LIMITATIONS OF THE INTERPRETATION OF MEASURES OF EFFECT SIZE IN EVALUATIVE RESEARCH[1]

Lee Sechrest and William H. Yeaton, University of Michigan

A problem that researching psychologists have been aware of for years but that has had increasing attention over the past decade or so is how to determine just how large an effect is produced by an experimental intervention. Especially for psychologists working in applied areas it is important to know more than that a treatment produces a statistically significant main effect. It is also important to gauge the effect size produced and to know what might be expected in the way of change if the treatment is implemented widely.

For purposes of the present discussion an experimental effect is simply the difference between measures obtained from experimental and control groups. Gilbert, McPeek, and Mosteller (1977), for example, in their study of outcomes of surgery expressed the effect in terms of "Innovations Minus Standards", meaning results obtained with new treatments minus results obtained with standard treatment. To be concrete, the effect produced by an analgesic is the difference in measured distress between the experimental and placebo drug groups. If, on the average, the control subjects report headaches of 68 on a 100-pt. scale and the experimentals report headaches of only 53, then the experimental effect is 68-53= 15 points of pain reduction. Another example would be provided by a special reading program that increased the reading level of an experimental group by 1.75 grades during a time in which a control group gained 1.00 grades. Experimental effects are estimated by such comparisons but may be estimated by any other design, e.g. quasi-experimental, confidence in the estimate varying with quality of the design, the research generally, and the data. Extension of this idea to other kinds of data and variables is straightforward. An experimental effect could be expressed in terms of proportions of each group showing improvement. One could examine differences between correlations as well, e.g., studying the "effect" of sex on the relationship between ability and performance.

The problem posed by the example given is that we do not have any meaningful way of assessing the magnitude of the change produced. How important is a 15 point reduction in reported pain? Would a drug that decreased diastalic blood pressure by an average of seven millimeters be worth using in place of other drugs? Suppose an early childhood compensatory education program produced a mean IQ advantage of six points over a control group. Is that a meaningful and important advantage? Finally, suppose that in a particular job women experienced seven percent more lost time from injuries than men. Would that be a large enough difference to warrant preferential hiring of men? How big a difference would be enough? We cannot answer these or similar questions without some more directly interpretable measure of the magnitude of an effect than a mean difference and significance level.

Statistical approaches

We looked to statistical approaches to the solution of the problem of effect size estimation, with eventual disappointment. Perhaps for want of a better device, authors often resort to statistical significance as an index of effect size, often implying that there is at least some fairly direct relationship between the statistical significance of a finding and its importance in the real world. Thus, for example, it is fairly common to find authors noting that a finding is "highly" significant or "very" significant, or reporting p values to four, five, or even six decimal places. The desire to have some basis for interpreting results seems so strong that caution is dissolved in p values.

A widely prevalent concept of effect size involves the notion of accounting for variance. Several statistical indices utilize this notion of proportion of variance explained as an indication of the importance of research finding. Eta squared (Kerlinger, 1964), omega squared (Hays, 1973), epsilon squared (Kelley, 1935), Friedman's r_m (Friedman, 1968), and the coefficient of utility (Bolles & Messick, 1958) are examples of statistical attempts to evaluate the magnitude of experimental effects. Unfortunately, proportion of variance accounted for is an inherently deficient concept upon which to base inferences about the importance of treatment.

There are also some problematic statistical phenomena associated with the various measures of proportion of variance accounted for. The most problematic is whether the error term should include an estimate of total variance or merely the source plus error. The specific estimates of variance accounted for differ in respect to the error term and can result in marked discrepancies. We have discussed these problems in a previous paper (Sechrest and Yeaton, 1980), but in general we regard omega squared as the preferable (and conservative) estimate.

As a general proposition it can be stated that all measures of variance accounted for are specific to characteristics of the experiments from which the estimates are obtained, and therefore the ultimate interpretation of proportion of variance accounted for is a dubious prospect at best. There are, in fact, several determinants of the variance that can be accounted for within an experiment, and there are only very inexact ways of knowing or estimating the importance of these determinants.

