Conclusions of no difference are becoming increasingly important in evaluation research. We delineate three major uses of no-difference findings and analyze their meanings. (1) No-difference findings in randomized experiments can be interpreted as support for conclusions of the absence of a meaningful treatment effect, but only if the proper analytic methods are used. (2) Statistically based conclusions in quasi-experiments do not allow causal statements about the treatment impact but do provide a metric to judge the size of the resulting difference. (3) Using no-difference findings to conclude equivalence on control variables is inefficient and potentially misleading. The final section of the article presents alternative methods by which conclusions of no difference may be supported when applicable. These methods include the use of arbitrarily high alpha levels, interval estimation, and power analysis.

ANALYSIS OF NO-DIFFERENCE FINDINGS IN EVALUATION RESEARCH

GEORGE JULNES
LAWRENCE B. MOHR
University of Michigan

Evaluation research is increasingly concerned with conclusions that an intervention may not have had a meaningful impact, supported by statistical evidence suggesting that the null hypothesis may be true (e.g., Greenwald, 1975; Fagley, 1985; Cohen, 1987). This lack of meaningful impact is commonly referred to as "no difference"—a term best understood as meaning "no practical difference." Yeaton and Sechrest (1986, 1987) have assembled and integrated much of the literature on no-difference findings but point out that "inferential problems of no-difference results are not well understood" (1986: 838). We support these authors in their organization of the no-difference literature and seek to contribute to an understanding of the inferential problems. We believe that the relevant problems are made manageable when there is clarity as to the following.

AUTHORS' NOTE: *We gratefully acknowledge the comments of William A. Ericson, Charles Reichardt, Edward M. Gramlich, Timothy E. McDaniel, and Andrew M. Asher on an earlier draft.*

EVALUATION REVIEW, Vol. 13 No. 6, December 1989 628-655

1. The role of the no-difference findings: We delineate three prominent uses of no-difference findings and discuss the inferential problems peculiar to each use.
2. The type of error that is to be minimized: Whereas traditional conclusions of "difference" seek to minimize Type I errors, no-difference findings need to minimize Type II errors (i.e., failing to reject the hypothesis of no difference – the null hypothesis – when in fact it is false). While this distinction has been recognized for many years, it is often disregarded; moreover, the meaning of Type II errors varies in the context of the three uses of no-difference findings.

Underlying our position is our belief in the importance of the principle of conservatism in science – the belief that one should posit only those conclusions of which one is quite certain. This principle underlies the conventional use of the 5% level of significance instead of, say, the 30% level.

Our concern, however, is that this convention often misleads investigators in the direction of misusing nonsignificant results in such a way as to violate the principle of conservatism (see Cangelosi and Jesunathadas, 1986). In particular, we discuss several kinds of situations in which following traditional conventions would have the effect of lowering confidence in no-difference conclusions. The following analysis seeks to address the management of no-difference explorations, primarily within the context of this concern for the appropriate conservative stance. We begin, as noted, by delineating three common uses of no-difference findings. We then consider the implications of multiple comparisons and, finally, suggest alternative statistical procedures for maintaining conservatism in connection with no-difference conclusions. Our purpose is not to establish a definitive procedure (indeed, a combination of procedures may be useful), but to begin a dialogue and perhaps to stimulate a period in which several procedures are tried and their various attractions compared.

THREE USES OF NO-DIFFERENCE FINDINGS

Yeaton and Sechrest (1986) point out the many ways in which no-difference findings can support the validity of scientific conclusions. We suggest that decisions not to reject the null hypothesis (i.e., to accept a "no-difference" finding) are prominently used for the following purposes:

Case I. To permit the conclusion of no treatment effect in randomized experiments,

Case II. To permit the conclusion of no treatment effect in nonrandomized studies, and

Case III. To establish equivalence between groups: (1) on control variables in nonrandomized studies and (2) in the face of attrition in randomized experiments. Demonstrating equivalence is undertaken in order to make further analysis justifiable.

The purpose of delineating these three uses is that conclusions of no difference must be managed differently in each.

CASE I: NO-DIFFERENCE OUTCOMES IN RANDOMIZED EXPERIMENTS

For impact analyses (both experimental and quasi-experimental) a concern for no-difference findings occurs in such policy areas as the effect of limited versus radical mastectomy (Fisher et al., 1985), the efficacy of acupuncture (Gaw et al., 1975), the equivalence of control and experimental groups in a study of the treatment of angina (Gerstenblith et al., 1982), the effect of preventive police patrol on crime and public attitudes (Kelling et al., 1976), the effect of no-fault divorce laws on the divorce rate (Mazur-Hart and Berman, 1979), the effect of pretrial release without bail on appearance for trial (Botein, 1964-65), and the effect of desegregation on white flight (Pettigrew and Green, 1977). In each of these studies and a great many others, investigators have made a substantive or theoretical case for the absence of an effect.

In most randomized experiments, significance tests are used to determine whether the difference in outcome between experimental and control groups is sufficiently large to justify rejecting the possibility that it is due to chance. Such a conclusion is based on establishing an alpha level so small (e.g., 5% or 1%) that the risk of attributing efficacy to a treatment when the difference is due entirely to the randomization process or other chance effects is minimal. The smaller the alpha level, the more conservative the test (i.e., the lower the likelihood of Type I errors), and the more weight a significant result carries. Before proceeding, it is worth emphasizing the reason behind this conservative tradition in social science.

In some areas, such as manufacturing, it may be possible to specify exactly the relative costs in dollars (or some other common metric) of Type I and Type II errors. For example, it might be possible to estimate objectively the relative costs of remaking parts wrongly thought to be defective and of shipping defective parts (these two errors may be thought of as wrongly rejecting and accepting, respectively, the hypothesis that the parts are good). For such cases, it is appropriate to ignore statistical conventions, such as a probabilistic 5% decision rule, and establish a criterion that minimizes the expected cost, for example: "Remake the lot when the mean of a sample of parts deviates more than two-thirds of a standard error from the accepted norm, because the probability times the cost of remaking good parts when following that rule is less than the probability times the cost of shipping bad

ones." In the social sciences, however, the relative cost of Type I and Type II errors is more difficult to establish.

In the face of such uncertainty, we have statistical conventions that encourage us to withhold judgment when the data are not conclusive. In the traditional statistics, conventions direct us to withhold judgment when failing to achieve significance at, for example, the 5% or 1% alpha level. Mostly, in other words, we use conventions that consider Type I errors alone, despite the likelihood that they result in large risks of Type II error, because it is impossible or impractical to compare the probable costs of the two. Since a balance cannot be struck, the strategy elected is to choose the more serious type of error and minimize its probability. Thus we have a conventional, conservative bias against conclusions of difference or relationship.

Contrast this traditional use of significance testing with its use in conclusions of no difference, still in randomized experiments. As a hypothetical example, consider an evaluation that attempts to show that the traditional policies requiring newly released inmates to report to parole officers are unnecessary. Instead, it is proposed that a state program involving an honor system for newly released convicts (i.e., no requirement to report to a parole officer) will be just as effective in deterring future criminal activity. Suppose that 42 out of 100 randomly chosen subjects granted this honor system upon parole were later convicted of serious crimes, and that out of 100 randomly selected control subjects with traditional parole requirements only 32 were later convicted of serious crimes (a difference of proportions of 0.10). Given these data, one would conclude that the difference between the groups is not significant at the one-tailed 5% level (Z is dependent on the proportions involved; choosing 32% as a given, a difference of proportions of 0.12, rather than 0.10, would have been necessary for significance) and, thus, that the new parole program is just as safe as the old. But this failure to reach significance would in general not be reassuring: while the results may not be statistically significant, 42 is quite different from 32 and could signify an important community crime risk.

