How the magnitude of an experimental effect may be measured has been a matter of concern for at least two decades. The phenomenon of effect size is still not well understood, and it cannot be inferred from statistical significance. In recent years various ways of assessing the amount of variance accounted for have been proposed as measures of magnitude of effect. Other writers have proposed rules for standardizing effect size, with the interpretations of the measures depending largely on intuitions buttressed by some further general empirical norms. All the methods of assessing effect size have serious flaws that limit their usefulness. The various statistical procedures for estimating variance accounted for are based on different statistical models and can produce rather sharply differing results, depending on the model employed. All the methods suffer from the limitation that they reflect to too great an extent the particular characteristics of the study being reported and hence have limited generalizability.

MAGNITUDES OF EXPERIMENTAL EFFECTS IN SOCIAL SCIENCE RESEARCH

LEE SECHREST
WILLIAM H. YEATON
University of Michigan, Ann Arbor

A problem of which researching psychologists have been aware for years (e.g., Bolles and Messick, 1958; Savage, 1957) but that has had increasing attention over the past decade or so is how to determine how large an effect is achieved 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. However, even for theoretical problems it is at least enlightening and often sobering to find out how much an effect is at stake during the intricacies of a theoretical controversy. For example,

---

AUTHORS' NOTE: *Preparation of this article was supported by Grant Number 1 R01 HS02702 from the National Center for Health Services Research. The authors wish to thank Jacob Cohen, John Cotton, Ronald Gallimore, David Kenny, and Stephen West*

EVALUATION REVIEW, Vol. 6 No. 5, October 1982 579-600

0193-841X/82/050579-22$2.45

Smith (1980) found a .25 standard deviation sex bias effect in published studies in counseling and psychotherapy, a finding that accounts for less than 2% of the variance in results. Unfortunately, a relationship of that magnitude is likely to have little, if any, practical significance.[1]

A number of suggestions about ways of estimating the magnitude of an experimental effect have been proposed, and clearly some of them are useful—or at least better than relying solely on statistical significance as a criterion. Yet we believe that there are serious shortcomings with existing approaches and that alternatives need to be invented and investigated. In our view, the problems are of sufficient complexity that no one solution will suffice, and multiple approaches will be required.

STATISTICAL SIGNIFICANCE

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 a fairly direct relationship between the statistical significance of a finding and its importance in the real world.[2] 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, and even six decimal places, as if something of critical importance were contained in those several zeros preceding the final digit.

Statistical significance in fact depends on several factors quite unrelated to the magnitude of the experimental effect. First, *whether* a finding is statistically significant depends on the alpha level one sets (traditionally, .05). The statistical significance of a finding is also a function of sample size, although the relationship of significance level and sample size is nonlinear, being in general a function of $\sqrt{N}$. Consequently, studies with larger numbers of subjects yield smaller p values for equal experimental effects, though sample size tends to be inversely related to the percentage of variance accounted for (Craig et al., 1976). It is also the case that whether one obtains a significant effect at all is a function of the level chosen for beta, the arbitrarily defined probability

---

for their helpful comments on earlier drafts. An extended version entitled "Estimating Magnitudes of Experimental Effects" can be obtained from Journal Supplements Abstract Service, *1981, manuscript 2355. Requests for reprints should be made to Dr. Lee Sechrest, Center for Research on the Utilization of Scientific Knowledge, The Institute for Social Research, University of Michigan, Ann Arbor, MI 48106.*

of Type II errors and the resulting power ($1 - \beta$) of a particular experiment to reject the null hypothesis.

It is, of course, true that statistical significance is very much a function of effect size. All other things kept equal, a larger difference between means will be associated with a lower p value.[3] Again, however, the relationship is nonlinear, and it is hazardous to make comparisons of significance levels across experiments or even across treatments within an experiment. To take the latter, a less obvious case, if in comparison to a control group one treatment is significant at the .01 level while another treatment is significant at only the .05 level, it is tempting to conclude that the former is a stronger treatment. For that to be a legitimate conclusion, the difference between means would have to be larger and the error term the same. And to prove true generalizability, one would have to defend the proposition that the two treatments were implemented with equal care and precision, a point to which we will return.

WHAT IS AN EXPERIMENTAL EFFECT?

Perhaps it would be desirable at this point to clarify what we mean by an experimental effect and also to indicate how our arguments may be extended to nonexperimental research findings. In experimental research an effect is usually reflected in the difference(s) between measures of central tendency for different experimental groups, or at least that is the simplest case. Thus, in a simple two-group (E and C) design with posttest measures only, magnitude of the experimental effect is achieved by: Effmag = mean X_E - mean X_C. If the data are categorical and are cast in the form of a contingency table analysis, say for volunteer donors and nondonors within experimental and control groups, the experimental effect is the excess of observed versus expected donors in the experimental as compared to the control condition. If the data are analyzed by regression analysis, the experimental effect is the regression coefficient associated with an E versus C dummy variable. When we speak of experimental effects, we have in mind a raw estimate or value that cannot, until processed in some kind of index, be compared across studies, or even across different main effects within the same experiment. To be concrete, the effect produced by an analgesic is the difference in measured headache distress between the experimental and placebo drug groups. If, on the average, control subjects report headaches of 68 on a 100-point scale and

experimentals report headaches of only 53, then the experimental effect is 68 - 53 = 15 points of pain reduction.

