BEHAVIORAL ASSESSMENT 3, 93–100, 1981

Pergamon Press Ltd.
Printed in the USA

A Cautionary Note on the Use of Omega Squared to Evaluate the Effectiveness of Behavioral Treatments

CHRISTINE MITCHELL AND DONALD P. HARTMANN

University of Utah

Estimating the magnitude of treatment effects has been recommended as a solution to the problems associated with conventional hypothesis testing. In comparison to tradition statistical tests of treatment effectiveness, omega squared (ω^2) and related magnitude of effect statistics provide a graduated rather than a dichotomous judgmental aid, index the strength of the relationship between treatment and outcome, and are unaffected by aspects of statistical power related to sample size. Unfortunately, these correlational statistics also have characteristics that limit their interpretative value. The size and therefore the meaning of ω^2 varies as a function of (a) the reliabilities of both the treatment and outcome variables, (b) the subject, treatment, and other independent variables included in the investigation, and (c) the formula used to calculate ω^2. The interpretation of ω^2 also depends upon the conceptual clarity of the treatment variables. Finally, comparisons between ω^2's are problematic due to the limited information available concerning their sampling distribution. Indeed, the *uncritical* use of ω^2 to assess the effectiveness of behavioral treatments may represent a solution that is as troublesome as the problems the statistic was intended to remedy.

The evaluation of treatment outcome involves complex empirical and conceptual issues including the breadth, importance, and durability of performance changes; the cost and efficiency with which these changes are brought about; and consumers' satisfaction with these changes (see, for example, recent discussions by Hartmann, Roper, & Gelfand, 1977, and by Kazdin & Wilson, 1978). Despite the general appeal of these criteria, conclusions about behavioral interventions typically are more narrowly based on comparisons of treatment performance with one or more standards. These standards may include pre-treatment performance, alternative or no-treatment performance, normative performance, or ideal performance. Whichever standards are employed, decisions in group-design behavioral treatment studies often are based on the results of conventional statistical tests of the differences between means. These

Requests for reprints should be addressed to Donald P. Hartmann, Department of Psychology, University of Utah, Salt Lake City, UT 84112.

0191-5401/81/010093-08$01.00/0

tests only reveal the probability that a comparison is due to chance alone. They fail to assess and may even obscure the importance of treatments (e.g., Kazdin, 1980).

The inadequacies of these statistical tests (see Morrison & Henkel, 1970) have provoked a number of writers to suggest alternative or supplementary statistical procedures for evaluating treatment effectiveness. One class of supplementary procedures recommended by behavior therapists (e.g., Mahoney, 1977) are those that index the proportion of variance accounted for by treatments (or by other manipulated or measured variables such as chronicity of the client's problem or clinical skill level of the therapist). Given the ubiquity of variability in outcome across individuals, treatments that account for a substantial portion of the total outcome variance would be judged noteworthy or important. Conversely, treatments that account for a minor portion of outcome variance would be viewed as inadequate or as less important.

These variance accounting or magnitude of effects statistics range from the familiar squared correlation coefficient to the more exotic generalizability coefficient (Cronbach, Gleser, Nanda, & Rajaratnam, 1972). This paper will critically examine the utility of these procedures for assessing treatment effectiveness. While the focus is on omega squared (ω^2), the most frequently discussed magnitude of effect statistic, the comments are generally applicable to the entire class of procedures.

Calculating ω^2

Traditionally, ω^2 has been used to estimate the importance of treatments in a fixed-effects design from the ratio of two variances—the variance attributable to the treatment divided by the sum of all sources of variance in the design. In determining the effectiveness of social skill training in a treatment-control post-test only design, the value of ω^2 equals the ratio of treatment variance to treatment plus error variance [$\omega^2 = \sigma^2_t / (\sigma^2_t + \sigma^2_e)$]. These variance components are obtained from formulas for the expected mean squares for the design and from the calculated values of the mean squares for treatment and for error (see Table 1). Computational illustrations of ω^2 for a variety of designs are given in Dodd and Schultz (1973), Dwyer (1974), Fleiss (1969), Freidman (1968), Gaebelein, Soderquist, and Powers (1976), Golding (1975), Hays (1963), and Vaughan and Corballis (1969).