One might be tempted by convention to use a "more conservative" test, say the 1% level, but this tactic would backfire because then the new program would be considered safe unless 49 or more of the experimental subjects were convicted of serious crimes (a difference of proportions of 0.17!). Thus it appears that extending traditional strategies to the no-difference situation results in tests that are more liberal than the investigator intended. The problem is, as Blalock (1972) suggests, that the investigator is on "the wrong end of the hypothesis" (p. 161). In such cases, the investigator is attending

to Type I error when Type II error—failing to reject "no difference" when it should be rejected—is the real concern.

Rather, in such a case, state officials might reserve the term "safe" for a program that yielded 40 or fewer convictions as compared to 32 in the control group (a difference of proportions of 0.08, $p > 0.10$). Should one wish to be even more conservative, one might conclude that the program was safe only if there were fewer than 37 convictions (a difference of proportions of less than 0.05, $p > 0.25$).

The important point here is to notice the trend to increase confidence in all conclusions by restricting the range of significance to only the most compelling values: *rejection* of the null hypothesis is most compelling with extremely large differences—outlying scores that are significant at small alpha levels; "*acceptance*" of the null hypothesis—the no-difference conclusion—is most compelling with scores close to the null value, scores that are nonsignificant even at one-tailed alpha levels of 0.25 or even greater. Thus if we wish to be conservative in accepting no-difference findings, we need to do something comparable to the use of large alpha levels rather than small ones (we suggest three alternatives in the final section, below).

That a desire to be conservative sometimes requires small alpha levels and at other times large alpha levels results from our being concerned with different types of error. In the traditional case we wish to avoid calling an ineffective program effective (avoid Type I error—rejecting a null hypothesis when it is true); in the no-difference example we wish to avoid the conclusion that a treatment with appreciable impact had little or none (avoid Type II error—failing to reject a false no-difference hypothesis). This distinction between types of error is explained in any introductory statistics text; it is developed here because it appears to be misapplied in much of the no-difference literature: When a substantive or theoretical case that would be supported by no-difference findings has been made, chances are high that significance testing will then be applied by the investigator in the *ordinary* way, and therefore misapplied. In some cases (especially in Cases II and III, below) it is better to draw conclusions of no difference without inferential statistics of any sort (e.g., make judgments simply on the basis of the magnitudes of the differences), but there are alternative procedures, such as large alpha levels, that could legitimately be used more frequently to support relevant conclusions in no-difference situations. Unquestionably, these procedures apply in principle to Case I, randomized experiments.

In summary, it is necessary to be conservative in one's statements. Believing that science progresses best when only the most conclusive results are accepted as true, "*the researcher should lean over backwards to prove*

himself wrong or to obtain results that he actually does not want to obtain" (Blalock, 1972: 161, emphasis in the original). Thus when supporting a decision that one drug treatment is superior to another, for example, researchers should make it difficult to conclude on the basis of experimental evidence that there is a difference (small alpha level). When supporting a position that a particular treatment or policy is not superior, researchers should make it hard to conclude that the treatments are equivalent. If after making it difficult to conclude no difference the no-difference decision is still supported, then the conclusion carries some weight.

CASE II: NO-DIFFERENCE OUTCOMES IN QUASI-EXPERIMENTAL STUDIES

Most evaluations are not randomized experiments in design. When assignment to treatment groups is not by randomization, as in quasi-experiments and ex post facto studies (and, incidentally, in ordinary survey research in social science), the considerations regarding significance testing when it is important to hedge no-difference findings conservatively are not at all the same. This is not so much because of any special implications of no-difference results in quasi-experiments, but rather because of the general role of significance testing in that context. Basically, significance testing has little relevance for causal inference in such studies.

Mohr (1988: 90-96) explores the subject at length, including the presentation of certain qualifications of this basic conclusion (Mohr, 1988: 163-182). The essence of the matter is this: In impact analyses, a test of significance responds to the question: "What is the probability that results such as those observed could have been generated by random forces rather than a treatment effect?" The answer is based on the statistical model of sampling theory, which assumes assignment by a probability sampling procedure. In a randomized experiment, if one can rule out the effects of the randomization itself for at least some of the measured difference, relatively few reservations remain to an inference of at least some true experimental effect.

In the quasi-experimental case, however, chance plays a minor role. If a difference between the treatment and comparison groups is statistically significant, one may consequently be fairly certain that it is not due to chance. Such information, however, is rarely interesting. Since the groups were not assigned by a chance process, the test only rules out effects from such sources as random measurement error and events in the world that happened to occur to the subjects at random after their assignment to treatments in the study. The major worry by far in assessing the treatment effect, however, is that

there has been selection bias in the original assignment (or, in the case of before-after designs, effects from extraneous events occurring at about the same time as the treatment). These are nonrandom effects. Selection bias replaces randomization vagaries as the chief alternative to treatment effects in quasi-experiments, and this worry is not addressed in any way by the test of significance.

A nonsignificant result suggests that, had the two groups been equivalent before treatment, the measured outcome difference is not too large to have occurred as the result of certain random forces (measurement error, recent random events, etc.). One does not know, however, how nearly equal the two groups were to start on variables that matter for the outcome; one possibility is that the two groups were quite different and that the treatment has evened them out some, that is, it has had a substantial impact. *Although the emergent difference may be small, inference of little or no treatment effect is then inappropriate.*

One might try to overcome this potential bias by employing a pretest and other control variables, but the same difficulties remain in principle. An evaluation of the impact of a job training program might measure the before and after incomes of those who voluntarily participated in the program and those who voluntarily declined participation. Even if the pretreatment income of the two groups were exactly equal or statistically controlled (see Case III, below), one cannot conclude that the two groups were equal on all other meaningful—but unmeasured—variables, such as the motivation to work (see LaLonde and Maynard, 1987).

Thus when the design does not involve random assignment, defending a conclusion of no effect on the basis of a nonsignificant result requires defending the assumption that the treatment and comparison groups differed *only* by chance in their potential outcome scores before the treatment was administered, that is, that the assignment process was functionally equivalent to a randomization procedure. Defending this assumption is the raison d'etre behind some of the more sophisticated evaluation designs. The question becomes how effective these designs are in avoiding selection bias or some other distortion. Some of the more sophisticated nonrandomized designs appear reasonably adequate (e.g., the regression-discontinuity and random-comparison-group designs; see Mohr, 1988), others less so. The ultimate test is comparison of inferences from nonrandomized studies with those of parallel randomized experiments, using the same treatment subjects in both cases. Unfortunately, early comparison studies seem to indicate that results from the commonly used nonrandomized designs may easily be misleading, even when a pretest is used and very sophisticated analysis procedures are

later applied (e.g., Deniston and Rosenstock, 1973, or the series of studies described in Ashenfelter, 1986, Fraker and Maynard, 1987, and LaLonde and Maynard, 1987).

The foregoing analysis suggests that a no-difference finding based on statistical inference in a quasi-experiment would ordinarily have little to do with causality. However, the statistical model can serve to provide a metric for judging whether a difference is large or small (Blalock, 1960: 270-271; Mohr, 1988: 93-96). The metric is in terms of the probable outcomes of a hypothetical randomization process, as demonstrated by the significance test. If a positive difference, for example, is so small that even randomization would be expected to produce a larger difference, say, 25% of the time, as indicated by a nonsignificant result at the one-tailed 25% alpha level, then it may indeed be declared "small." There is no harm in this as long as it remains firmly understood that what is under consideration is the *difference* and not the *treatment effect.* The difference may indeed be small (in this metric or any other), but the treatment effect may still have been large in either the beneficial or harmful direction—offset, as we have noted, by selection bias in the quasi-experiment.