Extension of the above ideas to subject variables is straightforward. The "effect" of sex of subject on a dependent variable is the difference between the means for male and female subjects. However, it is also possible to think about correlational results within the same framework. A correlation between two variables is the "effect" of one on the other, and the difference between two correlations enables one to compare two different effects—for example, to answer the question whether the relationship of family size to intelligence is greater than the relationship of socioeconomic status to intelligence.

WHAT IS THE PROBLEM IN ESTIMATING EFFECT SIZE?

Having just defined what is meant by an effect, we can now turn to the problem of trying to see how big an effect is, of estimating the magnitude of an experimental effect. The problem is not a simple one, and it has generally been either ignored or treated as if it were simple. Some illustration may help.

1. Suppose a group of children exposed to an early childhood educational enrichment program show a five-point superiority over a control group on a 100-item achievement test. The experimental effect is five points, but how much are those five points worth? Is the program really a good one? Should congressmen vote money to implement it on a nationwide basis? The critical questions cannot be answered.

2. An investigator finds that there is a correlation of .32 between income and utilization of the services of mental health specialists. Is that correlation grounds for supposing that persons with low incomes are being seriously shortchanged in their access to mental health services? Is it a large enough relationship to demand action, or even further study? There is no way of saying—at least not right now.

3. How many lives would an experimental medical unit have to save to be considered impressive? A difference of one life could never be statistically significant, but would five lives or twenty lives be grounds for elation over an operation period of one year? Probably it all depends, but how much and on what (see Rhoads, 1978)?

The foregoing examples are meant to convey the sense of uncertainty, often bordering on the absurd, that must afflict any insightful and honest investigator if asked what his or her results really mean in a

practical sense. We believe that the same sense of uncertainty should often afflict even the investigator of a theoretical problem, but it is characteristic of theoretical research in psychology to ignore the issues of significance in other than a statistical sense. Because we do not wish to single out any particular investigation, we do not provide specific references here, but as an example note that one theoretical investigation with statistically significant results hung on the differences in ratings of objects on a five-point scale, with obtained differences being .22 and .25 points for two effects! One scarcely knows whether to conclude that the theory is woefully weak to produce such small differences or impressively powerful to be able to predict them.

A prevalent concept of effect size involves the notion of "accounting for" or "explaining" variance. Presumably a large experimental effect is reflected in a large index, of whatever nature, of variance accounted for.

WHAT DOES IT MEAN TO "ACCOUNT FOR" VARIANCE?

Nearly all statistics books are vague at best when it comes to explaining what is meant by "accounting for" variance. We have even encountered such circular explanations as that to account for variance means to explain it, with explaining it meaning to be able to account for it. Cohen and Cohen (1975), however, present a clear, understandable exposition of what is entailed in accounting for variance. Space does not permit elaboration here (see Sechrest and Yeaton, 1981c, for a detailed discussion), but put simply, accounting for variance means that one is able to reduce the variance in one's errors of predictions of scores by applying some knowledge more precise than that a person is a member of a population. In the experimental paradigm, knowing merely that a person was a subject in an experiment enables no better prediction of his or her standing on a dependent variable than the overall mean for the sample. However, if an experimental treatment accounts for some of the total variance, that means that by knowing which treatment condition a subject was in, a better prediction, the mean for the condition, can be made, and variance of errors of prediction will be reduced from the variance of the total sample to some lower value.

It will be noted that accounting for 25% or so of the variance in some scores does not make a great deal of difference in the standard deviation

of errors of prediction. Let us suppose that a population exists for which the mean intelligence score is 100, and let us suppose that the standard deviation of those scores is 15, which results in a variance of 225. Accounting for 25% of the variance reduces the standard deviation of errors of predicted intelligence from only 15 to 13, a feat more unimpressive than usually imagined.

ESTIMATING SIZE OF EFFECTS

The problem of estimating the magnitude of an experimental effect as we noted, has been recognized for some time, although not so widely that it has achieved any degree of prominence in the research literature (e.g., Soderquist and Hussian, 1978). However, as early as 1935, Kelly developed the correlation measure ϵ^2 (epsilon squared) which could be used to estimate effect size, though it was Bolles and Messick (1958) who made one of the earliest attempts to deal with the issue of estimating substantive significance, proposing the coefficient of utility U for use in statistically assessing the usefulness of specific experimental variables. Underlying the rule-of-thumb approaches soon to be discussed is the notion that some index of effect size can be devised for which useful, if arbitrary, comparisons may be made of the results of different experiments. We emphasize in advance, however, the arbitrary nature of the rules, for none of them speaks in a direct way to the issue of social or practical importance of findings, only to relative size of effects.

COHEN'S RULE OF THUMB

As much as any other person Cohen has been responsible for bringing the issue of effect size to the attention of the social science community (Cohen, 1977). In an article published in 1962, Cohen analyzed research reports in a complete volume of the *Journal of Abnormal and Social Psychology* in an attempt to determine the statistical power with which each analysis might confront the null hypothesis. Since this seminal article, several researchers have conducted power analyses in other disciplines (e.g., Brewer [1972] in education, Katzer and Sodt [1973] and Chase and Tucker [1975] in communications research, and Chase and Chase [1976] in applied psychological research). In Cohen's article

it was necessary for him to set an effect size in order to estimate power. Cohen simply stated that some effect sizes represented "small," "medium," and "large" effects and confessed that the values he chose were arbitrary but seemed "reasonable," urging readers to render their own judgment on the matter.

