Journal of Educational Psychology
1974, Vol. 66, No. 5, 688-701

ESTIMATING CAUSAL EFFECTS OF TREATMENTS IN RANDOMIZED AND NONRANDOMIZED STUDIES[1]

DONALD B. RUBIN[2]

Educational Testing Service, Princeton, New Jersey

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 that the use of carefully controlled nonrandomized data to estimate causal effects is a reasonable and necessary procedure in many cases.

Recent psychological and educational literature has included extensive criticism of the use of nonrandomized studies to estimate causal effects of treatments (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 effects. If taken as applying to all fields of study, this position is untenable. Since the extensive use of randomized experiments is limited to the last half century,[3] and in fact is *not* used in much scientific investigation today,[4] one is led to the conclusion that most scientific "truths" have been established without using randomized experiments. In addition, most of us successfully determine 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 (nonrandomized) and either (*a*) the cost of performing the equivalent randomized experiment to test all treatments is prohibitive (e.g., 100 reading programs 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 nonrandomized studies than to ignore these data and dream of the ideal experiment or make "armchair" decisions without the benefit of data analysis. Using the indications from nonrandomized studies, one can, if necessary, initiate randomized experiments for those treatments that require better estimates or that look most promising.

The position here is *not* that randomization is overused. On the contrary, given a choice between the data from a randomized experiment and an equivalent nonrandomized study, one should choose the data from the experiment, especially in the social sciences where much of the variability is often unassigned to particular causes. However, we will develop the position that nonrandomized studies as well as randomized

[1] I would like to thank E. J. Anastasio, A. E. Beaton, W. G. Cochran, K. M. Kazarow, and R. L. Linn for helpful comments on earlier versions of this paper. I would also like to thank the U. S. Office of Education for supporting work on this paper under contract OEC-0-71-3715.

[2] Requests for reprints should be sent to Donald B. Rubin, Division of Data Analysis Research, Educational Testing Service, Princeton, New Jersey 08540.

[3] Essentially since Fisher (1925).

[4] For example, in Davies (1954), a well-known textbook on experimental design in industrial work, randomization is not emphasized.

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 (E) treatment (e.g., a compensatory reading program) and the other half having been exposed to a control (C) treatment (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 nonrandomized study, a quasi-experiment, or an observational study. The objective is to determine for some population of units (e.g., underprivileged sixth-grade children) the "typical" causal effect of the E versus 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 versus C treatment on Y.

Defining the Causal Effect of the E versus C 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 initiated at t_1 and what would have happened at t_2 if the unit had been exposed to C initiated 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 versus C treatment will reflect this intuitive meaning.

First define a trial to be a unit 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 of these but rather of a third treatment, E + C).

Now define the causal effect of the E versus 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[5] 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 versus C treatment on Y for that trial, that is, for that particular unit and the times t_1, t_2.

For example, assume that the unit is a particular child, the experimental treatment is an enriched reading program, and the control treatment is a regular reading program. Suppose that if the child were given the enriched program initiated at time

[5] The measured value of Y stated with reference to time t_2 is considered 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 he must consider errors about some unmeasurable "true score." 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

t_1, 10 days later at time t_2 he would have a score of 38 items correct on a reading test; and suppose that if the child instead were given the regular program initiated at time t_1, at time t_2 he would score 34 items correct. Then the causal effect on the reading test for that trial (that child and times t_1, t_2) of the enriched program versus the regular program is $38 - 34 = 4$ more items correct.

The problem in measuring $y(E) - y(C)$ is 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 carryover effects (e.g., the effect of the first treatment wears off slowly) or general time trends (e.g., as the child ages, his learning ability increases), we cannot be certain that the unit's responses would be identical at both times.

Assume now that there are M trials for which we want the "typical" causal effect. For simplicity of exposition, 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 ($j = 1, 2, \cdots, M$); thus $y_j(E) - y_j(C)$ is the causal effect of the E versus C treatment for the j^{th} trial, that is, 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 versus C treatment for the M trials is the average (mean) causal effect for the M trials:

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

Even though other definitions of typical are interesting,[6] they lead to more complications when discussing properties of estimates under randomization. Hence we assume the average causal effect is 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.