CASE III: NO DIFFERENCE ON CONTROL VARIABLES

Perhaps the most common use of no-difference conclusions involves attempts to show that two groups are equivalent in terms of measured control variables (examples are too common to cite). This use does not address the question of treatment effect directly but is meant to support such an analysis: (a) in the case of justifying a comparison group in a quasi-experiment, whose initial similarity to the treatment group is always critical if a causal inference is eventually to be made, and (b) in the case of attrition from the experimental or control group in a randomized experiment. In the former case, no-difference conclusions are used to suggest that the treatment and comparison groups are so similar on measures believed related to the outcome that they can be viewed as equivalent except for the treatment (and all unmeasured variables). In the latter case, the point is to see whether the attrition has compromised the result of an earlier randomization procedure: if there is no difference on certain measured characteristics between those who left and those who remained, then the nature of the group has not been changed by the attrition, at least not on those measured variables. Otherwise, the basis for invoking the statistical model for hypothesis testing, which assumes random sampling, has apparently been undermined. Note with respect to attrition that if the design is quasi-experimental rather than experimental, the similarity of the

treatment and comparison groups is already uncertain; it is ambiguous whether attrition, if it occurred, has made matters better or worse. There is therefore little point in testing for the similarity between those who left and those who remained. Nevertheless, if the groups are assumed to have differed at the start only by chance, these comments would apply to the quasi-experimental case as well.

In both cases (a) and (b), the conservative logic demands that one use a test that makes it difficult to conclude no difference, something that a small alpha level (5% or 1%) generally does not do.

But furthermore, once the decision is made that the groups are "equivalent" on the basis of a no-difference finding, the information from the control variables is typically discarded. This disregard of valuable information is unnecessary and unfortunate. If there are even slight differences in control variables and these control variables are related to the outcome of interest, it is best to use this information about differences in the analysis. That is, rather than reduce the information about control variables to a dichotomy (difference versus no difference), it is better by far to use the actual values on the control variables in the analysis—for example, as independent variables in a multiple regression equation. Therefore, the significance test for no difference need not be performed at all.

The advantage of this approach is that the added information can give a more accurate picture of the treatment effect. For example, Smith (1976) reported results in which a treatment group was worse off on a pretest than the comparison group, but the difference was not significant. Although the difference appeared to be substantial, failure to reach significance at the 5% level was interpreted as indicating that the two groups were equivalent, an inappropriate no-difference conclusion. Furthermore, following treatment the treatment subjects were *better* off than the comparison subjects. Because the outcome difference was not significant, however, the results were interpreted as showing that the treatment had no effect. Had the information on the pretest been used in the analysis rather than discarded it would have indicated that the treatment group had moved from well below the comparison group to well above it—a difference that could signify a very substantial impact. (Note: Because the study was a quasi-experiment, it does not add force to the above observation to say that the result might well have been statistically significant: significance testing in Case II—impact analysis in quasi-experiments—is generally of dubious value.)

In sum, Case III represents an inefficient use of the no-difference inference. Just for information, one might examine the differences and even test them for significance (but not at the 5% level) to get a metric for "small" and

"not small," as in Case II. But it is both wasteful and inappropriate to discard the information at that point on the basis of a "no-difference" conclusion.

THE RELEVANCE OF MULTIPLE COMPARISONS FOR NO-DIFFERENCE FINDINGS

In the foregoing analysis, we have tried to show that conservatism in no-difference conclusions requires minimizing the risk of accepting a false null hypothesis (minimizing Type II error) and that this is accomplished, for example, by moving in the direction of higher alpha levels. We now consider the impact of multiple comparisons on the conservative stance in relation to no-difference findings. Multiple comparisons with no-difference conclusions can be divided into two groups: (a) those in which a single no-difference finding (or a very few) will lead to some action, and (b) those in which some action is taken only if all (or almost all) of the multiple comparisons produce no-difference conclusions.

In the first type, consider an experiment with the honor parole system in which there are multiple subpopulations of parolees (e.g., first offenders versus multiple offenders; nonviolent versus violent crimes). If one concludes no difference for a particular subpopulation, then the honor system will be used for that type of parolee. If there were 10 subpopulations in the experiment (e.g., 200 parolees in each subpopulation randomly divided into honor and traditional parole programs), one faces a greatly increased likelihood that one of the subpopulations will yield a nonsignificant difference just due to chance (principally, to "unhappy" randomization). The proper response is to make it even harder to conclude no difference for each subpopulation. In the framework of using the alpha level as the criterion, this would involve *raising* alpha (e.g., from a conservative level of 0.25, one-tailed, to an even more conservative level of 0.30).

Another example of this type involves comparing one treatment with several others: An inexpensive drug is compared to a variety of different surgical techniques. If the drug is "equivalent" to any of the techniques, it will become standard policy to use the drug whenever that particular technique was formerly indicated. Under such conditions, an experiment comparing the drug to 10 surgical techniques runs too great a risk of falsely concluding no difference unless one compensates by making it *harder* to conclude no difference for each comparison. Again, if one were using an alpha level as the criterion, becoming more conservative would dictate

raising the critical alpha level, thereby requiring values closer to the null value.

The second type of multiple comparison decision—needing to find no difference on all comparisons—is exemplified by an experiment on the honor parole system in which there is only one population examined but multiple measures of recidivism. Thus the eligible parolees are randomly assigned to the honor or traditional system and their subsequent behavior is evaluated in terms of 10 recidivism measures (e.g., number of convictions, time before first conviction, and seriousness of crimes): the program is considered a success only if all measures yield no-difference results. A corresponding drug study would conclude that a new drug was as effective as traditional surgery for a single disease only if it yielded no-difference findings for all 10 symptoms monitored.

Perhaps the most common use of this second type of multiple comparison decision, however, occurs when researchers attempt to conclude that two groups in a quasi-experiment are equivalent by virtue of no-difference findings on control variables (as in the case of attrition, or of differences between treatment groups). In the typical case, the groups are not declared equivalent unless all of the comparisons are nonsignificant. We have argued that groups should not be declared equivalent or nonequivalent on the basis of a control-variable comparison. The case is no different when there are multiple comparisons, as there usually are. Instead, the comparison factors should be used as control variables in a multivariate analysis. Still, the logic of multiple no-difference comparisons in such cases is instructive, and we will touch on it briefly, below.

First, it is important to recognize that using multiple comparisons to support a global conclusion of no difference results in a paradox. On the one hand, as is commonly understood, the probability of at least one Type I error is increased; that is, there is an increased likelihood that at least one comparison will be significant merely due to chance. On the other hand, and this is less commonly noted, the probability of at least one Type II error is also increased, so that the more independent comparisons there are in any set of nonsignificant differences, the more likely it is that at least one of them is nonsignificant by chance. Thus multiple comparisons simultaneously: (1) decrease the likelihood of a global conclusion of no difference, and (2) decrease confidence in such global conclusions when they are made.

Yeaton and Sechrest (1986) suggest that when one wishes to conclude no difference for multiple comparisons, it is appropriate to *lower* the alpha level for each comparison (see Davis and Gaito, 1984, for a general treatment of this issue.)

> From a different perspective, since many group characteristics are typically tested for initial differences (such as age, sex, aspects of medical history), it would be expected that a few spurious differences would be produced by chance alone. *If* such subgroups have adequate sample size and statistical power, then it is legitimate to cite the problem of experiment-wise-error rate and to opt for such standard approaches as lowering the alpha level of each comparison [Yeaton and Sechrest, 1986: 842, emphasis in original.].

There is an important sense in which Yeaton and Sechrest are correct: at any given level of significance, the more comparisons, the more likely that one of the comparisons will be statistically significant just by chance (Type I error). Such a chance difference would prohibit further analysis; thus, it is natural to lower alpha to make it more difficult to conclude that the groups are different.

But this advice merely highlights the tension between Type I and Type II errors. While there are ways to decrease each of these, (see Cohen, 1982), the two are inversely related: all else being equal, lowering alpha necessarily decreases the likelihood of Type I error and increases the likelihood of Type II error. Thus not only do multiple comparisons per se increase the probability of making at least one Type II error, but lowering alpha as a strategy constitutes a second, compounding mode of having the very same effect.

As discussed in Case I, if we knew the relative costs of Type I and Type II errors, we could dispense with statistical conventions. It is the lack of this objective knowledge that inspires conventions to manage Type I error in order to maintain confidence in traditional conclusions of difference. With conclusions of no difference, we need parallel conventions to manage Type II error rates, something that lowering alpha does not do.