Cohen wished to establish a set of metrics for effect sizes that would make it possible to compare effect sizes across experiments. Moreover, he aimed to make it possible to compare effect sizes across different statistics so that, for example, one could estimate whether an experimental result estimated by a chi-square test for difference between proportions is less or greater than a result estimated by a t-test for difference between means. Cohen's classification of effect sizes for difference between means as small, medium, and large was based on the ratio $(M_1 - M_2)/\sigma$. The specific values he chose were .25, .50, and 1.00 for small, medium, and large, respectively, though more recently, Cohen (1977) has reduced the initial ratios to .20, .50, and .80. These values for effect sizes are now being cited with some frequency in the literature, despite the fact that they have no compelling rationale other than that they seemed like a good idea at the time.

We note that what Cohen means by "small" effect is likely to be *really* a small effect size in any practical sense. A difference between means of only $.2\sigma$ represents about 1% of the variance in the dependent variable, and even a "large" effect of $.8\sigma$ represents only about 14% of the variance in the dependent variable. To put it another way, Cohen's small effect size would be reflected in a correlation of only .10 and a large effect in a correlation of only .37. Those are limited aspirations, indeed. Somewhat astoundingly, however, Cohen (1973b) notes that researchers are often implicitly testing for effect sizes *smaller* than what he has defined as small!

The big advantage of Cohen's rule of thumb is that effect size, by his procedure, is standardized and hence independent of specific population or sample values. One can compare the relative effects of manipulations as diverse as psychotherapy and demand characteristics on dependent variables as diverse as IQ and reduction in cigarette smoking. Since Cohen's rule of thumb is standardized, one can, for purposes of statistical power analysis, state an anticipated effect size and do power analysis without the necessity for estimating population variance that would otherwise be required.

Glass and his colleagues (Glass et al., 1981; Smith and Glass, 1977; Smith et al., 1980) have provided a practical application of Cohen's

method of standardizing results in their meta-analyses of psychotherapy and drug outcome studies. They converted relevant effects in several hundreds of studies to standard deviation units; that is, effects were expressed in terms of a fraction of a standard deviation of difference between experimental and control groups. From these quantitative syntheses they found, for example, that systematic desensitization results in an average effect of .48σ. The meta-analytic approach, however, does not address the fundamental question of the value of the resulting effect. We are still uncertain whether .75σ or .40σ or any other fraction of a standard deviation difference in a dependent measure is worth the money and effort required to produce it. What can a group of children whose math scores are at the fiftieth percentile do that a group whose scores are at the fortieth percentile cannot do?

FRIEDMAN'S r_m

Friedman (1968) attempted to establish a single generalizable index of magnitude of experimental effect by expressing the relationship between a statistical measure such as t, F, or χ^2 and sample size as a correlation. Beyond expressing the notion of effect size in a general form, Friedman's contribution is a table making possible the quick estimation of effect size. One merely needs values for an inferential statistic and for sample size to enter the table, and r_m may be read directly. To take but one example, a t of 2.50 with a sample size of 60 produces an r_m of about .31, indicating that a little less than 10% of the variance is accounted for.

ω^2 AND RELATED STATISTICS

Largely with impetus from Hays (1963, 1973) a statistic he named ω^2 (omega squared) has come into use in estimating proportions of variance accounted for in experiments involving parametric tests of differences between means—t and ANOVA. Hays noted that ω^2 is a population value to be estimated from sample data. The formulae for estimating ω^2 are considered to produce biased estimates in unknown degree. Hays also notes that ω^2 applies to ANOVA models with fixed effects. The formula for ω^2 will vary according to the specific design that is involved, but for a simple one-way ANOVA it is:

$$\text{est } \omega^2 = [SS_{bet} - (J - 1)MS_{with}] / [SS_{tot} + MS_{with}]$$

Hays also discusses η^2 (eta squared), a sample statistic useful for descriptive purposes within any one experiment and interpreted in the same way as ω^2. Actually, η^2 has traditionally been used to quantify curvilinear relationships (Peters and Van Voorhis, 1940). Its use to estimate proportion of variance accounted for by an experimental treatment (Cohen, 1965) is a direct extension of its capacity to express the relationship between variables not necessarily either ordered or linear in magnitude. Since η is a correlation ratio, η^2 is interpretable as proportion of variance in one variable accounted for by the other. Hays notes that since η^2 is a sample statistic, it is subject to capitalizing on chance and usually gives a larger estimate accounted for than does ω^2. The computational formula for η^2 is: $df_N(F)/[df_N(F) + df_D]$.

Hays describes an additional statistic for estimating proportion of variance accounted for, ρ_I (rho), the intraclass correlation. According to Hays, ρ_I^2 provides an estimate of variance accounted for in analyses involving a random effects model. It is also a population parameter.

Although Hays asserted that ω^2 is applicable only for analyses up to the two-way ANOVA, Fleiss (1969), Vaughn and Corballis (1969), Halderson and Glasnapp (1972), and Dodd and Schultz (1973) have extended the rationale and computation to include random and mixed models and more complex designs. Halderson and Glasnapp (1972) give generalized rules for estimating magnitudes of effects in factorial and repeated measures ANOVA designs. Refinement in the methodology, notably in the estimation of interaction terms in mixed models (Dwyer, 1974), and attention to assumptions underlying the model (Gaebelein and Soderquist, 1976) have subsequently been proposed.

