CRAFT NOTE

In this randomized study 160 members of the Evaluation Research Society acted as judges to assess the attributes of research that produce credibility. The study focused on acceptability of no-difference findings, a long ignored but important domain of research. In the context of a hypothetical study, four factors were tested to determine their influence on acceptability of both no-difference and difference findings: randomization/nonrandomization, one/three outcomes, power = .80/.60, and equivalence on baseline measures of all eight/all but two of eight. Experts were asked to judge degree of acceptability and to reject or accept findings in both a no-difference and a difference study. Randomization consistently enhanced the believability of outcomes whereas other factors exerted a less consistent influence. Limitations of the study were discussed.

ASSESSING FACTORS INFLUENCING ACCEPTANCE OF NO-DIFFERENCE RESEARCH

WILLIAM H. YEATON
University of Michigan
LEE SECHREST
University of Arizona

A major problem in planning research is the lack of knowledge about evidence required to convince an intended audience to accept research findings (Sechrest, 1985). Scientists have seldom stressed that one, ultimate goal of research is to persuade and convince a relevant audience that study findings are believable. In fact, one researcher (Campbell, 1982) has used the phrase "experiments as arguments" as a metaphor to emphasize this research goal.

AUTHORS' NOTE: *This work was supported by grant HS 04825 from the National Center for Health Service Research.*

EVALUATION REVIEW, Vol. 11 No. 1, February 1987 131-142

Our knowledge of what influences acceptance of most research is incomplete, and we know less still about what influences the particular kind of research upon which we wish to focus in this article, namely no-difference research. We will use the expression "no-difference research" to refer to those studies in which no-difference results are either an intended purpose of the research or are important should they emerge. "Difference research" refers to studies that attempt to show a difference between groups or variables. With both types of studies the pattern of findings, either difference or no-difference, is of intrinsic interest to the research question(s) at hand. What we wish to emphasize is that no-difference research is an important subset of research and that no-difference outcomes can often be viewed as a desirable outcome of research rather than as an undesirable by-product of difference research.

Despite its relative neglect, no-difference research is central to many policy questions. Often, we want to determine whether to discontinue policies that no longer work (that is, lead to no difference). Do our methods of providing assistance to the poor remain effective? Does a particular drug continue to provide relief from pain? Is a given training method still successful in producing quality teachers?

No-difference research is likely to be important not only in diverse policy settings but also in a number of areas within a given policy setting. For example, health research is riddled with questions in which no-difference answers are important. We ask whether cesareans increase the risks for both mother and child (Sears, 1985), whether more invasive surgeries for early forms of breast cancer enhance survival and quality of life (Fisher et al., 1985), and whether nonnutritive sweeteners increase our chances of having bladder cancer (Havender, 1983). In each instance, a no-difference conclusion has important implications, and research addressing these questions should allow relatively unambiguous conclusions.

Despite its potential importance, no-difference research has traditionally been very difficult to publish. This difficulty is most probably due to the fact that it is relatively easier to produce no-difference results than difference results (for example, use small samples, implement weak treatments at low integrity levels, utilize unreliable outcome measures; Sechrest and Yeaton, forthcoming). In addition, the null hypothesis is never, strictly speaking, likely to be true (Bakan, 1966). With a large enough sample size, any small difference will be significant. For these and other reasons, it is very difficult to identify studies that show a "true" lack of difference, which tends to tip the scales in favor of publishing articles reporting differences.

A reasonable way to increase the odds that no-difference studies are accepted by the research community is first to examine difference research for clues about how one might conduct high-quality no-difference research. Certainly, methodological rigor is likely to enhance the believability of difference findings and should reasonably affect the acceptance of no-difference research. Use of well-controlled experimental designs, adequate statistical power, and numerous, high-quality measures are likely to prove important in difference and no-difference research. Both kinds of research are probably most convincing when they are consistent with existing theory. We have identified other, plausible aspects (such as ideological foreclosure, preponderance of evidence, context of the findings) that may be associated with acceptance of no-difference findings. (The interested reader may refer to Sechrest and Yeaton, forthcoming, for a more detailed discussion.)

After identifying factors likely to be critical to the acceptance of no-difference research, one can imagine several, different contexts in which to study the role of these factors in no-difference research. One viable possibility is to identify a fairly large set of studies and to note the presence or absence of potentially important factors in each study. Ascertaining acceptance is somewhat problematic but in our early work we have utilized the Science Citation Index to locate studies that have cited particular no-difference papers and inspected these studies to determine their acceptance or rejection of no-difference findings. Given this measure of acceptance/rejection, it is possible to compare the relative occurrence rate of potentially important factors in accepted and rejected no-difference studies.