Consider the following scenario: The honor versus traditional parolee study is conducted, with random assignment and with multiple independent measures of recidivism. Wanting to be rigorous about the comparability of the two treatments, the researchers decide to assert equivalence only if differences on all types of recidivism are small, $p > 0.25$ for each measure considered independently. But the first outcome variable examined, "seriousness of crimes leading to new convictions," reveals higher average seriousness for the honor parolees (significant, $p < 0.21$). Fortunately, the groups were compared also on number of arrests, number of convictions, and time to first conviction, all nonsignificant ($p > 0.25$). Assume now that the researchers, following standard advice for multiple comparisons, had lowered alpha from 0.25 to 0.20 before the analysis. Since differences on *all* outcomes would then be nonsignificant, they would conclude that the treatments were equivalent.

Our intuition tells us that this "standard approach of lowering alpha" increases the risk of *wrongly concluding no difference* (Type II error); intuition would be even clearer if all five conviction comparisons had yielded $p < 0.21$ but were considered innocuous due to the lower alpha level. Lowering alpha should be proposed only if one believes that multiple comparisons make the overall conclusion of equivalence more conservative than originally desired. But in the no-difference situation, assuming that the several comparisons are independent, this belief would not be true: If two groups proposed to be equivalent are compared in terms of age, race, religion, and family background, lowering alpha means that larger differences in age, for example, will now be tolerated as nonsignificant. Why should this be so? If one decides on the basis of one comparison that a one-year age difference between groups is not negligible, multiple comparisons should not lead one to consider a two-year difference to be negligible.

In terms of the parole example, finding that the honor parolees are "equivalent" to traditional parolees in terms of number of arrests and time to first conviction does not compensate for possible differences in the seriousness of crimes committed. The effect of lowering alpha will be that more real differences in recidivism are overlooked; this potential increase in Type II error makes it more likely that an unsafe parole program will be pronounced "safe" and, hence, unwisely continued.

Thus the solution of lowering alpha decreases the risk of Type I errors—a definite problem when it is felt that all comparisons must be nonsignificant—but at the cost of an inevitable increase in the risk of Type II errors. When making multiple comparisons, it is necessary to decide which error type is more important. For conservatism in connection with the typical no-difference conclusion, it is Type II errors that must be given priority. In concluding no difference, researchers should be most concerned with the probability of falsely *accepting* the null hypothesis. With this orientation, the researcher must *manage beta* (the probability of Type II error), allowing alpha to become *larger* when necessary. As noted, this merely mirrors the more common practice: one always allows beta to become larger when moving to lower alpha levels in the traditional multiple comparison situation.

Because multiple independent comparisons do not inherently make a global conclusion of no difference more conservative, we recommend against lowering alpha as a general solution. Our perspective on multiple comparisons is developed further in the context of the statistical procedures described in the following section.

METHODS TO SUPPORT NO-DIFFERENCE FINDINGS

The foregoing discussion indicates why it is generally inappropriate to conclude no difference upon failure to reject the null hypothesis at the 5% level—such a test is too liberal. Accepting Yeaton and Sechrest's (1986) belief in the importance of no-difference conclusions, as we do, a concerted effort is required to develop alternative methods that maintain the conservative stance. Rather than endorse one particular method, we present three alternative approaches.

1. *Test the null hypothesis with a large alpha level (e.g., 0.25, one-tailed).*

This option is probably the simplest solution. We have used the one-tailed 25% level in this article, but it is arbitrary.[1] In the example of the honor parole system as originally developed above, the program would be concluded to be as safe as the traditional system if less than 37 honor inmates were later convicted (difference of proportions of less than 0.05). Note that such a procedure need not require that every program of this type with an effect greater than a difference of proportions of 0.05 be categorized as "unsafe." One could continue to reserve the conclusion "unsafe" for outcomes significant at the 5% level—in this case a difference of proportions greater than 0.12. If an effect greater than 0.05 but less than 0.12 were obtained, no judgment, safe or unsafe, would have received particular support (see Figure 1).

The resulting three zones, "no difference," "no judgment," and "difference," can help to clarify debates as to whether or not an intervention has had an impact. For example, Coleman et al. (1975a, 1975b) suggest that school desegregation led to "white flight." Assuming that a causal inference is pertinent in this ex post facto context, their conclusions should be tested with a small alpha level (e.g., 0.05 or 0.01). Pettigrew and Green (1977) suggest that desegregation has not caused white flight; their data should be analyzed with large alpha levels (e.g., 0.25 or higher). Both sides of the debate should agree that intermediate results (e.g., between 0.25 and 0.05) support neither position.

The high alpha-level solution leads to some important insights, but it shares with traditional 5% testing the problem of being defined in terms of alpha and thus Type I errors. While testing at the 25% level is more conservative than testing at the 5% level, neither test provides any estimate of the quantity that is statistically critical in the no-difference circumstance, the probability of Type II errors. In some cases the probability of Type II

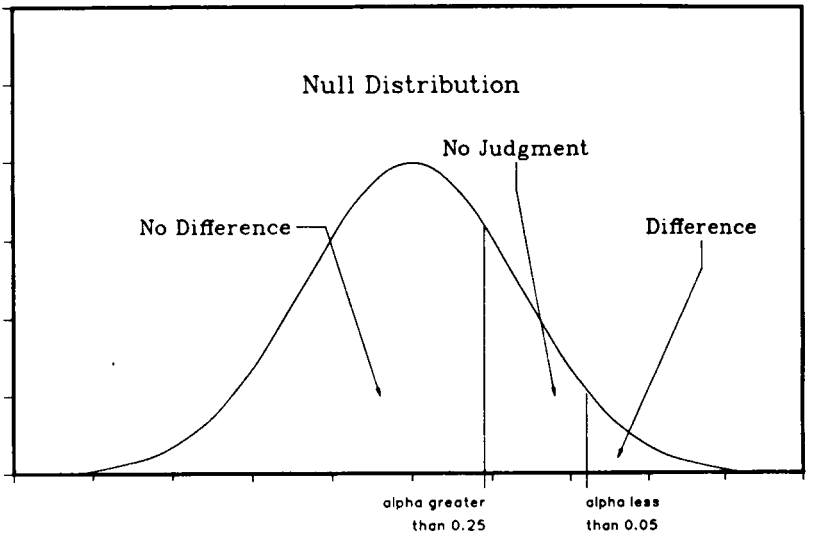

Figure 1: Three Zones of Judgment

errors could be quite large, in some quite small, but in all such cases it is unspecified. Thus this method does not allow one to reach standard statistical conclusions, such as: "The probability of being wrong when concluding no difference in this case is less than X%." Rather, this method primarily provides a metric with which to judge the size of the result in a randomized experiment (Case I), or in a quasi-experiment (Case II), with the caveat that in a quasi-experiment the metric applies to the "difference" found and not the "treatment effect." If a positive effect fails to reach significance at the one-tailed 25% level (i.e., an effect that small or smaller would be expected to occur 75% of the time due to chance), one can choose to conclude that the treatment effect (Case I) or the difference (Case II) is indeed small. While such a metric could be used for comparison-group-similarity analyses (Case III), we have argued that this is inefficient and potentially misleading and will not advise the Case III application for any of the techniques proposed for consideration in this section.

2. *Establish a confidence interval indicating a range of values for the true treatment effect, given the observed experimental outcome.*

Let us say that 95% of the time, chance alone (acting primarily through the randomization procedure in an experiment) would yield a difference of proportions of 0 ± 0.20.[2] Let us say also that the difference actually observed on the posttest in a randomized experiment was 0.15. Then the true treatment effect may be believed to be 0.15 ± 0.20 with 95% confidence.