COMPARISONS AMONG EFFECT SIZE ESTIMATES

Considerable energy has been expended to develop a set of guidelines to assist researchers in the choice of an appropriate statistical analysis of outcomes. Witness the number of statistical analysis and design textbooks available to students and faculty (e.g., Cochran and Cox, 1957; Kirk, 1968; Myers, 1966; Winer, 1962). With the exception of a few Monte Carlo studies (e.g., Carroll and Nordholm, 1975; Keselman, 1975), however, little comparable energy has been invested with estimates of effect size. The choice among these estimates is rather arbitrary.

An obvious consideration in the choice of an estimate is the knowledge of exactly what quantity is being estimated. As we noted, the

terminology "percentage of variance accounted for" is deficient in important ways as a descriptor of what is being estimated by effect size indicators. Furthermore, the computational formulae offer little intuition as to the actual quantities being estimated. This has the effect of inhibiting comparison, since we cannot ascertain if we are estimating fundamentally different quantities. Knowing that ω^2 is a population parameter and that partial η^2 is a sample statistic does make strict comparison impossible, though some "feeling" for the differences between these two estimates is desirable.

Further confusion is added when we learn that η^2 has been referred to in different ways by different researchers. Kennedy (1970) noted that Kerlinger (1964) defined $\eta_x^2 = SS_x/SS_{Tot}$ and claimed that Cohen (1965) and Friedman (1968) in their previous research had defined η_x^2 as $df_N(F_x)/[df_N(F_x) + df_D]$. However, Cohen (1973a) subsequently corrected Kennedy's use of the terminology eta squared for $\eta_x^2 = df_N(F_x)/[df_N(F_x) + df_D]$, recognizing this as a formula for partial eta squared. The confusion between eta squared and partial eta squared was alleviated considerably by Kennedy, who showed by algebraic simplication that partial $\eta_x^2 = SS_x/(SS_x + SS_e)$, thus making obvious the fact that the difference between the two eta square estimates is in the denominator of the two estimates; SS_{Tot} will change when any of the sources of variation in an experiment change, and the number of these sources will increase as the complexity of the experiment increases. However, $SS_x + SS_e$ only varies as a function of one additional source of variation, namely SS_e. Though there is no difference between these two estimates in the one-way ANOVA since $SS_{Tot} = SS_x + SS_e$, eta squared and partial eta squared will almost always differ in higher order ANOVA's. Even partial eta squares for different sources of variation are not comparable when they have different bases (denominators) and cannot legitimately be added together to obtain a total percentage of variance accounted for (Cohen, 1973a).

We wondered immediately if ω^2 was analogous to η^2 or to partial η^2. To answer this question, we consulted Table 1 in Vaughn and Corballis (1969), which gives variance components for fixed, mixed, and random designs in one- and two-way ANOVAs. Since $\omega^2 = \hat{\sigma}_x/\hat{\sigma}_{tot}^2$ for the one- and two-way ANOVA, ω^2 is analogous to η^2, since both denominators are expressed in terms of total variation. Additionally, it is apparent that Hays's ω^2 could be considered the fixed model case of the general components of variance approach. Previously we had wondered *why* ω^2 was relevant to fixed models and ρ_I, the intraclass

correlation, to random models as stated by Hays (1963). The reason is simply that ω^2 (an arbitrary symbol) refers to the fixed model case from which its computational formula is derived, while ρ_I is comparable to ω^2 except that it is the symbol chosen to designate the computational formula taken from the same components of variance approach when the model is random. A third symbol could just as easily have been chosen for those formulae derived from the components of variance approach when the underlying model is mixed.

To summarize briefly, both ω^2 and ρ_I can be considered special cases of the components of variance approach. And though it is a population parameter, ω^2 is more analogous to η^2, since its denominator (σ^2_{Tot}) is more similar to the denominator in η^2 (SS_{Tot}) than to the denominator in the partial η^2 ($SS_x + SS_e$).

One means of clarifying the important points made in this section on the comparison of effect sizes is to illustrate the extent of differences among effect size estimates with specific examples taken from the literature. Table 1 shows several effect size estimates calculated from data taken from Byrne and Rhamey (1965). η^2 and ω^2 are very comparable, as are partial η^2 and r^2_m. However, these two separate sets of estimates are discrepant. Since partial η^2 is based on a denominator using only source and error SS, sums of squares based on other main effects and interactions are not used in the calculation as they would be in effect size estimates based on total variation (η^2 and ω^2). Consequently, partial η^2 will be substantially larger than ω^2 when any other sources (main effects or interactions) account for considerable variance, as is the case in Byrne and Rhamey. Only when SS_{Tot} approximates $SS_x + SS_e$ (i.e., other sources of variance are close to zero) will these measures be comparable. Also obvious from Table 1 is that partial η^2_E, partial η^2_A, and partial η^2_{EXA} sum to more than 100%. This is also true of r^2_m, though this is not true of η^2 or ω^2. Since effect sizes are typically of small absolute magnitude, the undesirable feature of accounting for more than 100% variance would not likely be discovered by researchers.