In more complex investigations, additional variance components are included in the denominator for ω^2; furthermore, separate ω^2's can be calculated for each main and interactive effect included in the design. For example, in a study comparing systematic desensitization with no-treatment for moderate as well as severe agoraphobics, variance components would be calculated for treatments, problem severity, treatments by problem severity, and error. Separate ω^2's could be computed to assess the proportion of variance due to treatment vs. no-treatment, moderate vs. severe problem behaviors, and the interaction between these two factors.

TABLE 1

ANOVA SUMMARY TABLE EXTENDED TO INCLUDE THE INFORMATION REQUIRED TO CALCULATE ω^2 FOR A TWO-GROUP, BETWEEN SUBJECTS DESIGN (n = 10 IN EACH GROUP)

Source	*df*	*MS*	*E(MS)*	Variance Component	ω^2
Treatment	$k - 1 = 1$	85.0	$\sigma^2_e + n\sigma^2_t$	$4.0 = (MS_t - MS_e)/nk$	$.44 = 4.0/(4.0 + 5.0)$
Error	$k(n - 1) = 18$	5.0	σ^2_e	$5.0 = MS_e$	

These ω^2's would index the absolute, as well as the relative importance of these factors in accounting for treatment outcome.

Advantages of ω^2

In comparison to traditional hypothesis testing, ω^2 and related statistics have three major advantages that recommend their use to behavioral researchers. First, ω^2 indexes the *strength* of association between intervention and outcome (or between any independent and dependent variable). Second, instead of providing a dichotomous index of treatment success (significant vs. non-significant), ω^2 provides a *continuous* measure of the extent of the relationship between treatment and outcome. The values of ω^2 range from .00 (indicating that none of the variance in outcome is due to treatment) to 1.00 (indicating that all of the outcome variance is due to treatment). Third, unlike conventional hypothesis tests, ω^2 is not systematically affected by aspects of *statistical power* related to sample size. As a consequence of these properties, a small value of ω^2 would quickly expose a weak, though statistically significant treatment. In contrast, and perhaps more important for small-n behavior therapy research, a large value of ω^2 would indicate that even a statistically non-significant or borderline result represents a strong relationship between intervention and outcome. In such a case, an investigator may conclude that the treatment is worth further study, whereas the decision "non-significant" stemming from conventional statistical testing may result in discarding a promising technique. Thus, ω^2 may perform a valuable function in clarifying the meaning of both statistically significant and statistically non-significant outcomes.

Limitations of ω^2

While ω^2 has certain advantages that would appear to make it an attractive supplement to conventional hypothesis testing for evaluating behavioral treatments, the value of ω^2 is affected by a variety of factors which may make its meaning unclear. These factors, many of which also affect the values of conventional testing procedures and familiar correlational techniques, include the following: (a) the reliabilities of the treatment and outcome variables, (b) the specific design of the study, including the subject and treatment variables employed, and (c) the formula used to calcu-

late ω^2. Our discussion of these factors that limit the interpretive value of ω^2 does *not* include situations in which the statistical assumptions of ω^2 are violated. In fact, we assume that ω^2 is only calculated on data that meet basic ANOVA and hence ω^2 statistical assumptions.

Reliability. The value of ω^2 is affected by inconsistency in implementing treatment procedures and unreliability in measuring treatment outcome. These irregularities increase error variance and hence lower the value of measures of association such as ω^2. The error-free values of related magnitude of effect statistics are estimated by applying one of the correction for attenuation formulas (e.g., Cronbach et al., 1972; McNemar, 1969). These formulas permit researchers to make appropriate interpretations of magnitude of effects statistics when the reliability of the variables is known. Unfortunately, these formulas have not been applied to ω^2 to correct it for errors of measurement. If errors of implementation or measurement are substantial, the uncorrected value of ω^2 will seriously underestimate the importance of the treatment with which it is associated. If reliabilities are unknown or not corrected for, the meaning of ω^2 may be ambiguous.