Although the descriptive work discussed in the previous paragraph allows more informed speculation regarding the factors influencing acceptance of no-difference research, this approach will necessarily lead to ambiguous conclusions. Within a particular set of no-difference studies the presence (or absence) of a given factor (such as randomization) will be confounded by the unsystematic occurrence of other, potentially important factors (such as statistical power) influencing acceptance of these studies. The only way to avoid this inherent weakness is to design a research study that systematically alters the presence and absence of factors that may influence acceptance of no-difference results.

To accomplish this, one would first construct a description of a hypothetical study as a context for assessing acceptability of no-difference findings. Ware (1985: 705-706) addressed the need for direct research on acceptability in a recent commentary on future directions of medical

and health research. He noted: "What attributes of a research project produce credibility in your eyes? The issue deserves formal investigation. Experimental manipulation of different attributes of hypothetical studies giving the same results could be rated by expert panels."

The primary purpose of this study was to examine factors likely to be judged as critical to the acceptance of no-difference findings. Numerous factors were examined for potential inclusion in the study, but only four were chosen to ensure adequate sample size in each of the cells of the design.

These particular four factors were chosen because they represented potentially important study aspects likely to be critical in the interpretation of no-difference results and because our preliminary work in compiling descriptive characteristics of no-difference studies suggested that they would be important factors in acceptability by the research community. The four factors and a brief rationale for their inclusion were as follows: randomization, since previous studies (for example, Cohen, 1979) have indicated its importance in the judged scientific merit of research; statistical power (Freiman et al., 1978), since a no-difference conclusion and statistical nonsignificance will depend on sample size and the sensitivity of the study to detect real differences (that is, avoid Type II error); the number of measures reported (Cook and Campbell, 1979), because a pattern of no-difference findings among multiple measures would likely be more convincing than a single finding of no difference; and the equivalence of study groups on baseline variables, since nonequivalence is critical to both no-difference and difference conclusions whether or not random allocation has been utilized (Chalmers et al., 1983).

Given a decision to limit the study to these particular four factors, specific study values were chosen with the following rationale. First, a commonly accepted rule of thumb for acceptable statistical power is .80 (Cohen, 1977). A survey of 71 no-difference studies by Chalmers and his colleagues (Freiman et al., 1978) indicated that 50% of these studies had power of .60 or above (assuming alpha .05, a one-tailed test, and a 50% reduction in the clinical problem). Thus statistical power of .80 and .60 were designated to represent both desired and common values. Second, inspection of a subset of medical, no-difference studies indicated that eight was a commonly occurring number of baseline variables and that there was statistical nonequivalence on a couple of these variables in many instances. Third, one and three outcomes were chosen to reflect reliance in a single outcome as well as the case in which one or more quantitative and qualitative outcomes might be reported. Fourth, both

true experiments and quasi experiments were represented since they reflect what is probably the two most important kinds of research studies. We hypothesized that each of the four factors would have a significant influence on participant's judgments of study results. Given the lack of empirical evidence, we made no differential predictions regarding results in difference and no-difference studies, and for the same reason, we did not predict any significant interactions between any of the four factors used.

METHODS

PARTICIPANTS

Participants were randomly selected from the September 1983 membership directory of the Evaluation Research Society (ERS). Every fifth name was chosen from the alphabetical listing as a potential recipient of a questionnaire. The only exceptions were those who lived outside of the continental United States or members of the authors' research center. In the case of an exclusion, the next eligible name was chosen from the list.

A $5 incentive was offered to each potential participant for receipt of a completed questionnaire since payment has generally been found to increase response rates (Heberlein and Baumgartner, 1978). A subset of participants was provided space to respond to questions on prepayment checks to determine the effect of this format on rate of response (Yeaton and Sechrest, 1985). Since no difference in the quality of response was found between those participants who received their questionnaires in the standard way (with a promise of a $5 payment upon receipt of the completed questionnaire) and those who provided responses on their prepayment check, results in the two subgroups were aggregated.

To ensure adequate statistical power, a decision was made to sample a sufficient number of potential participants so that a total of 160 persons would complete questionnaires, 10 for each of the 16 cells in the study. Assuming that the four factors used in this study could be expected to account for 20% of the total variance of the four population means, with means assumed to be equally spaced over the range of the outcome variable, the power of an F test would be equal to .71, given an alpha level of .05 (Cohen, 1977).