Effect size estimates in Table 1 (taken from Vitalo, 1970) allow comparisons in the one-way ANOVA as well as this two-way case. Here, the proportion of variance accounted for by sources is smaller than proportions in the Byrne and Rhamey study. η^2, partial η^2, and r^2_m are indeed equal in the one-way ANOVA. Though ω^2 is more similar to η^2 than to partial η^2 and r^2_m, such a population parameter is not likely to approximate closely sample statistics when the sample size is

TABLE 1
Comparison of Effect Size Estimates in Three Different Studies*

Source of Variation	$\eta^2 = \frac{SS_x}{SS^x_{Tot}}$	$\omega^2 =$	*Partial* $\eta^2 = r^2_m$
	Percentage of Variance Accounted for		
I. Evaluation (E)	41.0	40.6	56.5
Attitudes (A)	23.1	22.5	42.3
E × A	4.3	3.1	11.9
II. (a) (1-way ANOVA)			
Conditioning	28.0	20.8	28.0
(b) (2-way ANOVA, between Ss)			
Conditioning (C)	.2	0.0	0.0
Interviewing (I)	5.9	2.2	6.4
C × I	7.4	3.7	7.9
III. (From a 3-way ANOVA)			
Achievement (A)	57.4	57.6	71.6
Company policy (C)	16.8	17.0	42.7
A × C	0.8	1.2	4.9

*Byrne and Rhamey, 1965 (I); Vitalo, 1970 (II); Lindsay et al., 1967 (III).

small, as is the case in Vitalo. Effect size estimates of sources in the two-way ANOVA are generally comparable, since other sources than those being tested account for small proportions of the total variance. Again, ω^2 is more discrepant from η^2 than in the Byrne and Rhamey data due to the smaller sample size.

Another interesting comparison among effect size estimates can be made by choosing published studies which have reported effect size estimates. For example, in Lindsay, Marks, and Gorlow (1967) the η^2 effect size estimate closely parallels respective ω^2 values due to the large sample size. However, the existence of main effects, which account for substantial portions of the total variance, causes ω^2 and partial η^2 to be discrepant.

Hard and fast decision rules regarding choice among effect size estimates are difficult to produce and perhaps undesirable. However, ω^2 appears to be the logical choice if the researcher wishes to be conservative in statements regarding percentage of total variance accounted

for. The fact that partial η^2 values summed across all the sources of variation in an experiment may total more than 100% should be considered a weakness of this statistic. It is also a bit uncomfortable to work with percentages that do not share the same base and that do not use 100% as a standard, even when the sum does not surpass 100%. However, given the typical research scenario in which the total SSs is made up largely of error variability, while other sources of variation contribute little to the total, the choice of an appropriate effect size estimate may be based on more practical considerations, such as computational ease. The best practice may be to report two or more estimates and let the reader judge the effectiveness of the results reported in the experiment.

HOW MUCH VARIANCE IS THERE TO BE EXPLAINED?

It seems generally and naively to be assumed by those who favor calculations of proportion of variance explained that the actual variance to be explained is 100%. That assumption is unwarranted, since it requires the additional assumption that the dependent measure is measured without error. For the most part, investigators seem conceptually to deal with total variance as if it consisted of two parts: that variance accounted for by the experimental factors and a residual part commonly called "error." Actually the total variance is better regarded as "partitionable" three ways: variance explained by the experimental factors, reliable variance not accounted for by experimental factors, and error, or unreliable variance. By definition, unreliable variance cannot be accounted for.

Consider, for example, a study of a helicopter patrol strategy for decreasing the incidence of specific crimes (e.g., Schnelle et al., 1977). Presumably, the number of crimes occurring during helicopter patrolling would be subject to errors reflected in a host of reliable factors not accounted for by the experimental manipulation (number of criminals in the area, time of year, unemployment rate, etc.). Consequently, what variance could even in principle be explained by the helicopter patrol intervention would be total variance minus the error variance. If the patrol strategy manipulation accounted for 20% of the total variance, when there was only 40% reliable variance, the seeming importance of

the experimental factors might be small even though, from the more insightful position where total reliable variance is known, the importance of experimental factors would be greatly enhanced.

The argument presented here is reflected in the psychometric relationship between reliability and validity of a measure, it being the case that the maximum achievable validity of measure is limited to $\sqrt{r_{tt}}$. Thus, if a measure has a reliability coefficient[4] of .81, the maximum validity coefficient that could be associated with a predictor of that measure would be .90, but it is the reliability coefficient, .81 (i.e., $.90^2$), that indicates the reliable variance to be accounted for. Therefore, if a dependent measure in an experiment had a reliability of .81, rather than estimating proportion of variance accounted for by an experimental variable against a maximum of 100%, the estimate should be done against the base of 81%. A variable that accounted for 20% of the total variance in such a situation would account for 25% of the reliable variance.

WHAT DETERMINES OUR ABILITY TO ACCOUNT FOR VARIANCE?

Now that the concept of accounting for variance has been explained in detail, it remains to be explained what determines variance accounted for. 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 were 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 variance accounted for within any experiment, and there are only inexact ways of knowing about or estimating the importance of those determinants.

The problem of interpreting a measure of variance accounted for begins with the fact that all such measures are essentially ratios of variance within some treatment to a more inclusive variance estimate ranging from treatment plus error up to total experimental variance. The fact that a ratio is involved should suggest immediately that estimates of variance accounted for might be unstable, since small changes in the denominator may well change estimates drastically due to decisions made about how an experiment will be carried out. (For a more extended discussion of the following factors that determine our ability to account for variance, see Sechrest and Yeaton, 1981c.)

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 (Glass and Hakstian, 1969). It should be easier to account for a lot of variance in the self-esteem scores of 15-year-old male delinquents in two experimental conditions than to account for the same proportion of variance in scores of two groups of delinquents whose only commonality is that they all live in the same county. Failure to replicate otherwise consistent results may be explained by the heterogeneity of the subject sample used in the study (e.g., Oakes, 1972).

