The nonequivalent-control-group design is important because true experimental designs are frequently either infeasible or undesirable and other quasi-experimental designs have only quite limited applications. This design, however, has been disparaged as nearly useless because it depends upon statistical methods that may give biased results when applied to it. The design is too important to let lie in this condition. It is suggested that slight modifications, a few of which have been offered by others, might render it more reliable. One such modification, quite simple, widely applicable, and highly restorative of internal validity, is suggested here. The bias associated with the standard design is presented as resulting from basic violations of the assumptions of statistical methods. Both reduction of the bias and estimation of its extent are shown to be possible if the comparison group is selected at random from the relevant population and used by itself, rather than in conjunction with the experimental group, for the preliminary estimation of parameters. The modified design is shown to have various advantages relative to its unmodified form and even, at times, relative to true experimental designs. A secondary purpose of this article, supportive of the first, is to clarify the analysis of evaluation designs by conceptualizing the issues in terms of ordinary least-squares regression.

On Rescuing the Nonequivalent-Control-Group Design

The Random-Comparison-Group Approach

LAWRENCE B. MOHR
University of Michigan

The primary purpose of this article is to contribute to the rescue of the nonequivalent-control-group design by specifying a modification that avoids the most damaging analytical difficulties of the standard approach. In order to do so, a

---

AUTHOR'S NOTE: *A prior version of this article was delivered under the title, "On Rescuing the Nonequivalent-Control-Group Design," at meetings of the American Political Science Association, Washington, D.C., August 1980.*

SOCIOLOGICAL METHODS & RESEARCH, Vol. 11 No. 1, August 1982 53-80

basis for expressing those difficulties is required. A secondary purpose is therefore to cast the logic and functions of a group of prominent preexperimental and quasi-experimental designs in the terms of ordinary regression analysis, which would seem to be more general, flexible, and parsimonious for this purpose than the analysis of variance, covariance, and correlation frameworks that have commonly been employed (Porter, 1973; Campbell and Boruch, 1975; Kenny, 1975; Cronbach et al., 1977; Reichardt, 1979).

A leading assumption behind this article is that the nonequivalent-control-group design, or "Design 10," as it is commonly known (see Campbell and Stanley, 1963: 40, 47-50), has great importance for program evaluation. The field has been strongly urged in recent writings to use true experimental designs rather than quasi-experimental designs such as Design 10 (e.g., see Gilbert, Light, and Mosteller, 1975; Campbell and Boruch, 1975). There can be little quarrel with this position, or with the judgment set forth in these same writings that true experiments, with effort and determination, could be used much more often than they are. But that use must still be limited. There are times when even the grimmest determination and the most tireless effort will not yield the political and practical control that is necessary to conduct randomized experiments. In addition, experimental design has its own internal validity problems in some situations and is therefore not always the design of choice even when it is a practical possibility (Cook and Campbell, 1979: 341-371; Chung, 1979). Alternatives to experimental design are therefore a necessity.

Among the quasi-experimental designs, there are three basic variants with fairly broad potential applicability: time series designs, regression discontinuity designs, and the nonequivalent-control-group design (see Campbell and Stanley, 1963: 37-43, 47-50, 55-57, 61-64). The first, time series, is limited because, except for a certain number of economic and demographic variables, time series data are simply not available for the measures of interest in most program evaluations (see Campbell, 1976). The regression discontinuity format demands controlled assignment to treatment groups just as an experiment does; the times

when an experiment is infeasible or undesirable, but controlled assignment on some basis other than randomization is possible, are few indeed. This leaves Design 10 as the last hope on a large number of occasions for obtaining respectable inferences about treatment effects—inferences that enable some reasonable degree of confidence.

The standard nonequivalent-control-group design improves on the two most prominent *pre*experimental designs, which are even less adequate, by combining them. Figure 1a shows one of these two, the simple before/after design, where the flow of time is from left to right (Campbell and Stanley, 1963: 7-12). The basic difficulty with this design is the threat of "history," i.e., one does not know but that something other than T_1 occurring between time 1 and time 2 caused some or all of the observed difference between the before measure ("pretest") and the after measure ("posttest"). The other preexperimental design, shown in Figure 1b, is the static-group-comparison design (Campbell and Stanley, 1963: 8, 12-13), where the dashed line divides the treatment group from a nonrandomized comparison group. The crippling threat in this case is "selection," i.e., one does not know but that the two groups were different to start with, so that observed differences after exposure to T_1 are untrustworthy indicators of treatment effects. Figure 1c, then, shows the nonequivalent-control-group design, which combines measurement at two time points, as in Figure 1a, with measurement on two groups, as in Figure 1b.

History is now controlled for because whatever happened to one group in the critical interval presumably happened to the other group, so that differences between them are not due to extraneous causes. The threat of history is therefore reduced to the threat of divergent history (sometimes called "intrasession" history), i.e., an extraneous event might have occurred only to one group after the initial observation, but not to the other (this possibility pervades much of program evaluation; in principle, it threatens true experiments as well as Design 10).

The selection threat is controlled in Design 10 by virtue of the pretest. One group may have started out ahead of the other, but the pretest permits the comparison of improvements, or

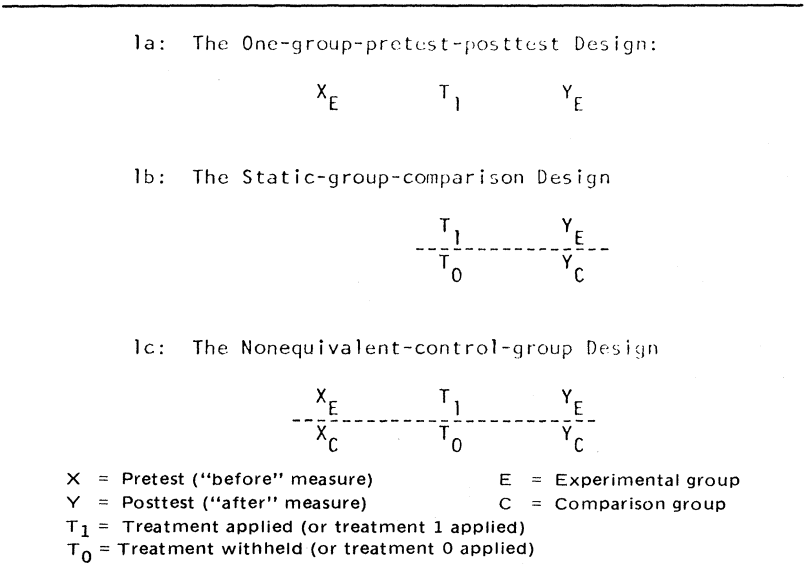


Figure 1: Two Preexperimental Designs and One Quasi-Experimental Design

"gains." Presumably, if the comparison group gained 3 points, or 4 points, or more, the experimental group should gain exactly the same amount and any difference in gain is due to the treatment. There is a problem with this reasoning, however, that has opened the door to relentless criticism of Design 10. The design, as a result, has lost much of its respectability.