A total of 208 questionnaires were sent during the first mailing. After approximately a month and a half, a second mailing was made to

complete those cells of the design that were not yet full. An additional 118 questionnaires were sent with the proviso that the number sent equal two times the number of respondents needed to fill a particular cell. Potential participants were chosen for this second mailing by identifying each nonrespondent in the 1983 Membership Directory and selecting the next two persons from the directory. Potential new respondents were assigned to the same cell of the design as the nonrespondent with whom they had been paired. The same exclusion criteria applied as described earlier. A more recent version of the ERS Membership Directory (October 1984) was used to update addresses. These two mailings filled most of the cells in the design.

Two months later, a brief reminder letter was sent to 8 potential participants from the second mailing to complete the few remaining unfilled cells. Where there were more respondents than necessary to fill a cell, the first 10 completed questionnaires were used.

All potential participants received a four-page questionnaire. On the first page, participants were introduced to the purposes of the research, provided general instructions, offered a $5 payment for participation, asked to complete responses to the description of the first, hypothetical study (no-difference research) before reading the second, hypothetical study (difference research), and told that further instructions would follow on page four. Pages two and three contained the same description of a research study designed "to evaluate a method to increase adherence to a prescribed treatment regimen." On page two it was stated that the findings of the research indicated that there was no statistically significant difference between groups. On page three the findings were described as statistically significant, and a statement to this affect was underlined on the questionnaire, emphasizing the difference between results on pages two and three.[1]

Pages two and three were identical for a given respondent except that a no-difference finding was reported on page two and a difference finding was reported on page three. Four important dimensions of the research context were systematically altered in one of two ways to establish 16 experimental conditions; random assignment or matching without random assignment; equivalence on all 8 or all but 2 of 8, relevant baseline variables; the existence of one or three important outcome measures; and the power of the statistical test, .80 or .60.

On both pages two and three, participants were instructed to answer two questions. On page two they were asked to circle a number from 1 to 10 that indicated "the degree to which you are persuaded by the findings of the study that there is no significant difference" between groups, and

to circle either "accept as equivalent" or "reject as equivalent." Two analogous questions were asked on page three for the difference case.

On page four, half the participants were requested to circle their answer on the answer sheet provided, return the sheet in a stamped, addressed envelope, and wait to receive $5 for their efforts. The other half were asked to transfer their responses from the answer sheet to the top of the enclosed $5 prepayment check on which they could circle their answers and endorse and cash the check.

To summarize briefly, four factors (randomization/no randomization; baseline equivalence, all eight measures/all but two of eight measures; one/three outcomes; power = .80/.60) were systematically altered to establish a $2 \times 2 \times 2 \times 2$ factoral design.

In the format used in this study, the same participants received a research scenario describing no-difference findings as well as a difference scenario. Since it would have required twice as many participants to provide difference and no-difference findings to separate groups of subjects, this option was judged too expensive. However, we instructed each potential participant to answer the questions on page two (the no-difference case) *before* turning to page three. Given our emphasis on no-difference results, this strategy restricted our strongest conclusions to the no-difference case, yet allowed us to report results pertinent to the difference case as well.

Two separate sets of analyses were conducted for both the difference and the no-difference case. Distinct ANOVA analyses were conducted for the 1-10 scale results in the difference and no-difference cases. Similarly, a logit analysis was conducted for the accept/reject results in both the difference and no-difference case.

RESULTS

A total of 329 questionnaires were sent, and responses from 160 and 165 participants who returned questionnaires were used in calculating study outcomes (response rate = 49%). Five questionnaires were omitted because they were returned late, had erased the cell identifier, or had provided unusable responses.

A multiple regression analysis was conducted on the 1-10 scale scores making it possible to test the statistical significance of all four factors in both the presence and absence of interaction terms. (See Table 1 for pertinent results.) In the no-difference case, the overall F was significant

TABLE 1
Factors of Acceptability: 1-10 Scale and Accept/Reject Scores in No-Difference and Difference Scenarios