Design. The second problem shared by magnitude of effect statistics is that their value may be manipulated by altering the design of the study, that is, by changing subject, treatment, or other independent variables. We will illustrate this point with two examples: First, when the range of levels of an independent variable is modified; and second, when independent variables are either added or deleted.

The value of ω^2 may be varied appreciably by altering the range of levels of the treatment variables included in the investigation (Dooling & Danks, 1974). Restricting the range of an independent variable may lower the value of ω^2 for that variable whereas using only extreme levels of the variable may raise the value of ω^2 (Glass & Hakstian, 1969). Consider an experiment on the effects of desensitization on snake avoidance. If the experiment is conducted three times, first on a sample of moderate and severe snake phobics, next on a sample of mild, moderate, and severe snake phobics, and finally on mild and severe snake phobics, then the ordering of the values for ω^2's associated with severity of snake phobia in the three experiments would be:

ω^2 (moderate vs. severe) $\leq$ ω^2 (mild vs. moderate vs. severe) $\leq$ ω^2 (mild vs. severe).

(We are assuming here that the relationship between the effect of desensitization and severity of snake phobia is linear and homoscedastic.) In the first experiment, the range of severity values is restricted, and so ω^2 for this factor is relatively small, while in the last experiment the range of values is most extreme, which results in the largest value of ω^2 for the severity factor.

The value of ω^2 can also be changed by adding or deleting independent

variables. Because ω^2 is defined as the ratio of treatment variance to total variance, the addition of any independent variable that increases total variance will produce a decrease in ω^2. Consider an investigation comparing desensitization with no treatment for severe tests anxious subjects. If this study yielded a variance component for treatments equal to 5.0 and a variance component due to error equal to 2.0, the value of ω^2 for treatments would be $5.0/(5.0 + 2.0) = .71$. If severity of test anxiety was added as another independent variable, with main and interactive variance components equal to 4.0, then the ω^2 for treatment would be reduced to $5.0/(5.0 + 2.0 + 4.0) = .45$. While the addition of independent variables can produce a decrease in ω^2, the control of variables typically present in the investigation can produce an increase in ω^2. Consider adding a variable such as the degree of relaxation achieved during the desensitization training which was uncontrolled in the original example. Assume that the portion of the error component due to the main and interaction effects of degree of relaxation was equal to 1.0 and the portion of the error component due to other uncontrolled factors was also equal to 1.0. If degree of relaxation was controlled in another study by using only subjects who could achieve some behavioral criterion of deep relaxation, then the error component would be reduced to 1.0 and the ω^2 for treatment would be $5.0/(5.0 + 1.0) = .83$. It can be seen from these examples that the value of ω^2 associated with treatments depends on whether other relevant variables are controlled by being held constant or are allowed to vary in the treatment groups. The effect of the presence or the absence of other variables on ω^2 thus adds additional ambiguity to the interpretation of this magnitude of effect statistic.

Because the magnitude of ω^2 depends on the independent variables and their range of levels included in the investigation, ω^2 cannot be safely generalized across studies in which these factors differ. In a fixed-effect design, ω^2 can only be generalized safely across experiments with identical independent variables and levels of these variables (Golding, 1975). Similarly, the results of a significant *F*-test in a fixed-effects design cannot be generalized to any investigation in which these variables and their levels differ.