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

For now assume that the 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 is discussed separately from the issue of estimating the typical causal effect for the trials under study. Also, for now we only consider the simple and standard estimate of the typical causal effect of E versus C: the average Y difference between those units who received E and those units who received C. After considering this estimate when there are only two trials in the study and then when there are $2N$ ($N > 1$) trials in the study, we will more formally discuss two benefits of randomization.

Two-Trial Study

Suppose there are two trials under study, one trial having a unit exposed to E and the other having a unit exposed to C. The typical causal effect for the two trials is

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

The estimate of this quantity from the study, the difference between the measured Y for the unit who received E and the measured Y for the unit who received C, is either

$$y_1(E) - y_2(C) \quad [2]$$

or

$$y_2(E) - y_1(C) \quad [3]$$

[6] 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 effects 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.

depending upon which unit was assigned E. Neither Equation 2 nor Equation 3 is necessarily close to Equation 1 or to the causal effect for either unit

$$y_1(E) - y_1(C) \quad [4]$$

or

$$y_2(E) - y_2(C), \quad [5]$$

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 in Equation 2 as that in Equation 3, so that the average or "expected" difference in Y between experimental and control units is the average of Equations 2 and 3,

$$\tfrac{1}{2}\,[y_1(E) - y_2(C)] + \tfrac{1}{2}\,[y_2(E) - y_1(C)]$$

which equals Equation 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 suppose 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 Equation 2 is about equal to Equation 3, and both are about equal to the desired typical causal effect in Equation 1. In fact, if the two units react identically in their trials, Equation 5 = Equation 4 = Equation 3 = Equation 2 = Equation 1, and randomization is absolutely irrevelant. 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. For example, when comparing the heat expansion rates (per hour) of a metal alloy in oxygen and nitrogen, an investigator might use 2 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, Equation 1.

A skeptical observer, however, could always claim that the experimental minus control difference is not a good estimate of the typical causal effect of the E versus 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 versus 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. Nevertheless, 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 versus 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 important causal variables have not been identified.

Another source of confidence that the experimental minus control difference is a good estimate of the causal effect of E versus 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.

The 2N Trial Study

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

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

Let S_E denote the set of indices of the E trials and S_C denote the set of indices of the C trials ($S_E \cup S_C = \{i = 1, 2, \cdots, 2N\}$). 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

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

where $\sum_{j \epsilon S_E}$ and $\sum_{j \epsilon S_C}$ 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). 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, and the two units react identically at the times of their trials; that is, the $2N$ trials are actually N perfectly matched pairs. We now show 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. 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.

In practice, of course, we never have exactly matched trials. However, if matched pairs of trials are very similar in the sense that prior to the initiation of treatments the investigator has controlled those variables that might appreciably affect Y, then $\bar{y}_d$ 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 (e.g., $2N$ children from the same school matched by sex and initial reading score into N pairs, with the same observed E − C difference in final score in each matched pair). Similarly, if the trials are not pair-matched but are all similar (e.g., all children are males from the same school with similar pretest scores) and if we observe that all $y_j(E)$ $j \epsilon S_E$ are about equal and all $y_j(C)$ $j \epsilon S_C$ 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 reading study, if more active children always received the enriched program, or in the heat-expansion study, if the first molded alloy was always 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 children or the first molded alloy. Hence, 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 child gets E or C). Of course, if we knew the skeptic's claim beforehand, 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 child, and in the other half assign C to the more active child, or include the child'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 child) and claim that $\bar{y}_d$ estimates the effect of this variable rather than τ, the causal effect of the E versus C treatment. Within the experiment there can be no refutation of this 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 it.

Two Formal Benefits of Randomization

If randomization can never assure us that we are correctly estimating the causal effect of E versus 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 that are important:

1. The average E − C difference is an "unbiased" estimate of τ, the typical causal effect for the $2N$ trials.
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 Sheffé (1959) and Kempthorne (1952).

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, Cochran & Cox, 1957), 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 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 randomly within matched pairs (the randomized blocks experiment, Cochran & Cox, 1957), any of the 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.

For each of the r possible allocations in the randomization set, there is a corresponding average E − C difference that would have been calculated had that allocation been chosen. If the expectation (i.e., average) of these r possible 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 ($j = 1, \cdots, 2N$) to the average E − C difference in half of the r allocations in the randomization set is $y_j(E)/N$ and in the other half is $-y_j(C)/N$. The expected contribution of the j^{th} trial to the average E − C difference is therefore

$$\tfrac{1}{2}[y_j(E)/N] + \tfrac{1}{2}[-y_j(C)/N].$$

Summing over all $2N$ trials we have, the expectation of the average E − C difference over the r allocations in the randomization set is

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

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 indicates 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 constraints on the symmetry of effects, the average E − C difference is *not* equally likely to be above τ as below it.