The problem is that the inference about gain scores assumes that the slope of posttest on pretest is linear and is equal to 1.0. Let us continue to assume linearity for the moment, for the sake of simplicity, and consider the alternative later on. Under the assumption that the slope is equal to 1.0, however, the expected gain or loss for any individual, no matter what his or her pretest score, is numerically the same. It is given by the "natural" growth for the group, or the quantity α in the linear model $Y = \alpha + \beta_{YX}X$, where X is the pretest score and Y the predicted posttest. Thus, the experimental group gain is expected to equal the comparison group gain precisely, even if the two

groups started out at different average pretest levels. If β_{YX} were assumed to be 0.5, however, or 2.0, or anything other than 1.0, this would not be the case. The gain would then be some α plus half of one's pretest score, or twice that score, and so on. In that event, the expected numerical gain of the experimental group would not necessarily duplicate that of the comparison group but would depend, rather, on where the two started out.

Another way to express this (and both ways are important for the sequel) is to consider β_{YX} to be a factor that weights or adjusts the gap between the pretest scores of the two groups so as to transform it into the posttest gap under the assumption of no treatment effect (the null hypothesis), i.e.,

$$(\overline{Y}_E - \overline{Y}_C) = \beta_{YX}(\overline{X}_E - \overline{X}_C) \qquad [1]$$

where $\overline{Y}$ and $\overline{X}$ indicate mean scores and the subscripts refer to experimental and comparison groups. In essence, this equation gives the expected change in Y for a specific change in X, namely, the difference between $\overline{X}_C$ and $\overline{X}_E$. If β_{YX} is assumed to be equal to 1.0, then the gap between the groups on the posttest is expected to equal precisely the gap on the pretest. If β_{YX} is assumed to be another quantity, however, the expected posttest gap is greater or less. In short, if one must simply assume the magnitude of the slope of Y on X, there are infinitely many possibilities; the number 1.0 is arbitrary and no more reasonable, in principle, than many others.

The solution, obviously, is to estimate β_{YX} from the data rather than assume it arbitrarily. For this reason, analytic techniques for parameter estimation, such as the analysis of covariance and regression, have assumed great importance in connection with Design 10. If, however, the analytic techniques available happened to be essentially inadequate to the task, then so would be the design. Critics (e.g., Campbell and Boruch, 1975; Kenny, 1975) have indeed made the point that regression and covariance analysis are not adequate or suitable for this particular task. These critics are essentially correct. The design, however, is so important that the matter should not rest there.

Something should be done to improve the situation if possible. All of the treatments of this subject (e.g., Porter, 1973; Reichardt, 1979; Chung, 1979; Cronbach et al., 1977; Campbell and Boruch, 1975) indicate that, given the basic or classic nonequivalent-control-group design, employing even the most sophisticated statistical techniques can yield little in the way of improvement in confidence.

It is possible, however, that with some practicable modifications in the *design itself*, even the simple techniques such as straightforward regression and covariance analysis may be relied upon with much greater confidence. That strategy represents the spirit of the present article (as it was the spirit of Campbell and Stanley's original treatise on quasi-experimental design). A few such modifications have been suggested by Reichardt (1979: 194-196), Garfinkel and Gramlich (1973: 291), and especially Chung (1979: 131-161), but because they are complex and rarely practical they are unlikely to be commonly used. The alternative to be offered in the present article consists in selecting the comparison group at random instead of arbitrarily and in basing parameter estimation primarily upon that group. This design does not eliminate all bias, but it does go an appreciable distance toward that goal and, in addition, is both simple and quite broadly applicable. Furthermore, it is a modification that serves well to illustrate the heart of the problem represented by the application of statistical analysis to the standard nonequivalent-control-group design.

The critique of analytic methods in connection with Design 10 has rested primarily on two grounds. The first general problem is that incorrect treatment effects will be inferred when certain measures are unreliable, and such unreliability must occur with significant frequency in program evaluations. For example, the measurement of the pretest may be unreliable, the pretest and the posttest may have different reliabilities, the experimental and comparison groups may test with differing reliabilities, and the tests used may have floor effects or ceiling effects (see Campbell and Boruch, 1975: 223-241, 255-272). These criticisms are quite valid and for some situations, quite damaging. It is also true

that they do not present a problem in randomized or experimental designs as they do in Design 10. The unreliability problem, having its own extensive literature, is not treated in this article. In some evaluations, badly needed estimates of reliability will simply not be available. At other times, either good estimates will be available so that the appropriate correction factor may be applied, or reliability will not be a potential problem at all. In the sequel, it is assumed that one of the latter two conditions prevails.

The ground that subsumes most of the balance of the criticisms, although not generally stated in these terms, is that *the assumptions of the analytic techniques are not satisfied in Design 10 applications.* Essentially, what is violated is the assumption of zero correlation between independent variables and the disturbance terms in a regression framework (Cain, 1975, makes this point without demonstration).

THE ASSUMPTION THAT $r_{Xu} = 0$

For simplicity, we will work mainly with the underlying 3-variable model,

$$Y_i = \alpha + \beta_{YX \cdot T} X_i + \beta_{YT \cdot X} T_i + u_i \qquad [2]$$

for the i^{th} individual, where Y is the outcome measure of interest in the evaluation (the posttest), X is a measure of the same variable before the treatment or program begins (the pretest), and T is the treatment variable, a dummy variable scored 1 for individuals in the treatment or experimental group and 0 for those in the comparison or control group. Generalization to the case in which control variables in addition to X are included is straightforward.

The focus of the evaluation is the estimation of $\beta_{YT \cdot X}$, the coefficient of treatment effect. This coefficient has an important equivalent expression or definition in the context of equation 2 as the quantity $\overline{Y}_E - \overline{Y}_{OE}$. The latter term, $\overline{Y}_{OE}$, will be called the

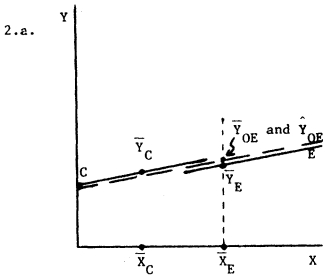
"null case" posttest mean for the experimental group, i.e., what the outcome would have been without the treatment, or presuming that the treatment had no effect whatever (when this value is estimated from the data, we will use the notation $\hat{Y}_{OE}$). The quantity $\overline{Y}_{OE}$ is of course never directly observable once the treatment has been applied. The coefficient $\beta_{YT \cdot X}$—the effect of the treatment—is thus simply the difference between the observed outcome, $\overline{Y}_E$, and the unobserved null case outcome, $\overline{Y}_{OE}$.