Built-in variance

First, the total variance to be accounted for will vary as a consequence of how much variance is built into the experiment. Thus, if experimental subjects are quite heterogeneous in factors associated with scores on dependent measures, there will be a larger total variance than if subjects are homogeneous. One of the vary characteristics that makes laboratory rats so useful as experimental subjects is that they have been bred to a state of maximum homogeneity with respect to a large number of characteristics, thus reducing the variance needed to be accounted for. Failures to replicate otherwise consistent results may often be explained by the heterogeneity of the subject sample used in the study. Such failures could be regarded simply as limits on the generalizability of findings, but unless the subject samples are carefully assessed and described

that interpretation might not be evident. Boruch and Gomez (1977) have commented forcefully on the decrease in effects achieved in moving from initial tests to implementation, and some of that may stem from increased heterogeneity. We think it is interesting that decisions about whether effects have been replicated almost always hinge on differences between central tendencies; rarely is the question raised whether variances have been replicated.

Experimental precision

Another determinant of total variance in an experiment is the precision achieved in planning and implementing the experiment. Consider, for example, the almost certain difference between otherwise identical experiments when one of them involves only a single, very motivated experimenter, while the other involves several experimenters with little direct interest in the outcome. The second experiment would certainly have a greater total variance, a larger error term, and the apparent experimental effect would be smaller. There are many sources of imprecision that might cause the two experiments to differ even if the same experimental treatment is being employed. Degree of standardization of experimenter demeanor, clarity of instructions, calibration of apparatus, degree of control achieved with respect to the experimental manipulation, reliability of outcome measures, and many other factors will affect total variance to be explained, and, consequently, proportion of variance explainable by any given variable.

Should one want to compare the variance accounted for by two treatments within the same experiment, it would be important to recognize that the treatments may contribute differentially to error variance, that is, one treatment might be implemented with considerably more precision than the other. Consider as an instance an experiment in which a drug and a behavioral intervention are to be jointly tested. It may be possible to achieve more careful control over drug dosage than over the behavioral manipulation. In such a case, one might be seriously misled about the potential magnitude of the effect produced by the drug, since it would be judged not in terms of its own characteristic error but in terms of the total error associated with it and the behavioral manipulation.

Number of treatments

Another factor which determines the variance one can account for in an experiment is the number of treatments being tested within the experiment. Some of the formulae for calculating proportion of variance accounted for use in the denominator a term reflecting the SS for the effect in question plus SS for error. In general we would expect that the more effects that are being analyzed for, the smaller the error term would be. Thus, one could expect to account by any one variable for a larger proportion of the variance when one or more other variables are being simultaneously studies. Therefore, in multifactor experiments, omega squared which uses an estimate of total variance in the denominator, would always be smaller than eta squared and r_m, which use only source plus error.

Strength of treatments

The proportion of variance accounted for depends on the strength of the experimental treatment. A weak treatment could only account for a small proportion of the variance in most experiments while a strong treatment could account for a large proportion. It is, in practice, usually possible to distinguish between a strong treatment with a small effect, since in the social sciences. we rarely have any independent measure of the strength of treatment administered. In only a few cases do experimenters even attempt to determine the strength of treatment employed, other than by its effect on the dependent variable. When the attempt is made, it is often by means of a "manipulation check" whose meaning can be taken quite literally. To show, for example, that experimental and control groups differ as they should on a seven-point rating scale, i.e. significantly, gives us not a clue about the strength of treatment beyond the fact that it was different between the two conditions. How many scale points of difference between experimental and control groups means would be indicative of a moderately strong treatment? of a very strong treatment? Without some way of assessing the strength of treatment, it does not make a lot of sense to talk about the proportion of variance it accounts for since one would not know the potential value of the treatment. One field experiment that used a manipulation check was the Kansas City study of different levels of police patrol, (Kelling et al. 1975). Samples of citizens in different areas were questioned about their awareness of police presence, and those in the more intensively patrolled areas did indicate higher awareness, but only if they had been pretested and, presumably, sensitized.

Range of treatments

Still another limitation on interpretations of proportion of variance accounted for is that any treatment involving more than two levels, a simple estimate of proportion of variance accounted for can obscure far more than it reveals. If one were testing the effects of two alternative drugs for controlling blood pressure, even if one of the drugs were more effective than the other, relatively little of the variance in terminal blood pressure might be accounted for by the treatment effect. If, however, one added an untreated control group to the experiment, the treatment effect might seem almost magically to have been greatly increased. A particularly apt example has been provided by Levin (1967). He described an experiment with six experimental conditions analyzed by a one-way ANOVA, with the result that omega squared was 37%. However, subsequent analysis indicated that over 85% of the explained variation was attributable to the superiority of one group to all the others.