Calculation Formulas. Controversy still exists over the appropriate magnitude of effect statistic to use and over the correct formulas to use in calculating ω^2 if it is the statistic of choice. Glass and Hakstian (1969) trace the history of measures of association and mention several alternatives to ω^2 including ϵ^2 (epsilon squared) (Cohen, 1965) and η^2 (eta squared) (Friedman, 1968).[1]

Dwyer (1974), in discussing formulas for calculating ω^2, argues that the

[1]According to Fleiss (1969) depending on the design used, ω^2 may under-estimate the relationship between the outcome and treatment variables, while ϵ^2 and η^2 overestimate the relationship. However, Keselman (1975) reported that ω^2 was the most accurate estimator of the population association when compared with ϵ^2 and η^2 in a Monte Carlo study.

methods outlined by Vaughan and Corballis (1969) overestimate the variance components for certain designs and hence affect the value of ω^2 involving these components. Gaebelein et al. (1976) showed that Dwyer was assuming that the interaction effect in a design with one random and one fixed effect (e.g., randomly sampled therapists and fixed treatment) must sum to zero over the fixed effect; Vaughan and Corballis did not make this assumption. Unfortunately, statisticians do not agree on which of these is the correct assumption to make in a mixed model (Kempthorne, 1968; Peng, 1967; Searle, 1971). Since the formulas advocated by Dwyer (1974) and by Vaughan and Corballis (1969) may yield substantially different values of ω^2, researchers must be careful to indicate the specific formula used in their calculations.

The preceding three factors may require caution in interpretation because of their direct effects on the value of ω^2. The following two issues—conceptual clarity of the treatment variables and the unknown characteristics of the ω^2 sampling distribution—may also affect the interpretation of ω^2.

Complexity of the Treatment Variable. The meaning of ω^2 may be unclear when the treatment variable is complex and does not form a conceptual unit (Doolings & Danks, 1974; Glass & Hakstian, 1969). To make sense as a conceptual unit, a treatment variable should vary on a single dimension. Easily defined and understood treatment variables are ones such as the number of modeling trials or the length of exposure to phobic stimuli. More complex are treatment variables or factors that include a control condition plus any number of different complex technique conditions such as desensitization and participant modeling. Such variables differ on multiple dimensions, and the meaning of ω^2 associated with treatment variables of this kind may be conceptually unclear.

Sampling Variability. On occasion, two or more ω^2's may be compared to determine which one among a number of treatment variables are of greater importance in producing outcome. In making such comparisons, investigators should keep in mind that ω^2 calculated on sample values is a statistic, not a parameter. Like other statistics, ω^2 is subject to sampling fluctuations. Unfortunately, the standard error and other characteristics of the sampling distribution for ω^2 are not yet well understood.[2] Still less is known about the sampling distribution relevant to comparing two correlated ω^2's obtained from the same investigation. The difference be-

[2]Monte Carlo studies have provided some information regarding the sampling distribution of ω^2. These studies have reported that the standard deviation of ω^2 depended upon the magnitude of experimental effects, upon whether the variate was sampled from a normal or exponential distribution (Keselman, 1975), and upon the sample size (Carroll & Nordholm, 1975). For example, with $n = 15$, ω^2's as large as .28 were found under null conditions. Thus, when the sample size does not allow for a good estimate of variance components, the power of the F-test is low and ω^2 is also poorly estimated.

tween ω^2's might be evaluated by treating the ω^2's as r^2's and applying procedures that are appropriate for testing the difference between correlation coefficients (McNemar, 1969, p. 157; Steiger, 1980). Because this approach provides only rough approximations of statistical significance when used with ω^2, conclusions regarding the relative importance of treatment variables based on the comparison of ω^2's should be made with caution.

SUMMARY

Magnitude of effects statistics such as ω^2 have clear advantages that recommend their use to applied behavioral researchers. They provide a graduated measure of the strength of the relationship between treatment and outcome that is independent of sample size. Thus, ω^2 and related statistics are important supplements to conventional hypothesis testing procedures for evaluating the importance of treatments. However, these magnitude of effects statistics have limitations that require behavioral researchers to exercise considerable caution when employing them. The value of ω^2 is subject to attenuation resulting from lack of reliability in either the treatment or outcome variables. The selection of treatment factors and levels within these factors which occur in fixed-effects designs makes ω^2 a poor estimate of the generalized degree of association. There are also disagreements among statisticians about the appropriate statistic to use to estimate magnitude of effects and about the correct formulas to use in calculating ω^2. The poorly defined variables sometimes found in treatment research makes interpretation of ω^2 difficult. The sampling distribution of ω^2 has not been identified, so the statistic cannot be used inferentially. The number and seriousness of these problems forces us to conclude that the *uncritical* use of magnitude of effects statistics as a cure for the problems of conventional hypothesis testing methods of assessing treatment effectiveness may very well represent a remedy as troublesome as the original problems.