Let us begin with an assumption that has been worded as some variant of, "the within-group slope predicts between-group confounding" (Porter, 1973: 43; Porter and Chibucos, 1974: 444, 1975: 249). What this assumption actually means is that if the experimental and comparison group averages are different to start with, i.e., different on X, the pooled within-group slope of Y on X in the data accurately projects the quantity $\overline{Y}_{OE}$.[1]

The reader is referred to Figure 2. In Figure 2a, the two groups are shown separately rather than combined, and the dashed line represents the unobserved but "true" or "population" null case slope of Y on X. As Figure 2a is illustratively constructed, one sees that in this hypothetical case the assumption is satisfied, since the extension or projection of the comparison group slope intersects the vertical through $\overline{X}_E$ precisely at the point $\overline{Y}_{OE}$.[2] Furthermore, since that point coincides with $\overline{Y}_E$, one would infer that the program had no effect, and the inference would in this example be correct.

Figure 2b depicts the same sort of case, the only difference being to assume a true effect of the treatment; the whole regression line for the experimental group has been displaced upwards, to a higher Y intercept, because of the exposure of its members to the program being evaluated. This treatment effect would be captured accurately by the data analysis; it is seen graphically to be equal to the distance $\overline{Y}_E - \hat{Y}_{OE}$.

If, however, there were a bias in the pooled within-group slope, it would not accurately project the null case outcome for the experimental group. As is standard in regression theory, let us attribute such bias to a nonzero correlation between independent variable and disturbance, expressed notationally as $r_{Xu} \neq 0$. An example of this case is shown in Figure 2c (there, as



[The solid lines should be read as falling exactly upon (being collinear with) the dashed line.]

——— : observed slopes

-------- : unobserved true or population slope

E : Experimental Group

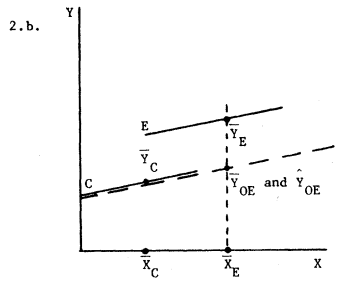
C : Comparison Group

X : Pretest

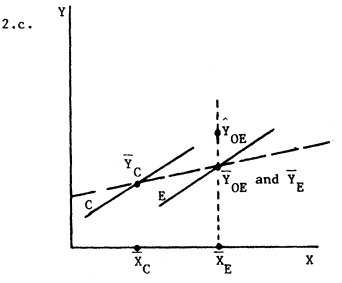
Y : Posttest

$\bar{Y}_{OE}$: True null-case experimental-group outcome

$\hat{Y}_{OE}$: Estimated null-case experimental-group outcome



(The comparison group line should be read as being collinear with the dashed line. Since $\bar{Y}_E \neq \bar{Y}_{OE}$, the figure represents a case in which there has been a true treatment effect.)



(Since $\bar{Y}_E = \bar{Y}_{OE}$, the figure represents a case in which there has been no treatment effect. Since $\hat{Y}_{OE}$ does not accurately estimate $\bar{Y}_{OE}$ in this hypothetical case, however, a treatment effect would be inferred.)

Figure 2

constructed, the correlation between X and u is positive in sign—within each group, the higher the pretest score the larger the algebraic magnitude of the disturbance).[3] In the figure, $\bar{Y}_{OE}$ would be erroneously estimated to be the projected point $\hat{Y}_{OE}$. In this way, bias in the within-group slope will clearly lead to an erroneous estimate of the effect of the treatment. Here, the bias makes a treatment that we assume by construction to have no effect, $\bar{Y}_E - \bar{Y}_{OE} = 0$, appear to have a negative effect—the

quantity $\overline{Y}_E - \hat{Y}_{OE}$. Standard regression analysis would correspondingly yield the same negative value for $b_{YT \cdot X}$ (note that the bias can have serious consequences even if small, depending upon how far apart the two group means are on the X axis). This phenomenon—a hidden phenomenon—occurs simply by happening to select groups for study in which the within-group slopes are misleading relative to the true null case regression of posttest on pretest. This kind of occurrence, whether referring to slopes or other values, will be called "sampling bias," as distinct from "sampling error." Since sampling bias results from arbitrary rather than random sampling, it is capable of grossly exceeding the range of ordinary sampling error. It is an extremely important concept in this context, since one way of expressing the primary problem with Design 10 is to say that, because of arbitrary selection, it inherently has the potential for an unknown degree of sampling bias.

A key conceptual step, then, is to consider the experimental and comparison groups in Design 10 to be *samples* from a population of interest. The problem is that they are arbitrary samples rather than random samples, which is why one cannot accept with any degree of confidence at all the closeness of statistics based upon them to the true parameters of interest. Given arbitrary rather than random sampling, there is no basis whatever for an expectation that $r_{Xu} = 0$. The existence, sign, and extent of sampling bias of this sort that may exist will simply be unknown.

A part of the suggested remedy is to *identify* the population of interest and *compose the comparison group by sampling randomly from that population.* In Design 10 applications, one usually has no control over the composition of the treatment group (the subjects may be volunteers, a natural group, a political jurisdiction, a group covered automatically, and so forth) and that is often why a true experiment cannot be conducted. The comparison group, however, is far more frequently a matter of choice by the investigator. There is no reason in most cases why the comparison group must also be arbitrary.

More needs to be said about determining the population from which to sample for the comparison group. This will be done

and examples will be provided as certain companion points are made. For the moment, however, let us assume that the "right" population can be identified and that it is some population that contains the experimental group.

Given that we compose the comparison group by sampling randomly from this population, the danger of bias from the source $r_{Xu} \neq 0$ is nullified if the estimate of β_{YX} is based on the random comparison group alone. This is the technique that we shall employ. It implies replacing equation 2 with a new, three-step procedure for calculating the treatment effect, as follows:

First, estimate α and β_{YX}, from the random-comparison-group data alone,[4] as the regression estimates a and b_{YX}:

$$Y_i = a + b_{YX}X_i + u_i \qquad [3]$$

Next, use these estimates and the *experimental* group pretest mean to estimate $\overline{Y}_{OE}$, as in note 2:

$$\hat{Y}_{OE} = a + b_{YX \cdot T}\overline{X}_E \qquad [4]$$

Note that all of the quantities on the right in equation 4 are known—a and $b_{YX \cdot T}$ from equation 3, since T in this context is inactive, and $\overline{X}_E$ from the observed data.[5]

Finally, a point estimate of the treatment effect in the random-comparison-group design, noted as b_T, is derived by simple subtraction:

$$b_T = \overline{Y}_E - \hat{Y}_{OE} \qquad [5]$$

where $\overline{Y}_E$ is taken from the data and $\hat{Y}_{OE}$ from equation 4.

Thus, the random-comparison-group approach eliminates bias in the slopes of the pretest and control variables as a source of error in estimating $\overline{Y}_{OE}$. However, that is not the only nor even the most important source of the problem. Preserving the estimating procedure specified in equations 3-5, we now consider the remaining source, still in the perspective of the assumptions for estimating the ordinary regression model.