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 reading programs, initial reading score is not a matching variable, and after the experiment is complete we find that the average initial score for the children exposed to E was higher than for those exposed to C. Clearly we would now believe that $\bar{y}_d$ probably overestimates τ even if treatments were randomly assigned.

In sum, the unbiasedness of the E − C difference for τ follows from the random assignment of treatments; it is a desirable property because it indicates that "on the average" we tend to estimate the correct quantity, but it hardly solves the problem of estimating the typical causal effect. As yet we have no indication whether to believe $\bar{y}_d$ is close to τ nor to any ability to adjust for important information we may possess.

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

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

Having the $\tilde{\tau}_j$ and the observed $y_j(E)$, $j \epsilon S_E$ and $y_j(C)$, $j \epsilon S_C$, we can calculate hypothesized values, say $\tilde{y}_j(C)$ and $\tilde{y}_j(E)$, for all of the $2N$ trials. For $j \epsilon S_E$, $y_j(E)$ 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 $j \epsilon S_C$, $y_j(C)$ 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, we can calculate hypothesized $\tilde{y}_j(E)$ and $\tilde{y}_j(C)$ for all $2N$ trials, and 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.

Suppose that we calculate the 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 the observed difference is in the tail of distribution and therefore not typical of the r differences, we might doubt the correctness of the hypothesized $\tilde{\tau}_j$.

Since the average of the r E − C differences is the hypothesized typical causal

effect, $\tilde{\tau}$, and the r allocations are equally likely, we can make the following probabilistic statement:

> Under the hypothesis that the causal effects are given by the $\tilde{\tau}_j$, $j = 1, \cdots, 2N$, the probability that we would observe an average E − C difference that is 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 that are 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, we either must admit that the observed value is unusual in the sense that it is 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 versus C treatment has no effect on Y whatsoever (i.e., $\tilde{\tau}_j \equiv 0$). Other common hypotheses assume that the effect of the E versus C treatment on Y is a nonzero constant (i.e., $\tilde{\tau}_j \equiv \tilde{\tau}_0$) for all trials.[7]

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 also have measured.

Presenting the Results of an Experiment as Being of General Interest

Before presenting the results of an experiment as being relevant, an investigator should believe that he has measured the causal effect of the E versus C treatment and not the effect of some extraneous variable. Also, he should believe that the result is applicable to a population of trials besides the $2N$ in the experiment.

Considering Additional Variables

As indicated previously, 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 because 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 the study. At times the variables will be ones which cannot causally affect Y even though in the study they may be correlated with the observed values of Y. An investigator who refuses to consider any additional variables 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 versus C treatment but 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, suppose in the reading study, with programs randomly assigned, we found that the average E − C difference in final score was four items correct and that under the hypothesis of no effects the significance level was .01; also assume that initial score was not a matching variable and in fact the difference in initial score was also four items correct. 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 $\bar{y}_d = 4$ items 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 zero or negative values for τ.

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 which was measured before the initiation of

[7] 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 $\tilde{\tau}_0$ 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 Lehmann (1959, p. 59) for the proof.

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. 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 x_j/N and the other half as $-x_j/N$; hence, averaged over all allocations, the j^{th} prior score has no effect.[8] But this result holds for any set of prior scores x_j, $j = 1, \cdots, 2N$, 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 variable "pretest reading score" or perhaps "IQ" properly scaled makes sense but "height in millimeters" does not. Also, why not use the prior variable "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 prior variables in order to use their values in an intelligent manner. For example, if the final score typically would equal the initial score if there were no E − C treatment effect (as with the length of the alloys in the heat expansion experiment), the gain score is perfectly reasonable. In the physical sciences, more complex models representing generally accepted functional relationships are often used; however, in the social sciences there are rarely such accepted relationships upon which to rely. What then does the investigator do in order to adjust intelligently the final reading scores for the subjects' varying IQs, grade levels, socioeconomic status, and so on? Apparently, 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 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). Hence, forming an intelligent adjusted estimate may not be simple even in a randomized experiment.