The aim of this confidence interval is to estimate what the treatment effect truly is rather than to declare "difference" or "no difference." For the parole study data, for example, a 95% confidence interval yields $0 \pm (1.96 \times 0.067)$, or 0 ± 0.133.[3] If the study results showed a difference of proportions of 0.10, this means that, with 95% confidence, the treatment effect was 0.10 ± 0.133.[4]

In sum, unlike method number 1, above, this procedure does use statistics to estimate a meaningful parameter, in this case providing a range in which the true treatment effect is believed to lie. Further, this range provides more information than a simple conclusion of "difference" or "no difference." However, often a decision must be made: Shall we conclude "no difference" or not? As long as the most extreme value in the range would not cause one to regret concluding no difference, such a conclusion is justified (see Dunnette and Gent, 1977; Blackwelder, 1982). However, when the range includes values so large in absolute magnitude as to be worrisome, it will be difficult to have confidence in such a conclusion.

Thus the classic application of confidence intervals is most useful when the interval is entirely within the range of acceptable difference or, perhaps, clearly outside it. In what may well be the most common case, that in which the interval includes both acceptable and unacceptable values, traditional confidence intervals (e.g., 95% intervals) do not allow explicit statement of the probability that the treatment effect is negligible (or its complement, the probability of error in accepting the hypothesis of no difference when it is false).

As an alternative to this standard application of confidence intervals, one could decide ahead of time how large a difference is still negligible – effectively no difference – and calculate the confidence level corresponding to the interval (based on observed results) that puts that particular difference just at the tip of its range. One might then find that the difference is a negligible one, for example, with 97% confidence, or perhaps only with 60% confidence. This use of confidence intervals is unorthodox, and it happens also to be illegitimate in a noteworthy sense. Since it arises most naturally out of the tradition of power analysis, we will reconsider it in that context, below.

Whatever manner of confidence interval were employed, if selection bias were added to chance as a source of the null case difference, there would be no basis for believing the true treatment effect to be within the interval, which

relates only to the results of chance. Thus this technique would be appropriate only for randomized experiments (Case I), as that is the only situation in which unhappy randomization is the primary concern with respect to bias. In a nonrandomized study (Case II), bias from self-selection or other assignment sources, whose magnitude cannot be estimated by statistical techniques, cannot be ruled out, leaving one with only small justification for invoking the random-sampling model that lies at the foundation of classical interval estimation. In attrition analysis (Case III), the difference between the two groups is what it is; unlike the problem of a masked treatment effect, there is no reason to wonder about such a thing as "the range in which it truly lies."

3. *Conduct a power analysis centering on the likelihood of detecting a true treatment effect of some predetermined size.*

The method of this subsection, power analysis, would again apply mainly to randomized experiments; its use to infer causality in quasi-experiments is appropriate primarily to the extent that, initially, the treatment and comparison groups may be considered as distinguished by chance differences alone.

Power is defined as the probability of detecting a true difference of a specified size and is equal to 1 minus the probability of Type II error. A power analysis would indicate the probability of making a Type II error as a function of three values: (1) the alpha level used as a criterion in a test of significance, (2) the sample size, and 3) an effect size considered small enough that a true effect of that size or smaller could be interpreted as "effectively no difference." (Power = f[alpha, n, effect size]; Tables are readily available; e.g., see Cohen, 1987.)

For example, if one could decide that 37 or fewer convictions among the honor parolees would be a negligible increase—thus interpreted as "no difference"—over the 32 arrests for the traditional parolees (a difference of proportions of 0.05; n = 100 per group), then one could calculate the probability of making a Type II error for any particular alpha level. One proceeds by assuming that the true effect is indeed 0.37 - 0.32. When alpha is set at 0.01 for a one-tailed test, the probability of *detecting* this true effect by obtaining a significant result is only 5%; thus, the likelihood of *missing the assumed true effect* (making a Type II error) is a very high 95%.

Note how the analysis is affected by the choice of alpha and by the maximum acceptable difference (sample size remaining at 100 per group). If a much larger effect could still be considered negligible, say something less than 0.20 instead of 0.05, then, assuming the undesirable treatment effect of 0.52 - 0.32 = 0.20 to be true, the probability that one's results would be

nonsignificant at the 0.01 level (one-tailed), leading to an erroneous no-difference conclusion, would go from 95% to 31%. Returning to a difference of proportions of 0.05 as the maximum acceptable difference, if alpha were set at 0.10 instead of 0.01, then the probability of concluding no difference in error would drop from 95% to 72%, and if alpha were set at 0.40 the probability of a Type II error would drop to 23%. If alpha were set at 0.40 with a maximum acceptable difference of 0.20 (0.32 versus 0.52), the probability of Type II error would be less than 1%.

This description of power analysis allows a more rigorous explanation of the problem of lowering alpha for multiple comparisons of no difference when all differences must be negligible. Assume that had there been only one comparison one would have used alpha = 0.25. Positing the true treatment effect for each comparison to be 0.42 – 0.32 = 0.10, beta for each comparison then would be 0.33. If because of multiple comparisons the significance level were lowered, the probability of Type II error would increase — if alpha were lowered from 0.25 to 0.10 in our example, beta would increase from 0.33 to 0.72! In the parole experiment, had the effect of the honor parole treatment actually been to create too large a difference in the single criterion, "seriousness of later crimes," lowering alpha would make this unacceptable impact more difficult to detect. Rather than facing a 33% probability of error in detecting a real, nonnegligible difference, researchers lowering alpha from 0.25 to 0.10 face a 72% probability of error. Given that one has advisedly adopted the criterion that all (or almost all) differences should be negligible, the fact that the two groups are equivalent on 5 or 10 other comparisons does not make this one real difference any less important to detect. Without an adjustment in power, lowering alpha simply forces one to consider a greater difference *acceptable,* contrary to one's original, free determination. Of course, the alpha level need not be the same for all variables; it might variously be set higher or lower according to the substance and importance of each particular comparison. Once the best level for each variable is determined, however, there is no basis for lowering it simply because of multiple comparisons.

In sum, as concluded above, lowering alpha is inappropriate as a response to multiple comparisons for conclusions of no difference even when all effects must be negligible. Instead, it is proper to leave alpha at least as large as original considerations (e.g., a power analysis) determined it to be.

Power analysis as described in general terms above solved for power (and thus, beta) as a function of alpha, sample size, and maximum negligible difference. These relations suggest three procedures that would be useful in various no-difference analyses.

First, if one wishes no-difference conclusions to be analogous to conclusions of difference, the primary concern is *managing beta.* This means that beta must be restricted to an acceptably small error rate (e.g., 20%; 80% is often considered to be the minimum sufficient power; see Fagley, 1985; Cohen, 1987). The analysis would therefore proceed (A) *to solve for the alpha level required to maintain the desired small Type II error rate* (alpha = f[beta, sample size, maximum acceptable difference]). This is Cohen's (1987) fourth type of power analysis, which he contends is rarely used. With this approach one could require a power of 80%, for example (20% chance of concluding no difference when there is a real effect of a given size), and determine the alpha level that allows such power.[5] This would seem to be a sensible procedure in the sort of case being considered here. That is, one ought to be able to fix upon a maximum acceptable treatment effect and upon an acceptable risk of error in concluding no difference if the effect is actually that large. One then runs the significance test at the alpha level that yields the appropriate risk of error. In the honor parole example, a real effect as big as 10 additional convictions would be found 80% of the time when the one-tailed alpha is set at 0.28 (corresponding to 36 honor parolee convictions, a difference of proportions of 0.04; see Figure 2).[6]

As suggested previously, using an alpha level of 0.28 (to restrict beta to 0.20) as a criterion for concluding no difference does not obligate one to conclude "difference" whenever the observed result is significant beyond the 0.28 level. Rather, one could continue to reserve conclusions of "difference" for those results that extend beyond the 0.05 or 0.01 alpha levels. By setting two criterion levels—one to conclude difference, one to conclude no difference—one demarcates three zones as in the first method suggested above, including a middle zone of suspended judgment. But by tying the criterion for *no-difference* conclusions to beta and arriving at alpha by that means, conclusions of no difference are based on the standard and quite defensible notion of risk of error—an element that is lacking when using procedure #1 (e.g., testing arbitrarily at the 25% alpha level without a power analysis).