	No Difference	*Difference*
Study Factors: 1-10 Scale		
$\overline{A}$	6.1	6.1
$\overline{A}$'	5.1	5.2
$\overline{B}$	5.7	5.7
$\overline{B}$'	5.4	5.6
$\overline{C}$	5.9	6.2
$\overline{C}$'	5.3	5.1
$\overline{D}$	5.6	5.7
$\overline{D}$'	5.5	5.6
Study Factors: Accept/Reject		
$\overline{prob}$ A	.300	.394
$\overline{prob}$ A'	.244	.281
$\overline{prob}$ B	.262	.338
$\overline{prob}$ B'	.281	.338
$\overline{prob}$ C	.300	.350
$\overline{prob}$ C'	.244	.325
$\overline{prob}$ D	.275	.388
$\overline{prob}$ D'	.269	.288

NOTE: A, randomization; A', no randomization; B, all baseline measures equivalent; B', all but two baseline measures equivalent; C, three-outcome measures; C', one-outcome measure; D, power, .80; D', power, .60. Overbars indicate means; and prob = probability.

(p less than .05, F = 2.52, df = 4/155), the randomization factor was significant at the .05 level, and the factor describing the number of outcomes reported was significant at the .10 level (the mean, randomization score was 6.1 whereas the mean, nonrandomization score was 5.1; the mean, three-outcomes score was 5.9 whereas the mean, one-outcome score was 5.3). When interaction terms were entered into the model, none of the overall F values were significant, and none of the main effect or interaction terms were significant.

A logit analysis was conducted in the no-difference case to determine if any models with the four factors, either with and without interactions, were significant. None of the models tested explained a significant amount of variation. (Among the models tested were the following: one that contained each of the four factors, one that contained each of the

six double interactions, and one that contained factors found significant on the 1-10 scale.)

Analogous analyses were conducted in the difference case. When only main effects were tested in the model, the randomization factor and the factor describing the number of outcomes reported were significant at the .01 level. The overall $F = 3.94$ ($df = 4/155$) was also significant at .01 (the mean, randomization score was 6.1 whereas the mean, non-randomization score was 5.2; the mean, three-outcomes score was 6.2 whereas the mean, one-outcome score was 5.1). When interaction terms were entered into the model, none of the main effect or interaction terms were significant.

A logit analysis was also conducted on the accept/reject results in the difference case. Here, the model with both the randomization factor and the power factor explained a significant amount of the variation. (The probability of accepting results with randomization was .394; without randomization the probability was .281. The probability of accepting results with power = .80 was .388; with power = .60 the probability was .288.)

DISCUSSION

The use of a random allocation strategy was judged to be an important factor by evaluators presented with a description of hypothetical research with no-difference outcomes, at least when asked to indicate on a 1-10 scale the degree of their belief in the findings. The presence of randomization also increased the percentage of respondents who accepted rather than rejected study results. However, although the pattern of results was replicated with this accept/reject measure, these findings were not statistically significant.

Including multiple outcomes also influenced the degree of acceptance of no-difference results (though the level of statistical significance was only .10) and, again, the pattern but not its statistical significance was replicated with the accept/reject measure.

In the case of difference findings, randomization was judged to be an important factor for both the 1-10 and the accept/reject measures. Providing multiple outcomes had an important influence on the degree of belief in outcomes but did not have a substantial impact on the decision to accept or reject results. On the other hand, utilizing greater statistical power had a substantial influence on a decision to accept or

reject findings but did not influence the degree of belief in difference findings.

The most consistent finding from this research is that random allocation has a substantial bearing upon judgments of research findings. This effect was evident in both no-difference and difference research irrespective of the outcome measure. The influence of other factors was less consistent. For example, reporting multiple outcomes also held some importance but ensuring a high degree of statistical power had no obvious influence in the no-difference case despite its potential impact in interpreting results.

It is certainly not surprising that randomization would emerge as an important factor in the acceptance of no-difference studies. Prominent evaluation researchers (for example, Cook and Campbell, 1979; Riecken and Boruch, 1974) have promoted it as a critical procedure for reducing uncertainty and providing relatively unambiguous conclusions to difference studies. In addition, we have found that methodological experts judge randomnization to be the most important single tactic in planning no-difference research (Yeaton and Sechrest, 1986). Its importance was consistent across eight different research contexts (for example, previous studies are flawed, the intervention is expensive, there are strong ideological leanings toward the effectiveness of the intervention). In light of this, it is bothersome that participants rated randomization to be more important than baseline equivalence since this is the primary purpose of randomization. It is hoped that this article will help to educate researchers not to naively assume that randomization automatically eliminates the possibility of group inequality.