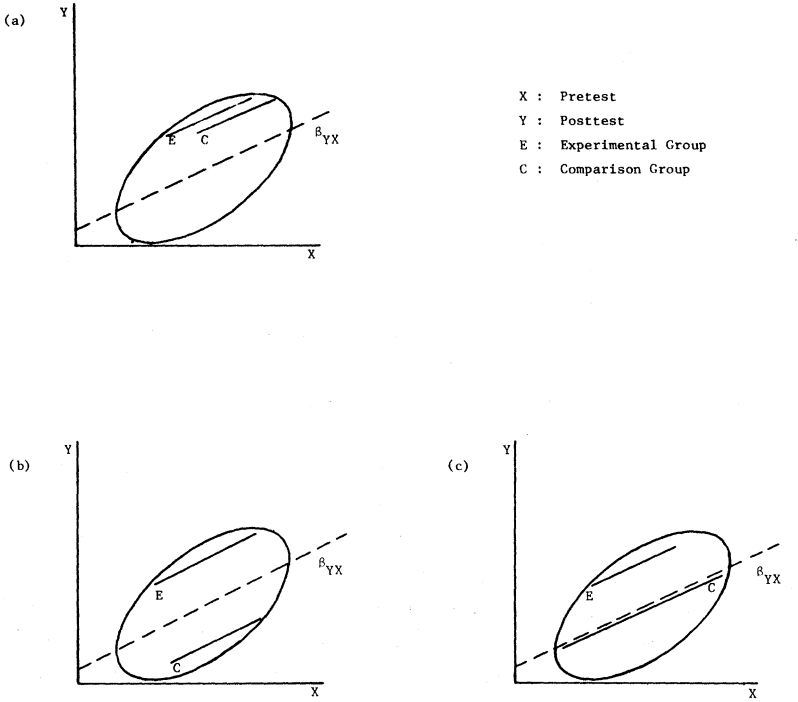
THE ASSUMPTION THAT $r_{Tu} = 0$

This is the assumption that the experimental and comparison group posttest means are given, within the constraints of ordinary sampling error, by expected values based on equation 2; neither group is characterized by disturbance terms that are systematically high or low. Another way of putting the same idea is to say that neither group begins, on average, either higher or lower than the other on unmeasured variables that affect the outcome.

With arbitrary selection, however, there is no basis for such an expectation. It holds legitimately only when the groups are established by randomization, as in a true experiment, or when both are taken as random samples from the same larger population. In Design 10, neither group is randomly selected. Thus, the essence of the remedy proposed by the random-comparison-group design is that some amount of random sampling may very constructively be introduced as a modification.

To explain, one expects statistically under this modification that the comparison group is representative of the population in unmeasured variables. That is, the intercept of the separate regression line for this group will be neither higher nor lower than it is supposed to be within the constraints of ordinary sampling error; the expected group mean would, in other words, fall on the dashed line in Figure 2. This assurance regarding the comparison group has the effect of *minimaxing* intercept bias in the estimation of treatment effect, that is, it minimizes the maximum possible error from the source $r_{Tu} \neq 0$. True, in the ordinary Design 10 approach, both groups, being arbitrary, could luckily be "off" due to sampling bias in the same direction, as in Figure 3a (where both have high average disturbances), and this would yield a very small bias. But they could also be off in opposite directions, as in Figure 3b, yielding a very large bias. In the random-comparison-group approach only one group can be unrepresentative because of arbitrariness, which essentially halves the maximum possible error from this source (Figure 3c).

Thus, using the random-comparison-group design, equations 3 and 4 would at least yield an unbiased estimate of $\overline{Y}_{OE}$ for



NOTE: The differences between the experimental and comparison groups could be due either to treatment effects or sampling bias. The text assumes that in these cases they are due to sampling bias.

Figure 3: Treatment Effect and Sampling Bias

a randomly selected experimental group. Unfortunately, equation 5 might still be biased from the source $r_{Tu} \neq 0$ because the experimental group is not random. It might well be unrepresentative of the population in its intercept: Because of certain variables that were not measured it may have the potential from the beginning for a higher or lower mean outcome than would have been expected from a random sample (see Figure 3c). To minimax the bias, in other words, is not to eliminate it. This irreducible potential for sampling bias[6] holds the design in the quasi-experimental category; it is still on less solid ground than a true experiment with respect to internal validity. However, it is significant that the only major source of bias remaining is

this intercept sampling bias in the experimental group.[7] Furthermore, even this problem is mitigated by the possibility of an *interval* estimate of program effect that makes allowances for the remaining sampling bias. Such an estimate is not available in ordinary Design 10 analysis.

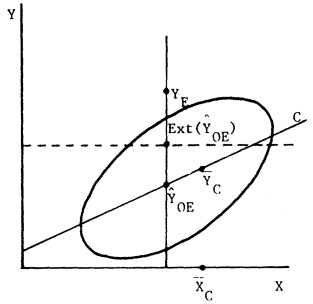
PROBABILITY-BASED ESTIMATES OF TREATMENT EFFECT

To speak of an interval estimate in the present context is to be in quite a different framework from classical inference, although the random comparison group does allow some elements of the latter to enter the procedure in critically important ways. A classical interval estimate depends conceptually upon the hypothetical outcomes of repeated sampling by the same random procedure. In the present context of possible sampling bias (as opposed to sampling error) in the experimental group, that basic idea is irrelevant; it is impossible to base parameter estimates on the outcomes of repeated sampling by some unknown arbitrary process. An ad hoc analytic procedure for arbitrary sampling in program evaluation must therefore be devised, taking advantage of the accompanying random comparison group. A suggested method is outlined in this section.

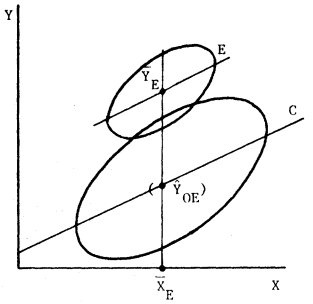
Figure 4a represents the illustrative outcome of an evaluation in which the treatment was applied to only one subject—a person, a city, a class in school, and so on. It appears from the plot that there has been a treatment effect, otherwise the outcome for that subject, Y_E, would likely have fallen within the ellipse that represents the error variance for the remainder of the population, of which that subject is a member. A point estimate of the treatment effect is given from equation 5 as

$$b_T = Y_E - \hat{Y}_{OE}$$

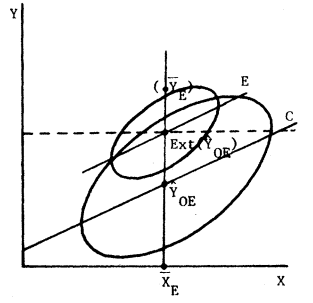
This of course assumes that there is no sampling bias—that the subject is precisely average in null case Y score for the pretest score X_E and the null case value is therefore assumed to lie



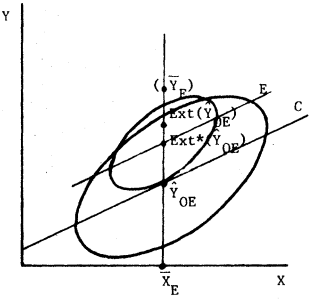
a. Treatment applied to one subject. Point estimate based on $\hat{Y}_{OE}$; interval estimate based on Ext($\hat{Y}_{OE}$).



b. Observed outcome; treatment applied to an experimental group.



c. Null case presumed for outcome in 4b when estimation is by Equation (7).



d. Null case assumed for outcome in 4b when estimation is by Equation (8).