Solving for alpha as carried out in variant (A) above requires specifying the maximum acceptable difference and the risk of Type II error. The observed results are then compared to the derived critical value to reach a yes or no conclusion as to no difference. A second alternative is (B) *to determine the probability of error associated with a critical value equal to the observed difference.* One proceeds by calculating the "Z" (or "t") value that corresponds to the observed difference and entering that "Z" value into the power equation. In effect, this use of power analysis is analogous to the very common practice of supplying the significance level actually achieved

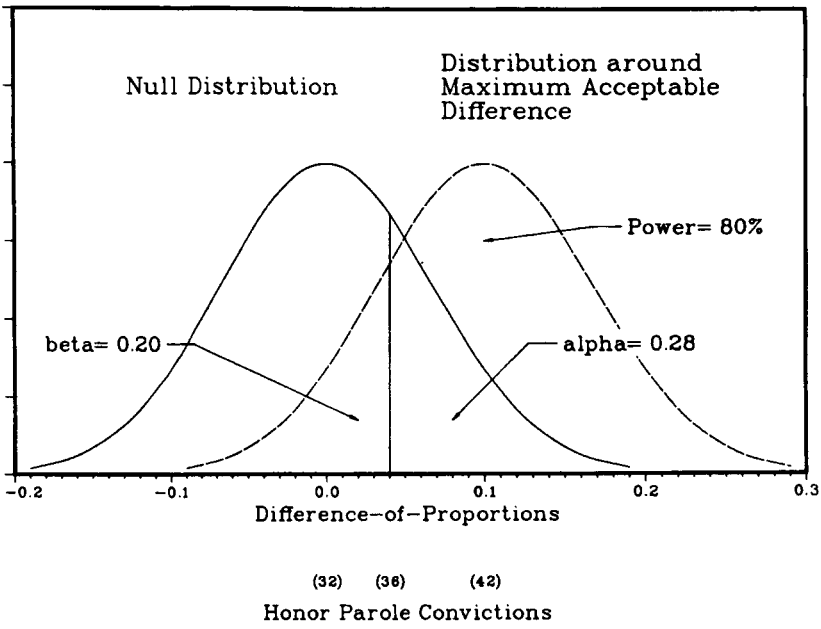

Figure 2: Managing the Type II Error Rate

by one's results in traditional testing, rather than just saying that the results were or were not significant at some predetermined level (e.g., noting post hoc that $p < .06$, or 0.08, or 0.01, or 0.001, etc.).

Figure 3 illustrates the situation where the actual results yielded 32 convictions among traditional parolees, 34 among honor parolees, and there was a declared maximum acceptable treatment effect of 10 additional convictions (42 convictions; difference of proportions of 0.10).

A difference-of-proportions test reveals that the 34 actual convictions corresponds to $Z = 0.30$, which, substituted into the power equation, yields a power of 87%. This means that if the true effect were 42 convictions, deciding ahead of time (which was not done here) to conclude no difference for 34 or fewer convictions would cause a Type II error only 13% of the time (beta = 0.13). Note that this is the maximum expected error; if the true effect were larger than 42 convictions, using 34 convictions as a criterion would yield even fewer Type II errors.

The phrase, "ahead of time," however, is critical, since the post hoc procedure of variant (B) does not yield a true probability of error (see

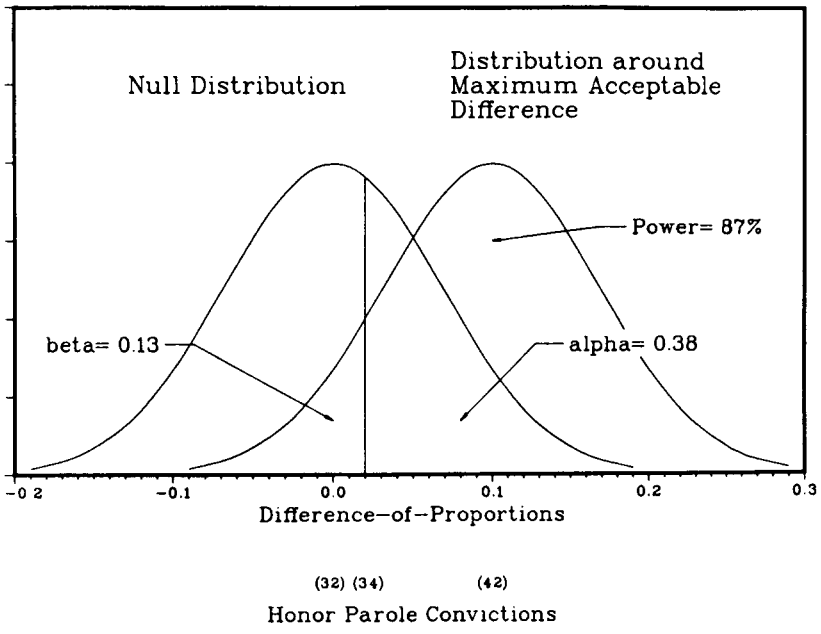

Figure 3: Calculating the Power Level Based on Observed Results

Guttman, 1981). For example, in traditional significance testing, rather than making the statement: "My experimental results are significant at the 0.06 level," which would be true only if the 0.06 level had been preestablished, a more acceptable statement would be: "If I had set alpha at 0.06, my results would have been statistically significant." This interpretation of "$p < 0.06$" is not a statement of probability or a statistical basis for a decision. Imputing such meaning to it is illegitimate. Its appeal is as an indication of the strength of the relationship or treatment effect (see the concept of "evidence" described by Goodman and Royall, 1988). Ceteris paribus, significance at the 0.001 level indicates a stronger relationship than significance at the 0.01 level, and so forth. It is, in other words a probability-based metric for strength, just as in quasi-experimental design, as noted in connection with Case II, above. The differences between the two uses are that it is applied here to the interpretation of the strength of a treatment effect (a causal relation) in a randomized experiment, and that the procedure discussed is post hoc (which,

no doubt, it would usually be), rather than one calling for preestablished levels.

In the power analysis context, rather than making the statement, "Setting alpha at 0.xx on the basis of the observed difference, the result is nonsignificant, beta < 0.13," a more acceptable statement would be, "If I had set alpha ahead of time at 0.xx, the result would have been nonsignificant, beta < 0.13." The larger the alpha level derived from the results, the smaller is that beta level, and, therefore, the weaker is the treatment effect in this metric; thus, the more justified is a conclusion of no difference.

Moreover, it is the logic of this variant of power analysis that may be seen as the root of the unorthodox application of confidence intervals suggested in the previous section (although the results of the two methods are not necessarily equivalent).[7] Consulting Figure 3, it is clear that a 74% confidence interval centered at 0.02 would just reach 0.10. One might therefore be tempted to make a statement such as: "I can conclude with 74% confidence that the true treatment effect is in this derived interval, and therefore that it is not as large as 0.10." This is, in other words, a conclusion of "no difference" with 74% confidence, since 0.10 is the maximum tolerable effect. This post hoc statement, however, has no true meaning in the language of probability; it may be communicative, but it is illegitimate in precisely the sense just noted with respect to ordinary significance testing and power analysis. We can conclude only that achieving a 74% confidence interval by this post hoc procedure is not as good as achieving a 95% interval, but is better than, say, a 65% interval. In other words, we have here an analogous metric in confidence-interval language for the weakness of the effect (the larger the confidence level, the further away is the result from the maximum negligible effect). Note that a 95% confidence level would be attained only with a power of 97.5% and a beta of 2.5%, a very tiny effect indeed under the present illustrative assumptions.