Real world variance

One final problem in the interpretation of proportion of variance accounted for has to do with its "external validity", that is, relationship to any "real world" context in which one might want to draw inferences about the probable effect on some intervention. The problem is that variance within an experiment may not be the same variance as occurs in the "real world". To begin with, the variance that exists within an experiment depends largely on how the experimenter plans and implements the experiment. When an experimenter studies interpersonal attractiveness as a function of attitudinal similarity and phys-

ical attractiveness, all other sources of variance in interpersonal attractiveness are controlled out of the experiment to as great an extent as possible, thus reducing the error term (that is, the amount of unexplained variance) to a value below that likely to exist in an extra-experimental context. Undoubtedly the same sort of phenomenon occurs in most experiments, whether in the laboratory or field.

A second factor limiting direct comparisons of proportion of variance accounted for in the laboratory and in the real world is the great likelihood that treatments tested are not representative of those found outside the experiment.

Experimental treatments may be either more or less extreme than those common in the real world. For example, experimental studies of punishment with human subjects could never include any as extreme as are found in the real world, not even if the experimental punishment and real world comparisons were restricted to verbal abuse. On the other hand, experimenters can arrange treatments that are more extreme than those likely to be encountered by most of the subjects they would be studying, a case in point being Milgram's (e.g., 1963) studies of obedience in which subjects were enticed into delivering what were apparently severe electric shocks to another person. Un-representative treatments simply do not allow meaningful generalization to real-world instances. One of us (Sechrest) has recently reviewed a proposal to study effects of interviewer training that includes at the high end an amount of training quite unlikely to be encountered in any research organization. The results of the study could prove quite misleading if couched in terms of proportion of variance in interviewer skill accounted for by training.

A third limitation of generalizing from an experiment to the real world about proportion of variance accounted for is that because a variable can be shown to account for variance, it should not be assumed that it does account for that variance. What is the real-world counterpart for an experimental treatment consisting of being told that another punative experimental subject agrees with you exactly on a ten item attitude questionnaire? We do not want to be thought to be arguing that attitude similarity has no effect outside the experimental social psychology laboratory, but we would argue that the fact that attitude similarity can be made to affect responses in the laboratory to some degree does not mean that attitude similarity has the same effect, let alone to the same degree, the the extra-experimental world.

Empirical approaches

It does not appear to us that any purely statistical method for assessing magnitude of effects is going to be satisfactory if one enters the realm of practical decision making. There are innumerable alternative possibilities that might be considered, but we are interested in developing empirically based procedures for deciding when experimental effects are big enough to be important. We have been working on this problem for some time and no simple solutions have emerged. Instead we have suggested a host of partial solutions to the problem though all can be placed into the two major categories of approaches we have termed judgemental and innovative. Included among these approaches are the following:

1. Expert judgments. One could ask experts in a field to indicate whether they regard an effect achieved in an experiment as important or even ask experts to scale the importance (magnitude) of the effect. That is essentially what Gilbert et al. (1977) did in their study of surgical outcomes when they relied on statements from the medically sophisticated investigators about whether the improvement produced by the innovation was large, small, or inconsequential. We have found that experts in smoking research can judge the strengths of smoking interventions in such a way that their judgments correlate reasonably well (about .50) with obtained outcomes (Yeaton and Sechrest, in press).

2. Absolute and relative standards of treatment effectiveness. One might for some interventions specify a standard that, if achieved, would justify testing or even implementation of a treatment. One such standard is "normalization" (Kazdin, 1977). For many interventions all that is expected is that a person (or group, or organization) be brought to a normal state, in which case, the issue of effect size is finessed. Ciarlo (1977), for example, has community norms for a number of psychological variables, and treatment of various types of psychiatric patients is judged successful if those patients are brought within normal range on those variables. In a study of a treatment designed to improve the speaking ability of disadvantaged adolescent girls, Minkin, Braukmann, Minkin, Timbers, Timbers, Fixsen, Phillips, and Wolf (1976) found that after treatment videotapes of the treated girls received ratings comparable to those of "normal" girls of the same age (a Turing machine solution).

3. Treatment effect norms. If data were available for enough tests, one could establish norms for achieved effects, and treatments could then be evaluated by whether they produce larger or smaller effects than those usual in the field. We have attempted to assemble such norms for smoking treatments (Sechrest & Yeaton, in press), and Smith and Glass (1977) have done so for psychotherapies. We warn, however, that the almost utter lack of standardization in methods of assessing outcomes and reporting results makes assembling of norms very difficult, even where studies are numerous.