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 equally likely allocations in the randomization set. However, 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 it 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 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

[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 treatment), we would have no assurance that the adjusted E − C difference is an unbiased estimate over the randomization set.

all causal effects are zero, $\tau_j \equiv 0$, 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. Such a search should be more difficult given that randomization was performed, but 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 is 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 additional variables 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 that do not causally affect Y 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 somewhat arbitrary model to adjust for those variables he feels are important. An example which demonstrates that it is not always simple to interpret significant results in a randomized experiment with many prior variables recorded is the recent controversy over the utility of oral-diabetic drugs.[9]

Generalizing Results to Other Trials

In order to believe that the results of an experiment are of practical 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 sixth-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 tell us something about future sixth graders who might be exposed to this 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 $\binom{M}{2N}$ ways of choosing the $2N$ trials is equally likely to be selected. If T 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 T 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 T: Because of the randomization and random sampling, each trial is equally likely to be an E trial as a C trial and thus contributes $y_j(E)/N$ to the E − C difference as often as it contributes $-y_j(C)/N$. It also follows that under a hypothesized set of causal effects, τ_j, $j = 1, \cdots, M$, 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 τ 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 essentially unlimited, for example, all future sixth-grade students), some additional probabilistic results follow. 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, but whether the adjustment actually adjusts for the

[9] See for example Schor (1971) and Cornfield (1971).

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

additional variables(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 *future* trials; we never have a random sample of trials from the future but at best a random sample from the present; in fact, experiments are usually 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 τ_j, $j = 1, \cdots, M$, is given by the usual conditional significance level.

Even though the trials in an experiment are often not very representative of the trials of interest, investigators do make and must be willing to make this assumption of subjective random sampling in order to believe their results are useful. When investigators carefully describe 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.[11]

Presenting the Results of an Nonrandomized Study as Being of General Interest

The same two issues previously discussed as arising when presenting the results of an experiment also arise when presenting the results of a nonrandomized study as being relevant. However, the first issue, the effect of variables not explicitly controlled, is usually more serious in nonrandomized than in randomized studies, while the second, the applicability of the results to a population of interest, is often more serious in randomized than in nonrandomized studies.

Effect of Variables Not Explicitly Controlled

In order to believe that $\bar{y}_d$ in a nonrandomized study is a good estimate of τ, the typical causal effect for the $2N$ trials in the study, we must believe that there are no extraneous variables that affect Y and systematically differ in the E and C groups; but we have to believe this even in a randomized experiment. The primary difference is that without randomization there is often a strong suspicion that there are such variables, while with randomization such suspicions are generally not as strong.

Consider a carefully controlled nonrandomized 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 was 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 you were shown the units without being told which received E, only half the time would you correctly guess 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 procedure is analogous to assuming subjective random sampling in order to make inferences about a target population. Until an obviously important variable is found that systematically differs in the E and C trials, the belief in subjective randomization is well founded.

Now consider a nonrandomized study in

[11] See Cochran (1963) on regression and ratio adjustments. These are appropriate whether the sample is actually random or not.

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

which an obviously important prior variable is found that systematically differs in the E and C trials. We must adjust the estimate $\bar{y}_d$ and the associated significance levels just as we would if the study were in fact a properly randomized experiment. An obvious way to adjust for such variables is to assume subjective randomization (i.e., the study was randomized and the observed difference on prior variables occurred "by chance"), and use the methods discussed in the previous section "Considering Additional Variables" appropriate for an experiment (i.e , gain scores, adjustment by a known function, covariance adjustment).

The main problem with this approach is that having found an important prior variable that systematically differs in the E and C trials, we might suspect that there are other such variables, while if the study were randomized we might not be as suspicious of finding these prior variables. Additionally, the various methods of adjustment that yield unbiased estimates given randomization have varying biases under different models without randomization. Even though an unbiased estimate is not (as we have seen) the total answer to estimating τ, it is more desirable than a badly biased estimate. Recent work on methods of reducing bias in nonrandomized studies is summarized in Cochran and Rubin (1974). Much work remains to be done, especially for many prior variables and nonlinear relations between these and Y.