EXPERIMENTAL PRECISION

Another determinant of total variance in an experiment is the precision achieved in planning and the integrity maintained in implementing the experiment (Sechrest et al., 1979; Yeaton and Sechrest, 1981a). Consider, for example, the almost certain difference between otherwise identical experiments when one of them involves only a single, 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, and the apparent experimental effect would be smaller. Note, however, that there is no necessary effect on means of the experimental groups; hence, subtracting one mean from another might well suggest the same effect in the two experiments. There are many sources of imprecision that might cause two experiments to differ even if the same experimental treatment is being employed. Degree of standardization of experimenter demand, clarity of instructions, calibration of apparatus, degree of control achieved with respect to strength of 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.

An interesting instance is provided by two experiments (Brady et al., 1976; Vitalo, 1970) involving the same experimental treatment and a generally serious attempt at replication. Brady et al. state that "the only known deviation from Vitalo's (1970) study is that the number of subjects was increased from 28 to 32."[5] A critical difference in the results

of the two studies was that Vitalo reported an F of 13.10 ($p < .005$) for an experimenters x conditions interaction, while Brady et al. obtained an F of only 1.54 (n.s.). The problem becomes clear when the SS for error is examined, for it is 2.40 in the first study and 19.88 in the second. (All other SSs in the source table were comparable.) For whatever reason, and despite their seemingly careful attempt to replicate Vitalo's experiment, Brady, Rowe, and Smouse produced a considerably larger amount of unexplained variation in the within-subjects part of their experiment.

Even if one wanted to compare the variance accounted for by two treatments within the same experiment, it is important to recognize that they may contribute differentially to error variance. Consider an experiment in which a drug and a behavioral intervention are to be jointly tested. It may be possible to achieve more control over the 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 behavior 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. 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 is being simultaneously studied (Kennedy, 1970).

The various indices of variance accounted for utilize different denominators and hence are differentially susceptible to the effects of multiple factors in experiments. Specifically, ω^2 uses an estimate of total variance in the denominator, while η^2 and r_m use source plus error. Therefore, ω^2 will always be smaller than the other indices in multifactor experiments and probably is the index to be preferred.

STRENGTH OF TREATMENTS

Of great theoretical and practical interest is the fact that proportion of variance accounted for obviously depends on the strength of the

experimental treatment (see Sechrest and Redner, 1979; Sechrest et al., 1979; Yeaton and Redner, forthcoming). A weak treatment could account for only a small proportion of the variance in most experiments, while a strong treatment could account for a large proportion. The problem in interpreting proportion of variance accounted for is that we rarely—at least in the social and behavioral sciences—have an independent measure of the strength of the treatment administered. For example, suppose one wished to know whether attitudinal similarity or physical attractiveness is a stronger determinant of interpersonal attraction. One could probably do little better than merely to describe the manipulations used to produce the levels of each factor and conclude that for the levels tested one or the other factor seemed to account for more variance in interpersonal attraction. For the more important theoretical question of which is generally more important, no statement can be made. There is no common metric for the two variables, so one cannot say how much physical attractiveness is equal to how much attitudinal similarity; consequently one could not say whether the treatments were of even approximately equal strength.

In only a few cases do experimenters attempt to determine the strength of a 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 only be taken literally. To show, for example, that experimental and control groups differ as they should on a seven-point rating scale gives no 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 makes little sense to talk about the proportion of variance it accounts for.

RANGE OF TREATMENTS

Still another limitation on interpretations of proportion of variance accounted for is that for any treatment involving more than two levels of an estimate of proportion of variance accounted for, far more can be obscured than revealed. 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 pressures might be accounted for by the treatment effect. If, however, one added an untreated control group to the experiment, the

treatment effect might, with seeming magic, be doubled. Glass and Hakstian (1969) have addressed this problem and note that it had previously been discussed by Sir Ronald Fisher (1946). A particularly apt example, however, has been provided by Levin (1967). He described an experiment with six experimental conditions analyzed by a one-way ANOVA, with the result that ω^2 was 37%. However, subsequent analyses 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 interpreting proportion of variance accounted for has to do with its "external validity" that is, its relationship to any "real-world" context in which one might want to draw inferences about the probable effect of some intervention. The variance that exists within an experiment depends largely on how the experimenter plans and implements the experiment. In the laboratory, when an experimenter studies the probability of a guilty verdict as a function of the physical attractiveness of a defendant, all other potential sources of variance in the determination of the verdict are controlled out of the experiment to as great an extent as possible, thus reducing the error term (unexplained variance) to a value below that likely to exist in the extraexperimental context. We are not arguing that physical attractiveness has no effect outside the social psychology laboratory; but we do argue that the fact that physical attractiveness can be made to affect responses in the laboratory in some degree does not mean that physical attractiveness has the same effect, let alone to the same degree, in the extraexperimental world.

THE SEARCH FOR EFFECTIVENESS CRITERIA

It appears to us that no purely statistical method for assessing magnitude of experimental effects is going to be satisfactory, at least if one leaves the fairly abstract world of theory building and enters into the realm of practical decision making. On the other hand, it is clearly not going to be satisfactory to continue as if all significant effects were important or to rely on haphazard or intuitive judgments. There appear to be no simple solutions for a whole variety of reasons, some of which would be remediable by changes in editorial policies and in ways in

which investigators report their findings. At present it does not appear to us likely that any single procedure or set of rules will soon emerge. What is more likely is that the demands and customs prevalent in different research areas will result in differing opportunities to develop empirical rules for estimation of effect size. Some rules are likely to involve a degree of arbitrariness and good judgment, while others will probably be normative at some level.