E : Experimental Group
C : Comparison Group
X : Pretest
Y : Posttest
$\hat{Y}_{OE}$: Estimated null-case experimental-group outcome, with E treated as though it were a randomly chosen group having pretest score $\bar{X}_E$.

Ext ($\hat{Y}_{OE}$) : Extreme ($\hat{Y}_{OE}$). Conservatively estimated null-case experimental-group outcome.

Ext*($\hat{Y}_{OE}$) : Not-quite-so-conservatively estimated null-case experimental-group outcome.

Figure 4: Interval Estimates of Program Effect

precisely on the regression line. One can have little confidence in so strong an assumption. In many cases, however, one may have a great deal of confidence in the relaxed assumption that the true value Y_{OE} lies *somewhere* within the ellipse, in which case one can infer a range of treatment effects on a probability basis. The modified formula is based on the assignment of an "extreme" value for $\hat{Y}_{OE}$, on either the positive or negative side, to make the inference more conservative, as follows:

$$\text{Ext}(\hat{Y}_{OE}) = \hat{Y}_{OE} + Z_{\epsilon} s_{u_C} \pm e_{\hat{Y}} \qquad [6]$$

where Z is an appropriate positive or negative value from the standard normal table, s_{u_C} is the estimated standard error of the regression, based on the comparison group, and the final term, $e_{\hat{Y}}$, will be explained momentarily. If the value Z = 1.96 is used, for example, one is estimating "with 97.5% confidence" that the null case Y score would have been no more than 1.96 error standard deviations above the regression line. "Confidence" in this case reflects the fact that, given X_E, no more than 2.5% of the population has the propensity for more extreme Y scores than our subject is assumed to have without the treatment. The new, or "extreme" point, $\text{Ext}(\hat{Y}_{OE})$, is also shown in Figure 4a for comparison. In contrast to equation 5, the estimate of treatment effect "at the .975 level" then becomes

$$\tilde{\upsilon}_T \begin{cases} \geqslant \hat{Y}_E - \text{Ext}(Y_{OE}), \text{ positive program effects} \\ \leqslant \hat{Y}_E - \text{Ext}(Y_{OE}), \text{ negative program effects} \end{cases} \qquad [7]$$

One may use a higher confidence level, but at the expense of accepting a smaller program impact.

The final term in equation 6, $\hat{e}_Y$, is taken from classical inference and allows for sampling error in the coefficients a and b, equations 3 and 4. That is, it allows for sampling error in estimating population parameters from the random comparison group. If, as will frequently occur, the whole population rather

than a random sample is used, this term may be ignored. Otherwise, it is taken from the "prediction" function of classical inference. The issue in the present application is that of error in the "prediction" of a Y value corresponding to X_E, that is, the prediction of $\hat{Y}_{OE}$. In the present illustrative context with one independent variable, the proper expression for the confidence interval is (see Johnston, 1972: 154)

$$e_{\hat{Y}} = t_{\tilde{\epsilon}} s \sqrt{\frac{1}{n} + \frac{(X_E - \bar{X}_C)^2}{\Sigma(X_i - \bar{X}_C)^2}}$$

where ϵ need not be the same as ϵ in equation 6. In most applications, the above expression will yield a very small number, but increasing in magnitude with decreasing comparison group sample size and increasing distance of X_E from $\bar{X}_C$. The general extension to the multivariate case is given in Johnston (1972: 153).

In the more usual case, the experimental group will consist of many subjects rather than one, and the inference will pertain to means rather than single values. The observed outcome for this case is illustrated in Figure 4b.

One might treat this exactly as in the previous example, with the *mean* Y score employed as though the group were one subject. The null case outcome for the experimental group mean, $\text{Ext}(\hat{Y}_{OE})$, would again be given by equation 6, but we may rewrite equation 7 for the more general case by allowing the observed mean on Y to replace the observed single value:

$$b_T \begin{cases} \geqslant \bar{Y}_E - \text{Ext}(\hat{Y}_{OE}), \text{ positive program effects} \\ \leqslant \bar{Y}_E - \text{Ext}(\hat{Y}_{OE}), \text{ negative program effects} \end{cases} \quad [8]$$

Equation 8 is the basic estimating equation for the random-comparison-group design. In truth, however, this equation will

now in many instances represent an *overly* conservative assumption. If Z = 1.96 is used, for example, it assumes that the typical experimental group subject with the posttest score $\overline{Y}_E$ would in the null case have been 1.96 standard errors above the mean, $\hat{Y}_{OE}$, or just inside the border of the large ellipse in Figure 4b. All experimental group subjects with the same pretest score $\overline{X}_E$, but with outcomes *above* $\overline{Y}_E$, however, are assumed to be extreme outliers in the null case, who have practically no counterparts in the remainder of the population and so were not captured at all by the random sample comparison group. The null case, that is, is imagined to look like Figure 4c. Subjects in the small ellipse who are above the large one are assumed to be very extreme outliers. If there is indeed a great danger that the experimental group could be quite unusual in type (and therefore, that its members could be extreme outliers compared to the rest of the population, e.g., if the program were carried out on nearly all the large metropolitan areas in a state, leaving primarily the smaller cities for the comparison group), then perhaps the conservatism of equation 8 is truly warranted, or perhaps the random-comparison-group design cannot constructively be used at all.

In other cases, however, although one cannot assume that the experimental group is effectively a random sample, one can at least assume that nearly all of this group would have fallen somewhere within the boundaries of the large ellipse—that it does not contain many subjects so extreme as to have few or no counterparts in the remainder of the population. One may then employ a moderated extreme Y score, as follows:

$$\text{Ext}^*(\hat{Y}_{OE}) = \text{Ext}(\hat{Y}_{OE}) + Z_{\epsilon^*} s_{u_E} \qquad [9]$$

where $\text{Ext}(\hat{Y}_{OE})$ is calculated as in equation 6.

For example, in establishing $\text{Ext}(\hat{Y}_{OE})$ on the right-hand side, let us again choose Z_ϵ to be 1.96, representing the .975 level of confidence. One might reasonably and still fairly conservatively choose the new term Z_{ϵ^*} to be −1.28, representing the .90 level

in the *small* ellipse (note that S_{u_E} in the above equation refers to the experimental group—we are taking advantage of the experimental group variance), with the sign chosen as appropriate to move the assumed null case distribution closer to the population regression line, as in Figure 4d—in this case, negative. Indeed, one may quickly see by comparing Figures 4c and 4d that the effect of equation 9 is simply to move the experimental group ellipse for the null case assumption to a somewhat more central population location. The inference of treatment effect might take the verbal form: "Assuming that as many as 10% (from $Z_{\varepsilon^*} = -1.28$) of the experimental group may have been outliers—beyond the .025 level without the treatment—one can say with 97.5% confidence that the treatment effect was as large as b_T^*," where b_T^* is given as follows:

$$b_T^* \begin{cases} \geq \bar{Y}_E - \text{Ext}^*(\hat{Y}_{OE}), \text{ positive program effects} \\ \leq \bar{Y}_E - \text{Ext}^*(\hat{Y}_{OE}), \text{ negative program effects} \end{cases} \qquad [10]$$

If the experimental group variance is approximately as large as that of the population, so that $Z_{\varepsilon^*} S_{u_E}$ in equation 9 almost nullifies $Z_\varepsilon S_{u_C}$ in equation 6, one may then in many applications infer a treatment effect almost as large as that given by the simple and liberal point estimate in equation 5, depending upon the need to hedge against null case outliers.