REFERENCES

Carroll, R. M., & Nordholm, L. A. Some characteristics of Kelley's ϵ^2 and Hays' ω^2. *Educational and Psychological Measurement*, 1975, **35**, 541–554.

Cohen, J. Some statistical issues in psychological research. In B. B. Wolman (Ed.), *Handbook of clinical psychology*. New York: McGraw-Hill. 1965.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley, 1972.

Dodd, D. H., & Schultz, R. F., Jr. Computational procedures for estimating magnitude of effects for some analysis of variance designs. *Psychological Bulletin*, 1973, **79**, 391–395.

Dooling, J. D., & Danks, J. H. Going beyond tests of significance: Is psychology ready? *Proceedings of the Psychonomic Society*, 1974, **53**, 15–17.

Dwyer, J. H. Analysis of variance and the magnitude of effect: A general approach. *Psychological Bulletin*, 1974, **81**, 731–737.

Fleiss, J. L. Estimating the magnitude of experimental effects. *Psychological Bulletin*, 1969, **72**, 273–276.

Friedman, H. Magnitude of experimental effects and a table for its rapid estimation. *Psychological Bulletin*, 1968, **70**, 245–251.

Gaebelein, J. W., Soderquist, J. A., & Powers, W. A. A note on the variance explained in the mixed analysis of variance model. *Psychological Bulletin*, 1976, **83**, 1110–1112.

Glass, G. V., & Hakstian, A. R. Measures of association in comparative experiments: Their development and interpretation. *American Educational Research Journal*, 1969, **6**, 403–414.

Golding, S. L. Flies in the ointment: Methodological problems in the analysis of the percentage of variance due to persons and situations. *Psychological Bulletin*, 1975, **82**, 278–289.

Hartmann, D. P., Roper, B. L., & Gelfand, D. M. An evaluation of alternative modes of child psychology. In B. L. Lahey & A. E. Kazdin (Eds.), *Advances in clinical child psychology* (Vol. 1). New York: Plenum, 1977.

Hays, W. L. *Statistics for psychologists*. New York: Holt, Rinehart & Winston, 1963.

Kazdin, A. E., & Wilson, G. T. *Evaluation of behavior therapy*. Cambridge, Mass.: Ballinger, 1978.

Kazdin, A. E. *Research design in clinical psychology*. New York: Harper & Row, 1980.

Kempthorne, O. Discussion of Searle's paper. *Biometrics*, 1968, **24**, 782–784.

Keselman, H. J. A Monte Carlo investigation of three estimates of treatment magnitude: Epsilon squared, eta squared, and omega squared. *Canadian Psychological Review*, 1975, **16**, 44–48.

Mahoney, M. J. Instructions to contributors. *Cognitive Therapy and Research*, 1977, **1**, unpaged (inside back cover).

McNemar, Q. *Psychological statistics* (4th ed.). New York: Wiley, 1969.

Morrison, D. E., & Henkel, R. E. (Eds.). *The significance test controversy—A reader*. Chicago: Aldine, 1970.

Peng, K. C. *The design and analysis of scientific experiments*. New York: Addison-Wesley, 1967.

Searle, S. R. *Linear models*. New York: Wiley, 1971.

Steiger, J. H. Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, 1980, **87**, 245–251.

Vaughan, G. M., & Corballis, M. C. Beyond tests of significance: Estimating strength of effects in selected ANOVA designs. *Psychological Bulletin*, 1969, **79**, 391–395.

RECEIVED: 3-24-80 REVISED: 8-28-80
FINAL ACCEPTANCE: 8-30-80.