The final variant of power analysis to be considered here is (C) *to solve for the maximum expected true treatment effect, given the required power and taking the observed results as the critical value* (maximum difference = f[alpha, sample size, beta]). This formulation asks: "How large might the true treatment effect be, such that the difference of proportions actually obtained, or an even smaller one, would nevertheless appear 20% of the time through sampling error, that is, through the offsetting vagaries of randomization?" If there were 32 traditional convictions, 38 honor convictions, and a required power of 80%, the maximum treatment effect expected to yield such results would be 0.118, or less than 12 additional convictions (see Figure 4). One

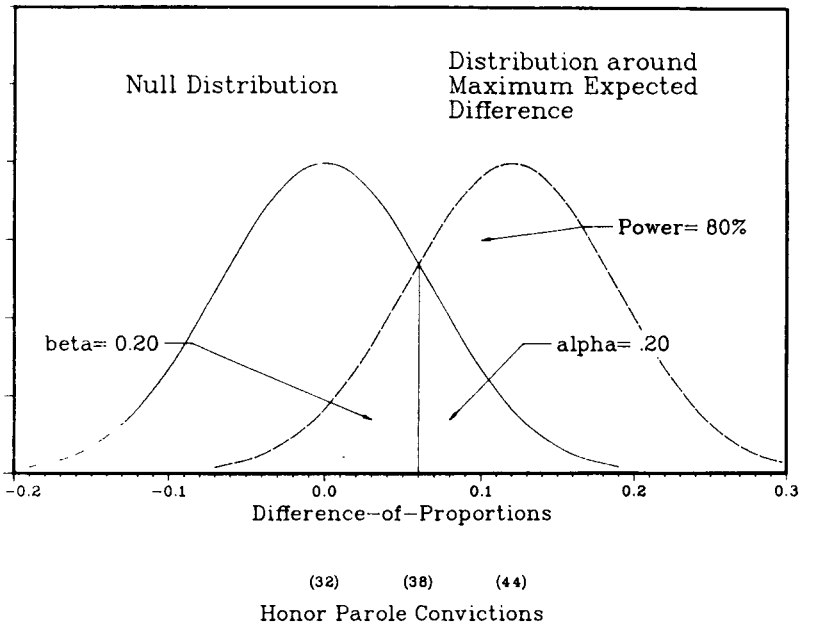

Figure 4: Maximum Expected Effect Based on Observed Results

would then have to decide if a treatment effect of 12 additional convictions (total of 44 honor convictions) would be acceptable.

One disadvantage of all power analyses is the added complexity of dealing with equations and tables not part of standard statistical practice. However, variant (C) is comparable to finding the difference of proportions that is 0.84 standard errors (the 20% tail, reading directly from the "Z" table) above the observed result of 0.38 – 0.32 = 0.06 ("comparable to" rather than "equivalent to" for the reason elaborated in note 7). That is, to assume that the true treatment effect is 0.84 standard errors above 0.06 has the effect of placing the observed difference of 0.06 at the point where, given that treatment effect, 80% of experimental differences would be larger than 0.06 (because of randomization results that may mask the true effect) and 20% would be smaller. Calculating this maximum expected true effect with multiples of the standard error is a relatively simple procedure that does not require the power tables or equations and that could be used to bracket the results of almost any experiment.

None of the general procedures reviewed in this section—high alpha, confidence interval, or power analysis—appears ideal, but each is an improvement over the policy of supporting no-difference conclusions with a failure to reject the null hypothesis at the 5% level. Traditional testing wherein one tries to *reject* the null hypothesis has been criticized as being too lenient, one argument being that "you don't have to predict the size of the effect, only that it will be bigger than nothing" (see Meehl, 1967). Leniency is also an issue when using small alpha levels or reducing alpha in no-difference conclusions. All else constant, the smaller the alpha, the lower the power and the greater the range of results that will be interpreted as confirmatory: Make alpha small enough and almost any result will do.

SUMMARY

Yeaton and Sechrest (1986) have argued that no-difference conclusions are an important part of science, warranting further attention. We have attempted to contribute to the dialogue on no-difference conclusions by addressing the methodology involved. We began by delineating three major uses of the no-difference conclusion and we then discussed its meaning and analysis within the context of each of those three uses:

Case I: Estimating the treatment effect in randomized experiments. If the studies are well designed and embody sufficient power, it is possible to conclude that a treatment had no (or little) effect with confidence in one's conclusions. Here, the question is not *whether* to use classical inference, but how (another possibility, of course, is to use Bayesian techniques; see Selwyn et al., 1981; Goodman and Royall, 1988). Unlike traditional uses of statistics, confidence in no-difference conclusions increases when one's procedure has the same sort of effect as moving from a small alpha level criterion (e.g., 0.05) to a large level (e.g., 0.25, one-tailed). If the risk of Type II error and maximum acceptable difference are specified, the precise alpha level at which it is appropriate to test may be determined. Alternatively, one could obtain a strength measure by determining the "probability of being in error" when concluding that the observed results represent no difference. Finally, one may use statistics to determine a range in which the true effect is estimated to lie; one must then decide if the outer reaches of that range may in context be considered "no difference," that is, small enough to be acceptable.

Case II: Estimating the treatment effect in a nonrandomized study. Without supporting information, it is risky outside of experimental design to

interpret a no-difference finding as good evidence that the treatment had no effect. It is quite possible that the two groups were unequal before treatment and that the treatment had a large impact that made the two groups appear similar. Only to the extent that one can treat the groups as though established by randomization can the analysis of Case I be made to apply to this category. It seems appropriate that a defense of such an assumption be made when it is used.

Case III: Using control variables to conclude that groups were equivalent before treatment or that attrition has not biased a randomized study. This use shares with Case II the problems of inference that arise when one cannot presume randomized assignment. If the study actually was a randomized experiment, then the use of statistics in this particular way is only a check against "unhappy" randomization. While such a check is legitimate, it is an inefficient and potentially misleading use of information and should be replaced by using the control variables in a multivariate analysis in which whatever differences do exist (as some almost inevitably must) are used to explain and adjust variation in the outcome variable.

Multiple comparisons were shown to have different effects on the conservatism of the analysis depending on the type of decision rule used, that is, whether decision and action require all (or nearly all) comparisons to show no difference, or require that merely one (or a few) show no difference. However, in both decision types, the primary concern must be to manage beta, not alpha, a focus that argues against using multiple comparisons as a justification for lowering the alpha-level criterion.

Having made a case for these various aspects of the methodological handling of common no-difference scenarios in evaluation research, many aspects of which we presume will be controversial, and having suggested what we would consider to be several of the most likely statistical procedures to be implemented in such circumstances in practice, we look forward to a period of debate and application in which the many issues involved will move toward resolution by further scrutiny and trial.

NOTES

1. Twenty-five is a round number. Its associated Z-score of 0.67 (also a round number) will generally reflect a small difference. We caution the reader, however, that "small" may not be sufficiently conservative with regard to Type II error, particularly in the context of small sample sizes or large variances.

2. This procedure requires an estimate of the parameter P (in our case, the proportion of population convictions "without treatment") because the width of the confidence interval is

dependent on the variance of the sampling distribution which, in the case of the difference of proportions, is in turn dependent on p, the estimated population proportion.

3. We estimate P based on the proportion of convictions in the traditional parole group, in this case 0.32. That is, we consider this value to be a best estimate of the proportion of population convictions "without treatment." In many cases, including the present one, the meaning of "without treatment" may not be very clear. An alternative to making a commitment to a specific meaning of "without treatment" is to use the most conservative available estimator of P, that is, the estimator whose observed value is closest to 0.5.