IMPORTANCE OF THE PRETEST

It is clear from the above that the magnitude of S_{u_C}, or the height of the large ellipse in Figure 4b, is important for the power of the design, particularly with estimation as in equations 5-8. If $\bar{Y}_E$ fell on the inside of the ellipse, for example, one might not be able with confidence to infer any treatment effect at all. If in the same case one could shrink the ellipse, however, so that $\bar{Y}_E$ were now on the border or on the outside, its position might

safely be attributed to the treatment rather than sampling bias. This suggests that the addition of good independent variables to narrow the error variance would be desirable. Such variables—age, sex, and so forth, depending on context—are almost always available.

In this question of reducing the size of the error variance, the pretest is extremely important. It has a unique potential for accomplishing that objective. Generally speaking, although not universally, "before" is not only the best predictor of "after" by far, but is a very good predictor in absolute terms as well. In the case of an education evaluation discussed by Kenny (1975: 357), for example, the correlation between the posttest and the pretest alone was .75; in a similar case analyzed by Magidson (1977: 410), on the other hand, the multiple correlation of the posttest with several standard predictors, a pretest being unavailable, was only .33. The pretest, in short, is a critical component of any strategy to make the best use of the random-comparison-group design. It cannot function similarly in Design 10 because there the error variance in the comparison group may just as easily be the result of sampling bias as may any other statistic; it is fundamentally untrustworthy, whether large or small.

CURVILINEARITY AND INTERACTION

If the posttest is a curvilinear function of the pretest or other important independent variables, additional problems clearly arise in ordinary Design 10 analysis. If the experimental and comparison groups are fairly far apart on the pretest, for example, so that they occupy quite different portions of the curve, they may well manifest significantly different within-group linear slopes of Y on X, making further analysis impossible, or they may yield a pooled within-group slope that is incorrect and leads to incorrect inferences of the treatment effect. Conceptualizing the problem as misspecification due to omitted terms (e.g., X^k), the treatment variable T is almost certainly related to such terms when it is related to X, leading directly to bias from the source $r_{Tu} \neq 0$. The random comparison group is advantageous in this connection. In composing the comparison group by sampling

randomly from the relevant population, one obtains a broad, realistic range on all independent variables and may use those data well to discover and model any curvilinearity (including interaction terms) that might exist (see Reichardt, 1979: 155). The parameters of the model estimated in this fashion then augment or modify equations 3 and 4 and one proceeds from there as in the linear case. It stands to reason that analysis based on curvilinear modeling could profitably be carried out more frequently in evaluation studies (e.g., Garfinkel and Gramlich, 1973: 291-293), but the proper data base then becomes extremely important. Two arbitrarily selected subgroups do not in general form a good basis for modeling curvilinearity; one broadly based random sample of the right population, however, is optimal.

A weakness of Design 10 that is frequently noted is selection-X interaction. It will be referred to here as selection-T interaction to be consistent with the notation employed throughout. In essence, it refers to a treatment effect that occurs in the group selected only because of that group's particular values on certain unmeasured variables—an effect that would not occur in the comparison group or in other populations of interest. The first point to be made about selection-T interaction is that Design 10 does not in fact differ greatly from true experimental design in regard to this threat to validity (Campbell and Stanley, 1963: 50). In both cases, the effect produced is real enough, but its generalizability is in question. In both cases, the group selected may have characteristics that make it unusually sensitive to the treatment, only in the experimental design this statement will apply to the control group as well, whereas in Design 10 it may not. Since Design 10 treatment groups are more likely to be composed of volunteers, they are more susceptible, on the whole, to this sort of interaction (Campbell and Stanley, 1963: 50), but that differs with the selection method of the individual study.

It should also be noted that a selection-T interaction could produce a nongeneralizable slope of Y on X (for any X, i.e., for other control variables as well as the pretest). Generally, there is not great interest in this slope for its own sake, but there is a potential interaction problem with respect to its use in calculating the treatment effect. In the ordinary Design 10, if b_{YX} in the experimental and comparison groups are significantly different,

apparently connoting some sort of interaction, there is little choice but to terminate the analysis. Thus, one of the greatest dangers of selection-T interaction is that of wasting the study entirely because one cannot know which slope to believe.[8] This danger is eliminated by the random-comparison-group modification, since in that design only the slopes in the comparison group are utilized in the calculation of the effect of the program.

Equations 9 and 10 raise the issue of an interaction effect that is not a matter of concern in other designs. In "taking advantage of the variance" of the experimental group's posttest scores, there is the underlying assumption that the null case error variance for that group would have been the same as the observed variance, i.e., that there is no interaction of the treatment with unmeasured traits of the individuals *within* the experimental group such that the observed variance of Y is either increased or decreased, as a result, from what the null case variance would have been. The quantity s_{ue} estimates the null case; it helps to locate where the experimental group mean would have been without the treatment. If the observed posttest variance underestimates the null case variance, equations 9 and 10 underestimate the treatment effect; the group average without the treatment would actually have been "closer in" than we assume. If the null case variance is being overestimated, then so is the treatment effect. Thus, if a variance-enlarging interaction is suspected as a possibility, the more conservative estimate given by equation 8 is to be preferred.

SELECTION OF THE COMPARISON GROUP

Which population to use as the "relevant" population can be a difficult question to answer. Recognizing that there must often be some arbitrariness here, the general answer is: the population to which one is satisfied to generalize the results of the experiment. Arbitrariness enters because any experimental group belongs to infinitely many populations. For example, Ann Arbor, Michigan, belongs to the population of cities in Michigan, cities in the United States, cities of approximately 100,000 residents, university communities, and so forth. It is difficult to imagine a satisfactory set of rules that would enable

this decision to be made without resort to the judgment of the evaluator. It has seemed to be true that the choice almost always becomes narrowed to one or two possibilities by the nature of the program being evaluated and the scope on which it might truly be applied. For example, if Michigan were contemplating a public drunkenness law requiring the police to take offenders to a detoxification center instead of to jail, and if a pilot program were being run in Ann Arbor, one would probably want to generalize to cities in Michigan. On the other hand, if one were conducting a summer program for new minority students at the University of Michigan, one would prefer a population of large state universities to the population of colleges in Michigan. In the last analysis, however, all populations to which the experimental group belongs are correct, and the choice is a matter of evaluation strategy: Which population will be most useful and persuasive as a standard of comparison in this particular case?

