial as a C trial and thus contributes $\frac{1}{N}\ y_j(E)$ to the E - C difference as often as it contributes $-\frac{1}{N}\ y_j(C)$. It also follows that under a hypothesized set of causal effects, $\tilde{\tau}_j$, $j = 1,\ldots,M$, the significance level (the probability that we would observe a difference as large as or larger than $\bar{y}_d$) given that we have sampled the 2N trials in the study is m/r where m is the number of allocations in the randomization set that yield estimates as far or farther from $\tilde{\tau}$ than $\bar{y}_d$.[10]

If we let M grow to infinity (a reasonable assumption in many experiments when the population to which we want to generalize results is

---

[10] Even though we have hypothesized $\tilde{\tau}_j$ for all trials we cannot calculate hypothesized $\tilde{y}_j(E)$ and $\tilde{y}_j(C)$ for the unsampled trials, and thus the probabilisitc statement is conditional on the observed trials.

essentially unlimited, e.g., all future 6th grade students), the stating of probabilistic results is facilitated. For example, the usual covariance adjusted estimate is an unbiased estimate of T (not necessarily τ) over the random sampling plan and the randomization set, even though whether the adjustment actually adjusts for the additional variable(s) still depends on the appropriateness of the underlying linear model.

Hence, given random sampling of trials the ability to generalize results to other trials seems relatively straightforward probabilistically. However, most experiments are designed to be generalized to <u>future</u> trials and we never have a random sample of trials from the future but at best a random sample from the present. Generally, in fact, observational studies probably have more representative trials than experiments many of which are conducted in constrained, atypical environments and within a restricted period of time. Thus, in order to generalize the results of any experiment to future trials of interest, we minimally must believe that there is a similarity of effects across time and more often must believe that the trials in the study are "representative" of the population of trials. This step of faith may be called making an assumption of "subjective random sampling" in order to assert such properties as (a) $\bar{y}_d$ (or $\bar{y}_d$ adjusted) tends to estimate the typical causal effect T and (b) the plausibility of hypothesized $\tilde{\tau}_j$, $j = 1,\ldots,M$, is given by the usual conditional significance level.

As indicated above, this subjective random sampling is quite possibly easier to believe in an observational study with data drawn from many sources than in an experiment performed under controlled conditions. Even so, investigators do make and must be willing to make this step in experiments in order to believe their results are useful; when investigators carefully indicate their sample of trials and the ways in which they may differ from

those in the target population this tacit assumption of subjective random sampling seems perfectly reasonable. If there is an important variable that differs between the sample of trials and the population of trials, an attempt to adjust the estimate based on the same kinds of models discussed previously is quite appropriate even if the sample is actually a random sample.[11] If the sample is not actually a random sample, and the model for this adjustment is reasonable, such an adjustment should make the assumption of subjective random sampling even more plausible.

6. Subjective Randomization and Observational Studies

Now consider a carefully controlled observational study--a study in which there are no obviously important prior variables that systematically differ in the E trials and the C trials. In such a study there is a real sense in which a claim of "subjective randomization" can be made. For example, if the study were composed of carefully matched pairs of trials, there might be a very defensible belief that within each matched pair each unit was equally likely to receive E as C in the sense that if I showed the units to you without telling you which received E, only half the time would you guess correctly which received E.[12] Under this assumption of subjective randomization the usual estimates and significance levels can be used as if the study had been randomized; this step is analogous to the step of assuming subjective random sampling in order to make inferences about a target population.

---

[11] See Cochran (1963) on regression and ratio adjustments.

[12] Perhaps this is all that is meant by "randomization" to some Bayesians under any circumstances (see Savage, 1954, p. 66).

If an obviously important prior variable were found to differ systematically in the E and C trials, we would of course have to adjust the estimate and the associated significance levels; but until such a variable is found a belief in subjective randomization in some cases might seem well founded. In addition, from the discussion in Section 4 it should be clear that given such a variable these adjustments would have to be made even if the study were properly randomized, and any adjustment based on a model is somewhat dependent upon the appropriateness of the assumptions of the model whether the data are randomized or not. If the model for adjustment is appropriate, one can no longer object to the belief in subjective randomization because of the adjusted variable.