We have begun to explore several alternative methods and to assess the interrelation of these approaches (Sechrest and Yeaton, 1981a, 1981b; Yeaton and Sechrest, 1981a, 1981b). These initial efforts may provide the first steps toward the development of a set of useful tools for thinking about the outcome of experiments. The ability of these methods to discriminate between large and small experimental effects may be reflected in the acceptability of the procedures to investigators and decision makers. That these methods be impressive enough for acceptance is a telling test of our success in estimating magnitudes of experimental effects.

NOTES

1. However, see Sechrest and Yeaton (1981a) for an explanation of why small differences at the means of two distributions may in some circumstances be important at the extremes.

2. A particularly cogent treatment of the test of significance and problems in its interpretation was provided some years ago by Bakan (1966), in a paper still worth reading.

3. Meehl (1967) has noted the paradox in the differences between the approaches and methods of physics and psychology: The better the methods employed in physics, the greater the probability that the experiment will *disprove* the hypothesis, while in psychology hypotheses are stated in terms of deviations from the null.

4. Space limitations prevent further explanation, but we note that it obviously makes a great deal of difference which reliability coefficient one chooses to estimate variance to be accounted for. Cronbach et al. (1972) present a particularly cogent discussion of the issues involved, and their work should be consulted.

5. Tversky and Kahneman (1971) have demonstrated that attempts to replicate experimental findings are quite likely to fail unless the replication experiment has a substantially larger N with the resulting increase in statistical power. Brady et al. were on the right track in increasing sample size but did not go far enough. In order to know whether they did or did not replicate Vitalo's findings, it would be necessary to have a table of means as well as an ANOVA table, since the direction of results could have been replicated even though statistical significance was not achieved. Unfortunately, Brady et al. did not give a table of means, nor do they report directions of effects in their text.

REFERENCES

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

BOLLES, R. and S. MESSICK (1958) "Statistical utility in experimental inference." Psych. Reports 4: 223-227.

BRADY, D., W. ROWE, and A. D. SMOUSE (1976) "Facilitative level and verbal conditioning: a replication." J. of Counseling Psychology 23: 78-80.

BREWER, J. K. (1972) "On the power of statistical tests in the *American Educational Research Journal*." Amer. Educ. Research J. 9: 391-401.

BYRNE, D. and R. RHAMEY (1965) "Magnitude of positive and negative reinforcements as a determinant of attraction." J. of Personality and Social Psychology 2: 884-889.

CARROLL, R. M. and L. A. NORDHOLM (1975) "Sampling characteristics of Kelley's ϵ^2 and Hays' ω^2." Educ. and Psych. Measurement 35: 541-554.

CHASE, L. J. and R. B. CHASE (1976) "A statistical power analysis of applied psychological research." J. of Applied Psychology 42: 29-41.

CHASE, L. J. and R. K. TUCKER (1975) "A power-analytic examination of contemporary communication research." Speech Monographs 61: 234-237.

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

——— (1973a) "Eta-squared and partial eta-squared in fixed factor ANOVA designs." Educ. and Psych. Measurement 33: 107-112.

——— (1973b) "Statistical power analysis and research results." Amer. Educ. Research J. 10: 225-229.

——— (1965) "Some statistical issues in psychological research," in B. B. Wolman (ed.) Handbook of Clinical Psychology. New York: McGraw-Hill.

——— (1962) "The statistical power of abnormal-social psychological research: a review." J. of Abnormal and Social Psychology 65: 145-153.

——— and P. COHEN (1975) Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences. Hillsdale, NJ: Lawrence Erlbaum.

CRAIG, J. R., C. L. EISON, and L. P. METZE (1976) "Significance tests and their interpretation: an example utilizing published research and ω^2." Bull. of the Psychonomic Society 7: 280-282.

CRONBACH, L. J., G. C. GLASER, H. NANDA, and N. RAJARATNAM (1972) The Dependability of Behavioral Measurement: Theory of Generalizability for Scores and Profiles. New York: John Wiley.

DODD, D. H. and R. F. SCHULTZ (1973) "Computational procedures for estimating magnitude of effect for some analysis of variance designs." Psych. Bull. 79: 391-395.

DWYER, J. H. (1974) "Analysis of variance and the magnitude of effects: a general approach." Psych. Bull. 81: 731-737.

FISHER, R. A. (1946) Statistical Methods for Research Workers. Edinburgh: Oliver and Boyd.

FLEISS, J. L. (1969) "Estimating the magnitude of experimental effects." Psych. Bull. 72: 273-276.

FRIEDMAN, H. (1968) "Magnitude of experimental effect and a table for its rapid estimation." Psych. Bull. 70: 245-251.

GAEBELEIN, J. W. and D. R. SODERQUIST (1976) "A note on variance explained in the mixed analysis of variance model." Psych. Bull. 83: 1110-1112.

GLASS, G. V and A. R. HAKSTIAN (1969) "Measures of association in comparative experiments: their development and interpretation." Amer. Educ. Research J. 6: 403-413.

GLASS, G. V, B. McGAW, and M. L. SMITH (1981) Meta-Analysis in Social Research. Beverly Hills, CA: Sage.

HALDERSON, J. S. and D. R. GLASNAPP (1972) "Generalized rules for calculating the magnitude of an effect in factorial and repeated measures ANOVA designs." Amer. Educ. Research J. 9: 301-310.