The practicality of drawing a random sample and obtaining data is, of course, one constraint upon the identification of a population. More experience is necessary to provide guidance in this area, but consultation and student projects have suggested that there are two bases on which sampling would generally proceed. The first, and by far the most common, turns out to be organizational—identifying, enumerating, and sometimes obtaining data through individual (or sets of) organizations and agencies such as schools, school systems, hospitals, employment services, welfare rolls, chambers of commerce, lobbying associations, police departments, courts, counseling services, and the like, with their lists and records both of staff and clientele. The second base is the general population survey, whether national, local, or in-between. In these cases, one must generally obtain the consultation and other services of one of the large survey research organizations.

At the current writing, the random-comparison-group design is being applied in two contexts, which are noted here for illustrative purposes.

(1) One study will evaluate the impact of Quality Circles—a participative management technique well known for its successes in Japan—upon clerical productivity in a large federal agency.

Application of the treatment, which involves considerable expense, will be to twelve work groups. Their selection is based primarily upon convenience of location. Since the productivity data are fairly readily available, the comparison group is a large sample of similar work groups in the entire agency.

(2) The other study will evaluate a pilot housing-subsidy program carried out in eight locations. The pretest and posttest data are available, with a bit of manipulation, in an annual housing survey regularly applied to a large random sample of the United States, partially on a panel basis. One comparison group will be a random sample of the panel respondents in the survey (so that both pretest and posttest are available) who meet the eligibility requirements for the subsidies. Another comparison group will be at a higher level of aggregation—the Standard Metropolitan Statistical Areas defined by the Census Bureau—so that the experimental group size is the number of cities (eight) subsidized, rather than the number of individuals.

The usual approach in selecting a comparison group for Design 10 has been to try to obtain one that is as much like the experimental group as possible. One might accomplish this, for example, by "matching" the two groups on X, subject for subject. This makes excellent sense because when the mean pretest scores of the two groups are equal, the expected gap in posttest scores from equation 1—$\beta_{YX}(\overline{X}_E - \overline{X}_C)$—is zero no matter what the magnitude of β_{YX}. All concern for accuracy in estimating that parameter is thereby obviated. But that is *all* that is gained through equating the means. It should not be overlooked that intercept sampling bias is still a prominent threat to validity.

To minimax bias from that source, a *random* sample, but one whose mean equaled $\overline{X}_E$, would seem to be the comparison group of first choice. There are several reasons, however, why it may not be possible to have a random comparison group whose mean equaled $\overline{X}_E$. For one, the precise composition of the comparison group on this dimension may not be controlled by the experimenter: It might not be possible to give the pretest before selecting the subjects, or it might be necessary to select the treatment and control groups concurrently, or it might be too expensive to continue sampling, testing, and rejecting until the proper composition was attained. Moreover, most evaluation studies deal

with several posttests, not just one (e.g., Waldo and Chiricos, 1977, used eighteen measures of recidivism), and it is far too much to expect that the comparison group have a matching pretest for all of them.

A strategy that copes with most of these difficulties is to take a broad-range random sample and, for each analysis carried out, to use only some portion of the sample whose pretest mean matches the corresponding treatment group mean. This might occasionally be an appropriate and desirable course to follow, although how to select objectively a subgroup with a particular mean is far from clear. But another exigency severely limits the applicability of this strategy, as well. As discussed above, it will usually be desirable to use several independent variables in addition to the pretest in any one analysis. Thus, if the means were equal for a particular comparison subgroup on one such variable they would no doubt still be unequal on all of the others, so that little would have been gained by matching even as an addition to the random-comparison-group design. Most often, then, a broad-range random sample of the relevant population, which in its entirety provides the best estimate of all of the slopes involved, will be the method of choice for composing the comparison group.

SUMMARY OF THE FUNCTIONS OF THE RANDOM COMPARISON GROUP

The nonequivalent-control-group design is capable at times of avoiding many of the political, practical, and internal-validity problems that would bedevil a true experiment (Chung, 1979). Preserving these advantages, the following four points pull together the additional advantages of selecting the comparison group at random rather than arbitrarily:

(1) It eliminates bias from the source $r_{Xu} \neq 0$. It provides statistically sound estimates of all slopes in the model. It avoids the paralyzing possibility of accepting significantly different slopes in the two groups when that is due either to sampling bias or to selection—T interaction, as well as of arriving at inaccurate slope estimates by pooling the error-prone within-group statistics.

(2) It minimaxes the bias in the point estimation of β_T that results from a correlation between treatment and error ($r_{Tu} \neq 0$). It also then provides a basis for estimating the treatment effect in light of the bias that may remain.

(3) It provides a sound basis for incorporating curvilinearity and interaction into the analysis—and even into the design—when they are truly present and the proper variables have been measured.

(4) External validity, or generalizability, is greatly enhanced by random selection of the comparison group from the population to which generalization is desired. To the extent that one can assume the absence of selection-T interaction, the treatment effect can be generalized to any subgroup in that particular population of interest.

NOTES

1. This assumes that b_{YX} (experimental) and b_{YX} (comparison) are equal, or that they are close enough in value to be pooled statistically, yielding a single estimate of β_{YX}. If they are not, then when the groups have not been created by randomization, strict adherence to statistical procedure dictates that the analysis must stop; the evaluation data must be largely wasted. This problem is reviewed further, below.

2. In Figure 2a and in the text, the two within-group slopes are assumed for convenience to be equal to one another in order to avoid having to draw yet another line to represent the pooled within-group slope. This simplification allows us to speak in terms of extending the comparison-group slope rather than the pooled slope. To explicate the figure statistically, $\bar{Y}_{OE}$ may be estimated from the data as $\hat{Y}_{OE}$, the point on the pooled within-group regression line that is obtained by specifying $\bar{X}_E$ for X_i in the sample (as opposed to population) version of equation 2:

$$\hat{Y}_{OE} = a + b_{YX \cdot T}\bar{X}_E$$

Here, the term containing the dummy variable T may be suppressed because to assume the null case is to assume that T has no effect on Y. The observed value Y_E is shown in standard texts (e.g., Johnston, 1972: 59) to be obtained by specifying $\bar{X}_E$ for X_i in equation 2 again, but without the null case assumption:

$$\bar{Y}_E = a + b_{YX \cdot T}\bar{X}_E + b_{YT \cdot X}\bar{T}_E$$

Since T_i for the experimental group is always 1, then $\bar{T}_E$ is also equal to 1. The coefficient $b_{YT \cdot X}$ is thus seen, by subtracting the former equation from the latter, to be equal to $\bar{Y}_E - \hat{Y}_{OE}$. This is the observed version of the population identity mentioned just above in the text.

3. This case (Figure 2c) exactly portrays the problem of random measurement error in the pretest, except that in that event the correlation r_{Tu} is negative (see Reichardt, 1979: 160-164).