Although it is tempting to extend the results of this study to other research contexts, there were several inherent aspects that may limit its generality. Only one brief, hypothetical research study was presented to each researcher, and the difference and no-difference aspect of the research were specifically pointed out in the materials provided. Many details of the research context were omitted for the sake of brevity, and identical descriptions were given simultaneously to all respondents for both the no-difference and difference cases. The decision to study the acceptance of no-difference research in an empirical manner necessarily eliminated the possibility of asking participants to examine "real" studies and rate factors whose context could not be manipulated. Thus since there were only two levels chosen for each of the four factors in this empirical study, the findings cannot technically be generated beyond these values. However, the values were realistic, based on characteristics of no-difference studies commonly found in studies of the literature and

in those at our disposal, and represented important design features in no-difference and difference research. Other potentially important factors, not included in this research, may also have affected judgments of acceptability (for instance, duration of follow-up, appropriateness of statistical analysis). However, such speculation can best be addressed by future research. Certainly, the current research provides one viable paradigm for addressing these issues.

Based on this study, the most conservative recommendations for researchers interested in maximizing the acceptability of their no-difference and difference research is that they utilize random allocation, report multiple outcomes, and use statistical power of at least .80. By randomly allocating participants to groups, baseline equivalence is likely to be achieved, and if it is not, a small amount of nonequivalence is unlikely to affect judgments of the findings. Although other practices may also influence the acceptability, suggestions regarding their use are beyond the scope of this research.

NOTE

1. In one instance on page three of the questionnaire, the word "equivalent" was inadvertently included instead of "different." However, several very specific contexts made clear the fact that a difference study was being described. For example, the statement that the there was a statistically significant difference between groups on the measure(s) provided in the study was underlined and explicitly contrasted to the no-difference scenario described on page two. In addition, the two response options included in the second question were "accept *as different*" and "reject *as different*" (underlines in original). Given the redundancy in both the questions and the context to indicate that the scenario on page three characterized a difference study, we judged the mistake to be minor. In addition, several participants noted the mistake and gave written explanation that they had responded to a difference scenario.

REFERENCES

BAKAN, D. (1966) "The test of significance in psychological research." Psych. Bull. 66: 423-437.

CAMPBELL, D. T. (1982) "Experiments as arguments." Knowledge 3: 327-1337.

CHALMERS, T. C., P. CELANO, H. S. SACKS, and H. SMITH (1983) "Bias in treatment assignment in controlled clinical trials." New England J. of Medicine 309: 1358-1361.

COHEN, J. (1977) Statistical Power Analysis for the Behavioral Sciences. New York: Academic Press.

COHEN, L. H. (1979) "Clinical psychologists' judgments of the scientific merit and clinical relevance of psychotherapy outcome research." J. of Consulting and Clinical Psychology 47: 421-423.

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

FISHER 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: 665-673.

FREIMAN, J. A., T. C. CHALMERS, H. SMITH, and R. R. KEUBLER (1978) "The importance of beta, the type II error, and sample size in the design and interpretation of the randomized control trial." New England J. Medicine 299: 690-694.

HAVENDER, W. R. (1983) "The science and politics of cyclamate." Public Interest 71: 17-32.

HEBERLEIN, T. A. and R. BAUMGARTNER (1978) "Factors affecting response rates to mailed questionnaires: a quantitative analysis of the published literature." Amer. Soc. Rev. 43: 447-462.

RIECKEN, H. W. and R. F. BORUCH (1974) Social Experimentation. A Method for Planning and Evaluating Social Intervention. New York: Academic Press.

SEARS, C. (1985) "Back from Caesareans? New data show normal delivery often safer." Amer. Health (June): 11.

SECHREST, L. (1985) "Experiments and demonstrations in health services research." Medical Care 23: 677-695.

SECHREST, L. and W. H. YEATON (forthcoming) "Role of no-difference findings in medical research."

WARE, J. E. (1985) "Commentary: monitoring and evaluating health services." Medical Care 23: 705-709.

YEATON, W. H. and L. SECHREST (1985) "Testing a strategy for increasing questionnaire return rate: providing space for responses on the prepayment check." (manuscript)

YEATON, W. H. and L. SECHREST (1986) "Assessing tactics in planning no-difference research." (manuscript)

William H. Yeaton is Assistant Research Scientist at the Institute for Social Research at the University of Michigan. His current research interests include medical technology assessment and evaluation research methodology, especially in the area of health.

Dr. Lee Sechrest is Chairman of the Department of Psychology at the University of Arizona. His current research interests include the relationship between the quality of research methods and research outcomes and the utilization of these findings with regard to policy.