In sum, with respect to variables not explicitly controlled, a randomized study leaves the investigator in a more comfortable position than does a nonrandomized study. Nevertheless, the following points remain true for both: (*a*) Any adjustment is somewhat dependent upon the appropriateness of the underlying model—if the model is appropriate the confounding effect of the prior variables is reduced or eliminated, while if the model is inappropriate a confounding effect remains. (*b*) We can never know that all causally relevant prior variables that systematically differ in the E and C trials have been controlled. (*c*) We must be prepared to ignore irrelevant prior variables even if they systematically differ in E and C trials, or else we can obtain any estimate desired by eventually finding the "right" irrelevant prior variables.

Generalizing Results to Other Trials

For almost any study to be of interest, the results must be generalizable to a population of trials. Typically, nonrandomized studies have more representative trials than experiments since these are often conducted in constrained environments. Thus, if the choice is between a nonrandomized study whose $2N$ trials consisted of N representative E trials closely matched to 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 nonrandomized evaluation of existing compensatory reading programs and an experiment having these compensatory reading programs randomly assigned to inmates at a penitentiary).

In a sense, all studies lie on a continuum from irrelevant to relevant with respect to answering a question. A poorly controlled nonrandomized study conducted on atypical trials is barely relevant, but a small randomized study with much missing data conducted on the same atypical trials is not much better. Similarly, a very well-controlled experiment conducted on a representative sample of trials is very relevant, and a very well-controlled nonrandomized study (e.g., E and C trials matched on all causally important variables, several control groups each with a potentially different bias) conducted on a representative sample of trials is almost as good. Typically, real-world studies fall somewhere in the middle of this continuum with nonrandomized studies having more representative trials than experiments but less control over prior variables.

Summary

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

causes and (*b*) 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.[13]

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, that is, 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 the belief in subjective randomization, that is, there is no important variable that differs in the E trials and C trials. With or without random sampling or randomization, if an important prior variable is found that systematically differs in E and C trials or in 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, because if we do, any estimate can be obtained.

In both randomized and nonrandomized 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 ones—either by matching or adjustment or both. When presenting the results to the reader, it is important to indicate the extent to which the assumptions of subjective randomization and subjective random sampling can be believed and what methods of control have been employed.[14] If a nonrandomized study is carefully controlled, the investigator can reach conclusions similar to those he would reach in a similar experiment.[15] In fact, if the effect of the E versus C treatment is large enough, he will be able to detect it in small, nonrepresentative samples and poorly controlled studies.

Basic problems in social science research are that causal models are not yet well formulated, there are many possible treatments, and in many cases the differential effects of treatments appear to be quite small. Given this situation, it seems reasonable to (*a*) search for treatments with large effects using well-controlled nonrandomized studies when experiments are impractical and (*b*) rely on further experimental study for more refined estimates of the effects of those treatments that appear to be important. The practical alternative to using nonrandomized studies in this way is evaluating many treatments by introspection rather than by data analysis.

[13] Even assuming a good estimate of the causal effect of E versus C, there remains the problem of determining which aspects of the treatments are responsible for the effect. 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.

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

[15] See, for example, the large scale evaluation of Salk vaccine described by Meier in Tanur et al (1972).

REFERENCES

Bancroft, T. A. *Statisical papers in honor of George W. Snedecor.* Ames: 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.), *The disadvantaged child.* Vol. 3. *Compensatory education: A national debate.* New York: Brunner/Mazel, 1970.

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

Cochran, W. G., & Cox, G. M. *Experimental designs.* (6th ed.) New York: Wiley, 1957.

Cochran, W. G., & Rubin, D. B. Controlling bias in observational studies: A review. Mahalanobis Memorial Volume *Sankhyā -A,* 1974, 1–30.

Cornfield, J. The University Group Diabetes Program. *Journal of the American Medical Association,* 1971, **217,** 1676–1687.

Davies, O. L. *Design and analysis of industrial experiments.* London, England: Oliver & Boyd, 1954.

Fisher, R. A. *Statistical methods for research workers.* (1st ed.) New York: Hafner, 1925.

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

Lehmann, 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, N.J.: Prentice-Hall, 1971.

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

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

Schor, S. The University Group Diabetes Program. *Journal of the American Medical Association*, 1971, **217**, 1671–1675.

Tanur, J. M., Mosteller, F., Kruskal, W. H., Link, R. F. Pieters, R. S., & Rising, G. R. *Statistics: A guide to the unknown*. San Francisco, Calif.: Holden Day, 1972.

(Received July 30, 1973)