4. Benefit, cost, and risk analysis. Effect sizes may in some cases be expressed in terms of benefits, especially if they can be monetized, cost effectiveness, or reductions in risk. In Seattle, for example, Mobile Cardiac Care Units were evaluated in terms of dollar cost per life save (about $3500) (Cobb & Alvarez, undated). Benefit-cost and cost effectiveness analyses are never easy, but when they can be done, they are highly informative. Risk analysis is not always straightforwardly interpretable, but risks have been quantified for many variables (e.g. Wilson, 1979), and it may be useful to know, for example, that the reduction in risk from not eating a half-dozen eggs is about the same as that achievable by not smoking 1.4 cigarettes, by not staying in New York or Boston two days, or by not taking a 300 mile automobile trip. Incidentally, another way of assessing a campaign to lower egg consumption is to note that halving egg consump-

tion over a life-time would increase anticipated life-span by ten days (Vaupel & Graham, 1980). We do not pretend that these possibilities are without fault, but we do believe that they merit serious consideration and further study. The need for quantifying estimates of the effects we achieve by experimental interventions is compelling, and statistical approaches appear to provide no solution.

[1]Preparation of this paper was supported by grant Number 1 R01 HS02702 from the National Center for Health Services Research.

REFERENCES

Bolles, R., & Messick, S. Statistical Utility in Experimental Inference. Psychological Reports 1958, 4, 223-227.

Boruch, R.F. & Gomez, H. Sensitivity, Bias, and Theory in Impact Evaluations. Professional Psychology, 1977, 8, 411-434.

Ciarlo, J.A. Monitoring and Analysis of Mental Health Program Outcome Data. Evaluation, 1977, 4, 109-114.

Cobb, L.A., & Alvarez, H., III. Medic I: The Seattle System for Management of Out-of-Hospital Emergencies. Unpublished manuscript, University of Washington and Harborview Medical Center, undated.

Friedman, H. Magnitude of Experimental Effect and a Table for Its Rapid Estimation. Psychological Bulletin, 1968, 70, 245-251.

Gilbert, J.P., McPeek, B., & Mosteller, F. Statistics and Ethics in Surgery and Anesthesia. Science, 1977, 198, 684-689.

Hays, W.L. Statistics for the Social Sciences. New York: Holt, Rinehart, & Winston (Second Edition), 1973.

Kazdin, A.E. Assessing the Clinical or Applied Importance of Behavior Change Through Social Validation. Behavior Modification, 1977, 1, 427-452.

Kelley, T.L. An Unbiased Correlation Ratio Measure. Proceedings of the National Academy of Sciences, 1935, 21, 554-559.

Kelling. G.L., Pate, T., Dieckman, D., & Brown, C.E. The Kansas City Preventive Patrol Experiment: A Technical Report. Washington, D. C.: Police Foundation, 1975.

Kerlinger, F.H. Foundations of Behavioral Research. New York: Holt, Rinehart, & Winston, 1964.

Levin, J.R. Comment: Misinterpreting the Significance of "explained variation". American Psychologist, 1967, 22, 675-676.

Milgram, S. Behavioral Study of Obedience. Journal of Abnormal and Social Psychology, 1963, 67, 371-378.

Minkin, N., Brankmann, C.J., Minkin. B.L., Timbers, G.D., Timbers, B.J., Fixsen, D.L., Phillips, E.L., & Wolf, M.M. The Social Validation and Training of Conversational Skills. Journal of Applied Behavior Analysis, 1976, 9, 127-139.

Sechrest, L, & Yeaton, W.H. Estimating Magnitudes of Experimental Effects. Manuscript submitted for publication, 1980.

Sechrest, L., & Yeaton, W.H. Assessing the Effectiveness of Research: Methodological and Conceptual Issues. In S. Ball (Ed.), New Directions in Evaluation Research. Jossey-Bass Monographs: San Francisco, in press.

Smith, M.L. & Glass, G.V. Meta-analysis of Psychotherapy Outcome Studies. American Psychologist, 1977, 32, 752-760.

Vaupel, J.W., & Graham, J.D. Egg in your bier? The Public Interest, 1980, Number 58, 3-17.

Wilson, R. Analyzing the Daily Risks of Life. Technology Review, February, 1979, 41-46.

Yeaton, W.H., & Sechrest, L. Empirical Approaches to Effect Size Estimation in Health Research. In P.M. Wortman (Ed.), Methods of Evaluating Health Services. Beverly Hills, California: Sage, in press.