No doubt, given a fixed set of $2N$ trials one would rather be able to randomly assign the treatments and not rely on the concept of subjective randomization. However, if the choice were between an observational study whose $2N$ trials consisted of N representative E trials closely matched with N representative C trials and an experiment whose $2N$ trials were highly atypical, it is not clear which we should prefer; in practice there may be a trade-off between the reasonableness of the assumptions of subjective random sampling and subjective randomization (e.g., consider a carefully matched observational evaluation of existing compensatory reading programs and an experiment having these compensatory reading programs randomly assigned to inmates at a penitentiary).

The basic position of this paper can be summarized as follows: estimating the typical causal effect of one treatment vs. another is a difficult task unless we understand the actual process well enough (a) to assign most of the variability in Y to specific causes and (b) to ignore associated but

causally irrelevant variables. Short of such understanding, random sampling and randomization help in that all sensible estimates tend to estimate the correct quantity, but these procedures can never completely assure us that we are obtaining a good estimate of the treatment effect. Even assuming a good estimate there remains the problem of determining which aspects of the treatments are responsible for the effect.[13]

In addition, almost never do we have a random sample from the target population of trials and thus we must generally rely on the belief in subjective random sampling, i.e., there is no important variable that differs in the sample and the target population. Similarly, often the only data available are observational and we must rely on belief in subjective randomization, i.e., there is no important variable that differs in the E trials and C trials. If an important prior variable is found that systematically differs in E and C trials or the sample and target population, we are faced with either adjusting for it or not putting much faith in our estimate. However, we cannot adjust for any variable presented or any desired result can eventually be obtained.

In both experimental and observational studies, the investigator should think hard about variables besides the treatment that may causally affect Y and plan in advance how to control for the important prior variables--either by matching or adjustment or both. When presenting the results to the reader it is clearly important to indicate the extent to which the assumptions of subjective randomization and subjective random sampling can be believed and

---

[13] Consider for example "expectancy" effects in education (Rosenthal, 1971) and the associated problems of deciding the relative causal effects of the content of programs and the implementation of programs.

what methods of control have been employed.[14] If an observational study is carefully controlled, the investigator can often reach conclusions similar to those he would reach in the corresponding experiment. In fact, if the effect of the E vs. C treatment is large enough, he will be able to detect it in small, nonrepresentative samples and poorly controlled studies.

Basic problems in educational research are that causal models are not yet well formulated, and in many cases the effect of the E vs. C treatment under study appears to be quite small. Given this situation, it seems reasonable to search for treatments with large effects by the use of observational studies and rely on further study for more refined estimates of the effects of those treatments that appear to be important.

---

[14] Recent advice on the design and analysis of observational studies is given by W. G. Cochran in Bancroft (1972).

References

Bancroft, T. A. Statistical papers in honor of George W. Snedecor. Ames, Iowa: The Iowa State University Press, 1972.

Campbell, D. T., & Erlebacher, A. How regression artifacts in quasi-experimental evaluations can mistakenly make compensatory education look harmful. In J. Hellmuth, (Ed.), Compensatory education: A national debate. Vol. III of The disadvantaged child. New York: Brunner/Mazel, 1970.

Cochran, W. G. Sampling techniques. New York: John Wiley, 1963.

Cochran, W. G., & Cox, G. M. Experimental designs, 6th edition. New York: Wiley, 1957.

Fisher, R. A. The design of experiments, 6th edition. New York: Hafner, 1925.

Kempthorne, O. The design and analysis of experiments. New York: Wiley, 1952.

Lehman, E. L. Testing statistical hypotheses. New York: Wiley, 1959.

Rosenthal, R. Teacher expectation and pupil learning. In R. D. Strom (Ed.), Teachers and the learning process. Englewood Cliffs, New Jersey: Prentice-Hall, 1971. Pp. 33-60.

Savage, L. J. The foundations of statistics. New York: Wiley, 1954.

Scheffé, H. The analysis of variance. New York: Wiley, 1959.