Behavior Research Methods, Instruments, & Computers
1996, 28 (1), 12-22

Measures of effect size

JOHN T. E. RICHARDSON
Brunel University, Uxbridge, England

Two different approaches have been used to derive measures of effect size. One approach is based on the comparison of treatment means. The standardized mean difference is an appropriate measure of effect size when one is merely comparing two treatments, but there is no satisfactory analogue for comparing more than two treatments. The second approach is based on the proportion of variance in the dependent variable that is explained by the independent variable. Estimates have been proposed for both fixed-factor and random-factor designs, but their sampling properties are not well understood. Nevertheless, measures of effect size can allow quantitative comparisons to be made across different studies, and they can be a useful adjunct to more traditional outcome measures such as test statistics and significance levels.

Most psychological researchers appreciate in abstract terms at least that statements describing the outcomes of tests of statistical inference need to be distinguished from statements describing the importance of the relevant findings in theoretical or practical terms. The latter may have more to do with the magnitude of the effects in question than their level of statistical significance. Cohen (1965) remarked that in research concerned with comparisons among treatment means, investigators nonetheless typically confined themselves to reporting test statistics such as t or F and did not attempt to derive measures of effect size. More specifically, Craig, Eison, and Metze (1976) surveyed the articles in three different psychological journals that had employed Student's t test; none of these reported a measure of effect size, and in many instances the "significant" effects proved on inspection to be relatively slight in magnitude. Craig et al. concluded that "researchers and journal editors as a whole tend to (over)rely on 'significant' differences as the definition of meaningful research" (p. 282). This situation does not seem to have altered in the intervening time.

This paper reviews research on the development and practical value of different measures of effect size. Classically, two different approaches have been taken in deriving such measures. One approach is based on the comparison of different treatment means, and the other approach evaluates the proportion of the variance in the dependent variable that is explained by the independent variable. Winer, Brown, and Michels (1991) noted that the first approach tends to be used in fixed-effects designs, where the treatments employed exhaust the population of interest. The second approach is typically used in random-effects designs, in which the treatments are regarded as a sample from some indefinite population of treatments, and in which it makes little sense to compute an effect size index by comparing the particular treatments that happened to be sampled.

The relevant publications on this topic extend back over much of this century, and are located in diverse sources in psychology, education, and statistics that may not be readily accessible to interested researchers. In this paper, therefore, I have endeavored to provide a tutorial overview of the subject, tracing the historical development of the measures of effect size encountered in the contemporary literature. At the same time, I want to argue that measures of effect size have a legitimate place in the advancement of current psychological theory and research; thus I will make practical suggestions about the strengths and weaknesses of particular measures.

I begin by considering the mean difference and the standardized mean difference between two independent populations, with the primary focus on the derivation and estimation of the latter as a measure of effect size and on its concomitant advantages and disadvantages. I will point out that this notion does not readily generalize to a situation in which there are three or more populations, and I will then suggest other measures based on the proportion of explained population variance. These measures represent various attempts to generalize the correlation coefficient to research designs in which the independent variable defines a number of discrete groups. This strategy can be employed regardless of whether the groups constitute a fixed set of treatments or only a particular sample from some indefinite population of treatments. Finally, I will make some comments concerning the application of measures of effect size in meta-analytic research: that is, the evaluation and comparison of the findings obtained across different studies in the research literature.

The author is grateful to Jacob Cohen, Richard Schweickert, and two anonymous reviewers for their comments on previous versions of this paper. Correspondence should be addressed to J. T. E. Richardson, Department of Human Sciences, Brunel University, Uxbridge, Middlesex UB8 3PH, United Kingdom (e-mail: john.richardson@brunel.ac.uk).

COMPARISONS BETWEEN TREATMENT MEANS

The Standardized Mean Difference

In the simplest situation, two samples of size n_1 and n_2 (where $n_1 + n_2 = N$) are drawn independently and at random from populations whose means are μ_1 and μ_2, respectively, and whose standard deviations are σ_1 and σ_2, respectively. Suppose that the two samples are found to have means of m_1 and m_2 and standard deviations of s_1 and s_2, respectively. The simplest index of effect size is the difference between the two population means, $(\mu_1 - \mu_2)$. This measure has two useful features. First, it is expressed in terms of the original units of measurement, and thus it is intuitively meaningful to researchers themselves (Wilcox, 1987). Second, although it is a parameter based on the underlying populations and hence is typically unknown, it has an unbiased estimate in the difference between the sample means $(m_1 - m_2)$ (Winer et al., 1991, p. 122).

Nevertheless, this index has a major drawback in that it depends on the specific procedure that has been employed to obtain the relevant data. In order to make meaningful comparisons among studies employing different procedures or to make useful generalizations about the relevant phenomena, it is necessary to measure the effect size in a manner that is not tied to arbitrary technical aspects of individual research studies. Cohen (1965) pointed out that this could be achieved if the difference between the two population means were standardized against the population within-treatment standard deviation. Assuming that $\sigma_1 = \sigma_2 = \sigma$, say, this yields an effect size index δ, defined as follows (Cohen, 1969, p. 18):

$$\delta = (\mu_1 - \mu_2)/\sigma.$$

In other words, σ is regarded as an arbitrary scaling factor, and δ is the mean difference that would obtain if the dependent variable were scaled to have unit variance within both populations (Hedges & Olkin, 1985, p. 76). Effectively, the magnitude of a treatment effect is judged in relation to the degree of error variability in the data (Winer et al., 1991, p. 121). Cohen (1965) proposed that "small," "medium," and "large" effects could be operationalized as effects for which the difference between the population means was 0.25σ, 0.5σ, and σ, respectively; subsequently, however (Cohen, 1969, pp. 22–24), he characterized them as effects for which $\delta = 0.2$, 0.5, and 0.8, respectively.