4. The technique involved is a variant on ordinary interval estimation. For convenience, call a difference of proportions a "difference." In the ordinary case, the logic of building a 95% interval estimate proceeds by asserting that (1) 95% of *all* sample differences are within 1.96 standard errors of a certain value that is the mean of the sampling distribution of such differences; and (2) the observed difference in the case of *this* sample is therefore probably within 1.96 standard errors of that value. Such reasoning does not apply to experimental design because "the observed difference in the case of *this* sample" does not represent the result of sampling alone, but rather the result of sampling (the randomization procedure) *plus* the subsequent treatment effect. Conceptually, this value does not appear in *any* sampling distribution. Rather, one builds an interval estimate of the true treatment effect by taking the observed difference as a midpoint and superimposing upon it the sampling distribution that represents the results of randomization – the sampling distribution whose true midpoint is zero, the difference expected to result from randomization. The values to the left of this newly labeled midpoint could reflect the true treatment effect because randomization vagaries could have added to it to yield the observed difference; the values to the right might also represent the true treatment effect because randomization counteracted it in some degree to produce the observed difference. The point is a conceptual one; here and below, the analysis is not affected.

5. Cohen (1987) provides tables for 0.01, 0.05, and 0.10 one- and two-tailed alpha levels; for other alpha levels, we have used the following formulae supplied by William A. Ericson:

$$\text{Power at h(where h} = \arcsin\sqrt{p_1} - \arcsin\sqrt{p_2}) = 1 - \Phi\,(-h\sqrt{n/2} + Z_{1-\alpha}) \text{ and } Z_{1-\alpha} = h\sqrt{n/2} + \Phi^{-1}(1-\beta).$$

6. Finding the alpha level (in this case, 0.28) that restricts beta to 20% for a specified maximum acceptable difference is functionally equivalent to testing an alternative null hypothesis – a null hypothesis of a specified difference (in this case, a difference of proportions of 0.10) – at the one-tailed 20% level (see Dunnette and Gent, 1977; Blackwelder, 1982).

7. If the relevant statistic were the difference of means and either the samples were large or the population normal, the two methods would be equivalent in result. With the difference of proportions, applying classical inference directly to the untransformed statistic (Cohen, 1987: 180; Blalock, 1972: 228-230), the variance of the sampling distribution (on which the estimate depends) differs with P, which in our example represents the proportion of parolees that would be convicted of crimes in the absence of treatment. Assume that P is estimated by pooling the control- and experimental-group values, p_1 and p_2. For power analysis, the estimate of P depends on the maximum acceptable treatment effect; for the confidence interval, it depends on the observed result. These values generally are not the same (in our example, 0.42 – 0.32 for power analysis versus 0.38 – 0.32 for the confidence interval). Thus the results of the procedure would ordinarily differ slightly between the two methods. Note that they could be the same, however, if P were estimated on the basis of the control-group proportion alone, as discussed in note 3.

REFERENCES

ASHENFELTER, O. (1986) "The case for evaluating training programs with randomized trials." Presented to the Education Sector of the World Bank Conference, Investing in People: New Directions for Education and Training. Hunt Valley, MD, January 7-10.

BLACKWELDER, W. (1982) "'Proving the null hypothesis' in clinical trials." Controlled Clinical Trials 3: 345-353.

BLALOCK, H. (1960) Social Statistics. New York: McGraw-Hill.

BLALOCK, H. (1972) Social Statistics (2nd ed.). New York: McGraw-Hill.

BOTEIN, B. (1964-65) "The Manhattan bail project: its impact on criminology and the criminal law processes." Texas Law Rev. 43: 319-331.

CANGELOSI, J. and J. JESUNATHADAS (1986) "The common misinterpretation of statistical insignificance." College Student J. 20: 115-120.

COHEN, J. (1987) Statistical Power Analysis for the Behavioral Sciences (rev. ed.). New York: Academic Press.

COHEN, P. (1982) "To be or not to be: control and balancing of type I and type II errors." Evaluation and Program Planning 5: 247-253.

COLEMAN, J. S., S. D. KELLY, and J. A. MOORE (1975a) "Recent trends in school integration." Presented at the annual meeting of the American Educational Research Association. Washington, DC, April 2.

COLEMAN, J. S., S. D. KELLY, and J. A. MOORE (1975b) "Trends in school segregation, 1968-73." Urban Institute. Washington, DC. (unpublished)

DAVIS, C. and J. GAITO (1984) "Multiple comparisons procedures within experimental research." Canadian Psychology 25: 1-13.

DENISTON, L., and I. M. ROSENSTOCK (1973) "The validity of nonexperimental designs for evaluating health services." Health Services Reports 88: 153-164.

DUNNETTE, C. W. and M. GENT (1977) "Significance testing to establish equivalence between treatments, with special reference to data in the form of 2 × 2 tables." Biometrics 33: 593-602.

FAGLEY, N. S. (1985) "Applied statistical power analysis and the interpretation of nonsignificant results by research consumers." J. of Counseling Psychology 32: 391-396.

FISHER, B. et al. (1985) "Five-year results of a randomized clinical trial comparing total mastectomy and segmental mastectomy with or without radiation in the treatment of breast cancer." New England J. of Medicine 312: 674-681.

FRAKER, T. and R. MAYNARD (1987) "The adequacy of comparison group designs for evaluations of employment-related programs." J. of Human Resources 22: 194-227.

GAW, A., L. CHANG, and L. SHAW (1975) "Efficacy of acupuncture on osteoarthritic pain. A controlled, double-blind study." New England J. of Medicine 293: 375-378.

GERSTENBLITH, G. et al. (1982) "Nifedipine in unstable angina. A double-blind randomized study." New England J. of Medicine 306: 885-889.

GOODMAN, S. and R. ROYALL (1988) "Evidence and scientific research." Amer. J. of Public Health 78: 1568-1574.

GREENWALD, A. (1975) "Consequences of prejudice against the null hypothesis." Psych. Bull. 82: 1-19.

GUTTMAN, L. (1981) "What is not what in statistics." Statistician 26: 81-107.

KELLING, G., T. PATE, D. DIECKMAN, and C. BROWN (1976) "The Kansas City preventive patrol experiment: a summary report," pp. 605-657 in Gene V Glass (ed.) Evaluation Studies Review Annual, Vol. 1. Beverly Hills, CA: Sage.

LALONDE, R. and R. MAYNARD (1987) "How precise are evaluations of employment and training programs." Evaluation Rev. 11: 428-451.

MAZUR-HART, S. and J. BERMAN (1979) "Changing from fault to no-fault divorce: an interrupted time series analysis," pp. 586-599 in L. Sechrest, S. G. West, M. A. Phillips, R. Redner, and W. Yeaton (eds.) Evaluation Studies Review Annual, Vol. 4. Beverly Hills, CA: Sage.

MEEHL, P. E. (1967) "Theory testing in psychology and in physics: A methodological paradox." Philosophy of Sci. 34: 103-115.

MOHR, L. (1988) Impact Analysis for Program Evaluation. Chicago: Dorsey.

PETTIGREW, T. and R. GREEN (1977) "School desegregation in large cities: a critique of the Coleman "white flight' thesis," pp. 363-412 in Marcia Guttentag (ed., with S. Saar) Evaluation Studies Review Annual, Vol. 2. Beverly Hills, CA: Sage.

SELWYN, M., A. DEMPSTER, and N. HULL (1981) "A Bayesian approach to bioequivalence for the 2 x 2 changeover design." Biometrics 37: 11-21.

SMITH, W. (1976) "Evaluation of the clinical services of a regional mental health center," pp. 343-355 in Gene V Glass (ed.) Evaluation Studies Review Annual, Vol. 1. Beverly Hills, CA: Sage.

YEATON, W. and L. SECHREST (1986) "Use and misuse of no-difference findings in eliminating threats to validity." Evaluation Rev. 10: 836-852.

YEATON, W. and L. SECHREST (1987) "Assessing factors influencing acceptance of no-difference research." Evaluation Rev. 11: 131-142.

George Julnes is Assistant Professor of Urban Studies and Public Administration in the College of Business and Public Administration, Old Dominion University. His teaching and research interests are in quantitative methods, program evaluation, organizational behavior, and analysis of the basic perspectives underlying social theory.

Lawrence B. Mohr is Professor of Political Science and Public Policy in the Department of Political Science and the Institute of Public Policy Studies, the University of Michigan. His teaching and research interests are in organization theory, quantitative methods, program evaluation, and the philosophy of social research.