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

——— (1963) Statistics for Psychologists. New York: Holt, Rinehart & Winston.

KATZER, J. and J. SODT (1973) "An analysis of the use of statistical testing in communication research." J. of Communication 23: 251-265.

KELLY, T. L. (1935) "An unbiased correlation measure." Proceedings of the National Academy of Sciences 21: 554-559.

KENNEDY, J. J. (1970) "The eta coefficient in complex ANOVA designs." Educ. and Psych. Measurement 30: 885-889.

KERLINGER, F. N. (1964) Foundations of Behavioral Research. New York: Holt, Rinehart & Winston.

KESELMAN, H. J. (1975) "A Monte Carlo investigation of three estimates of treatment magnitude: epsilon squared, eta squared, and omega squared." Canadian Psych. Rev. 16: 44-48.

KIRK, R. E. (1968) Experimental Design: Procedures for the Behavioral Sciences. Belmont, CA: Brooks/Cole.

LEVIN, J. R. (1967) "Comment: misinterpreting the significance of 'explained variation.'" Amer. Psychologist 22: 675-676.

LINDSAY, C. A., E. MARKS, and L. GORLOW (1967) "The Herzberg theory: critique and reformulation." J. of Applied Psychology 51: 330-339.

MEEHL, R. P. (1967) "Theory-testing in psychology and physics: a methodological paradox." Philosophy of Science 34: 103-115.

MYERS, J. L. (1966) Fundamentals of Experimental Design. Boston: Allyn & Bacon.

OAKES, W. (1972) "External validity and the use of real people as subjects." Amer. Psychologist 27: 959-962.

PETERS, C. C. and W. R. VAN VOORHIS (1940) Statistical Procedures and Their Mathematical Bases. New York: McGraw-Hill.

RHOADS, S. E. (1978) "How much should we spend to save a life?" Public Interest 51: 74-92.

SAVAGE, I. R. (1957) "Nonparametric statistics." J. of the Amer. Statistical Assoc. 52: 331-344.

SCHNELLE, J. R., R. E. KIRCHNER, Jr., J. P. CASEY, P. H. USELTON, Jr., and M. P. McNEES (1977) "Patrol evaluation research: a multiple-baseline analysis of saturation police patrolling during day and night hours." J. of Applied Behavior Analysis 10: 33-40.

SECHREST, L. and R. REDNER (1979) "Strength and integrity of treatment in evaluation studies," in Evaluation Reports. Washington, DC: National Criminal Justice Reference Service.

SECHREST, L. and W. H. YEATON (1981a) "Assessing the effectiveness of social programs: methodological and conceptual issues," in S. Ball (ed.) New Directions in Evaluation Research. San Francisco: Jossey-Bass.

——— (1981b) "Empirical bases for estimating effect size," in R. F. Boruch et al. (eds) Reanalyzing Program Evaluations: Policies and Practices for Secondary Analysis of Social and Educational Programs. San Francisco: Jossey-Bass.

——— (1981c) "Estimating magnitudes of experimental effects." J. of Supplement Abstract Service, #2355.

SECHREST, L., S. G. WEST, M. A. PHILLIPS, R. REDNER, and W. YEATON (1979) "Introduction—some neglected problems in evaluation research: strength and integrity of treatments," in L. Sechrest et al. (eds.) Evaluation Studies Review Annual, Vol. 4. Beverly Hills, CA: Sage.

SMITH, M. L. (1980) "Sex bias in counseling and psychotherapy." Psych. Bull. 87: 392-407.

——— and G. V GLASS (1977) "Meta-analysis of psychotherapy outcome studies." Amer. Psychologist 32: 752-760.

——— and T. I. MILLER (1980) The Benefits of Psychotherapy. Baltimore: Johns Hopkins Univ. Press.

SODERQUIST, D. R. and R. A. HUSSIAN (1978) "The utility of utility indices." Bull. of the Psychonomic Society 11: 136-138.

TVERSKY, A. and D. KAHNEMAN (1971) "Belief in the law of small numbers." Psych. Bull. 76: 105-110.

VAUGHN, G. M. and M. C. CORBALLIS (1969) "Beyond tests of significance: estimating strength of effects in selected AVOVA designs." Psych. Bull. 72: 204-213.

VITALO, R. L. (1970) "Effects of facilitative interpersonal functioning in a conditioning paradigm." J. of Counseling Psychology 17: 141-144.

WINER, B. J. (1962) Statistical Principles in Experimental Design. New York: McGraw-Hill.

YEATON, W. H. and R. REDNER (forthcoming) "Measuring strength and integrity of treatments: rationale, techniques, and examples," in R. Conner (ed.) Methodological Advances in Evaluation Research. Beverly Hills, CA: Sage.

YEATON, W. H. and L. SECHREST (1981a) "Critical dimensions in the choice and maintenance of successful treatments: strength, integrity, and effectiveness." J. of Consulting and Clinical Psychology 49: 156-167.

——— (1981b) "Estimating effect size," in P. M. Wortman (ed.) Methods for Evaluating Health Services. Beverly Hills, CA: Sage.

Lee Sechrest is Director of the Center for Research on the Utilization of Scientific Knowledge at the University of Michigan. His research interests include the relationship between the quality of research methods and research outcomes and the utilization of these findings with regard to policy.

William H. Yeaton is a research investigator at the Center for Research on the Utilization of Scientific Knowledge at the University of Michigan. His current research interests include evaluation research methodology and the assessment of outcomes of evaluation research, especially in the area of health.