4. An observation frequently made at this point is that it seems a pity to sacrifice statistical efficiency by rejecting the experimental group as part of the basis for estimating

equation 3. It would seem that a Bayesian approach, for example, would be desirable. There may be no objection to such an approach at times, especially when experimental groups are drawn from the same population more than once, so that one can gradually home in on the parameters of interest, but it is probably best not to rely upon it in principle. Since sampling bias in the experimental group is precisely the problem to be gotten around, it is best not to bring that potential sampling bias itself into the estimating procedure. In compensation, one may strive to make the comparison group large enough so that coefficient variances are quite small (this can often be done at relatively small expense in program evaluation since that group gets no treatment). Unfortunately, it is not possible to use a mean-square-error criterion to estimate the tradeoff between bias and efficiency. Neither the means nor the variances of sampling distributions involving the experimental group can be employed in such a calculation because the sample is not a random one. With arbitrary samples, bias and sampling variance can be huge, but in any case they are unknown. In a subsequent section, however, it will be shown that the experimental group *variance* can be advantageously employed, even if its size cannot.

5. We continue to treat the case of a single control variable, the pretest, for simplicity. The extension to any number of additional control variables, X_k, is straightforward; one merely uses the individual-level comparison group data on these variables in the generalization of equation 3 and the experimental group means on those variables in the generalization of equation 4.

6. The term "sampling bias" as used here—in the context of means—is meant to convey something slightly different from "selection bias." The latter, as commonly employed (e.g., Campbell and Boruch, 1975: 227, 230), refers simply to the selection of experimental and comparison groups with different pretest means. As used here, "sampling bias" refers to outcome rather than selection and is the analog of "sampling error" for the case of nonrandom samples. It signifies the (nonrandom) selection of a sample—with any given means, $\overline{X}_1, \ldots, \overline{X}_k$, on the measured independent variables—whose null case mean on Y is different from that expected of a *random* sample having the same given means, $\overline{X}_1, \ldots, \overline{X}_k$.

7. Other threats to validity such as curvilinearity and interaction between the experimental group and the treatment are examined below.

8. One cannot escape this difficulty by using an "interaction model" in Design 10 analysis, i.e., by including a term in equation 2 in which the dummy T is multiplied by X:

$$Y_i = \alpha + \beta_{YX \cdot T(XT)} X_i + \beta_{ZYT \cdot X(XT)} T_i + \beta_{Y(YX) \cdot XT} X_i T_i + u_i$$

The problem remaining is that one cannot know in principle whether to ascribe the effects of the new term (a) to the treatment or (b) to sampling bais. For the former possibility, one would assume that the comparison group slope of Y on X is correct and that the treatment causes the experimental group slope to differ. This means that the treatment effect differs by pretest score, so that it is given for the i^{th} experimental group subject by the expression: $(\beta_{YT \cdot X(XT)} + \beta_{Y(XT) \cdot XT} X_i)$. For the latter of two possibilities, the sampling bias in the slope of Y on X may lie either in the comparison group or the experimental group or both, and the inference is thoroughly ambiguous. Even if one knew that the comparison group slope were correct, one would still not be able to determine the *extent* of sampling bias in the experimental group slope of Y on X. The root of the problem with respect to differences in slopes is that the treatment is indecipherably confounded with sampling bias in slopes. The problem does not arise in true experiments because of the irrelevance of sampling bias; there, equal slopes are expected if there is no treatment effect, the interaction model is appropriate, and the effect of the treatment is given by the expression in parentheses, above.

REFERENCES

CAIN, C. G. (1975) "Regression and selection models to improve nonexperimental comparisons," pp. 297-318 in C. A. Bennett and A. A. Lumsdaine (eds.) Evaluation and Experiment. New York: Academic.

CAMPBELL, D. T. (1976) "Focal local indicators for social program evaluation." Social Indicators Research 3: 237-256.

——— and R. F. BORUCH (1975) "Making the case for randomized assignment to treatments by considering the alternatives: six ways in which quasi-experimental evaluations in compensatory education tend to underestimate effects," pp. 195-296 in C. A. Bennett and A. A. Lumsdaine (eds.) Evaluation and Experiment. New York: Academic.

CAMPBELL, D. T. and J. C. STANLEY (1963) Experimental and Quasi-Experimental Designs for Research. Chicago: Rand McNally.

CHUNG, C-K. (1979) "The random design and the non-equivalent control group design in evaluation: a comparison." Ph.D. dissertation, University of Michigan.

COOK, T. D. and D. T. CAMPBELL (1979) Quasi-Experimentation: Design and Analysis Issues for Field Settings. Chicago: Rand McNally.

CRONBACH, L. J., D. R. ROGOSA, R. E. FLODEN, and G. G. PRICE (1977) Analysis of Covariance in Nonrandomized Experiments: Parameters Affecting Bias. Occasional Paper. Stanford, CA: Stanford Evaluation Consortium, School of Education, Stanford University.

GARFINKEL, I. and E. M. GRAMLICH (1973) "A statistical analysis of the OEO experiment in educational performance contracting." J. of Human Resources 8 (Summer): 275-305.

GILBERT, J. P., R. J. LIGHT, and F. MOSTELLER (1975) "Assessing social innovations: an empirical base for policy," pp. 39-194 in C. A. Bennett and A. A. Lumsdaine (eds.) Evaluation and Experiment. New York: Academic.

JOHNSTON, J. (1972) Econometric Methods. New York: McGraw-Hill.

KENNY, D. A. (1975) "A quasi-experimental approach to assessing treatment effects in the nonequivalent control group design." Psych. Bul. 82, 3: 345-362.

MAGIDSON, J. (1977) "Toward a causal model approach for adjusting for preexisting differences in the nonequivalent control group situation: a general alternative to ANCOVA." Evaluation Q. 1 (August): 399-420.

PORTER, C. (1973) Analysis Strategies for Some Common Evaluation Paradigms. Occasional Paper No. 21. East Lansing: Office of Research Consultation, School for Advanced Studies, College of Education, Michigan State University.

PORTER, A. C. and T. R. CHIBUCOS (1975) "Common problems of design and analysis in evaluative research." Soc. Methods & Research 3 (February): 235-257.

——— (1974) "Selecting analysis strategies," pp. 415-464 in G. Borich (ed.) Evaluating Educational Programs and Products. Englewood Cliffs, NJ: Educational Technology.

REICHARDT, C. S. (1979) "The statistical analysis of data from the nonequivalent control group design," pp. 147-206 in T. D. Cook and D. T. Campbell. Quasi-Experimentation: Design and Analysis Issues in Field Settings. Chicago: Rand McNally.

WALDO, G. P. and T. G. CHIRICOS (1977) "Work release and recidivism: an empirical evaluation of a social policy." Evaluation Q. 1 (February): 87-108.

Lawrence B. Mohr is Professor in the Department of Political Science and Research Scientist in the Institute of Public Policy Studies at The University of Michigan. His prior research has been in organizational behavior and research methods.