The most natural manner to estimate δ would be to substitute unbiased estimates of its numerator and denominator. As just noted, the difference between the sample means, $(m_1 - m_2)$, is an unbiased estimate of $(\mu_1 - \mu_2)$. Under the assumption of homogeneity of variance, an unbiased estimate, s, of the common population standard deviation, σ, is given by

$$s^2 = \frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{(n_1 + n_2 - 2)}.$$

This yields an estimator $d = (m_1 - m_2)/s$. This is not, however, an unbiased estimate of δ.[1] More specifically, Hedges (1981) showed that the expected value of d is equal to $\delta/c(m)$, where

$$c(m) = \frac{\Gamma(m/2)}{\sqrt{(m/2)} \cdot \Gamma[(m-1)/2]}$$

where $m = (n_1 + n_2 - 2)$, and where $\Gamma(x)$ is the gamma function. Hedges provided exact values of $c(m)$ for $m \leq 50$, and he also pointed out that it was closely approximated by the function $[1 - 3/(4m - 1)]$. Although it approaches unity when m is large, it is appreciably smaller than unity when m is small, indicating that d seriously overestimates δ (see also Hedges & Olkin, 1985, pp. 78–80, 104–105).

Hedges then observed that the bias inherent in d could easily be removed by defining a new estimator $d' = d \cdot c(m)$. Not only is the latter an unbiased estimator of δ, but it also has a smaller variance and hence a smaller mean square error than d. In this sense, d' dominates d as an estimator of δ (see Hedges & Olkin, 1985, p. 81). Finally, Hedges showed that when $n_1 = n_2$, d' is the unique uniformly minimum variance unbiased estimator of δ. Hedges and Olkin (1985, p. 79) pointed out that d' was itself a sample statistic, and that its sampling distribution was closely related to the noncentral t distribution. Specifically, if $\tilde{n} = n_1 n_2/N$, then $\sqrt{\tilde{n}} \cdot d'$ has a noncentral t distribution with noncentrality parameter $\sqrt{\tilde{n}} \cdot \delta$ and $(n_1 + n_2 - 2)$ degrees of freedom. Asymptotically, the sampling distribution of d' is normal with a mean equal to δ and a variance equal to $[N/(n_1 n_2) + \delta^2/(2N)]$ (p. 86). Hedges and Olkin (1985, pp. 81–82) showed that d' was neither the maximum likelihood estimator of δ, which is given by $d \cdot \sqrt{[N/(N-2)]}$, nor even the minimum mean square error estimator of δ (since a "shrunken" estimator can be specified that has uniformly smaller mean square error than d'). Nevertheless, they considered that d' had good properties for small sample sizes and should be used as the basic estimator of effect size for data obtained from a single study (p. 83).[2]

Several reviewers have attributed the index δ to Glass (1976) (e.g., Hedges & Becker, 1986; Wilcox, 1987; Winer et al., 1991, p. 122). This is clearly incorrect; Cohen (1965) discussed the basic notion informally and then formally as an effect size index (Cohen, 1969). Glass's particular contribution was to point out that an estimate of δ could itself be used as a dependent variable in order to evaluate the consistency and the magnitude of a particular phenomenon across different studies in the literature. Smith and Glass (1977) used this approach to argue for the efficacy of psychotherapy on the basis of the effect sizes obtained in 375 different studies. Subse-

quently, Glass, McGaw, and Smith (1981) provided a more extended account of the meta-analysis of social research, and nowadays measures of this sort are fairly widely used, most notably in the investigation of gender differences (see, e.g., Hyde, 1981; Hyde, Fennema, & Lamon, 1990; Hyde & Linn, 1986). Glass and his coauthors did not mention Cohen's earlier work in any of these publications, but Cohen was patently the intended target of their criticism that "*there is no wisdom whatsoever in attempting to associate regions of the effect-size metric with descriptive adjectives such as 'small,' 'moderate,' 'large,' and the like*" (Glass et al., 1981, p. 104, italics in original).

One distinctive characteristic of Glass's (1976) account was that it was concerned with the comparison of one or more treatment groups with a single control group. The accompanying illustrations showed hypothetical distributions of the treatment groups expressed in terms of percentiles of the control group. In other words, Glass standardized differences between the group means against the standard deviation of the control group alone (see also Glass et al., 1981, p. 29). If the latter group is arbitrarily designated as Group 2, the estimate of the population effect size δ for Group 1 would be $(m_1 - m_2)/s_2$. However, as Glass et al. (1981, p. 106) themselves noted, various choices of the standard deviation with which to scale the differences between the group means can result in substantial differences in effect size.

Glass's (1976) original paper contained no justification for this way of computing a standardized mean difference. According to Hedges (1981), Glass's own concern was that the standard deviations of different samples would vary by chance even if the variances of the underlying populations were homogeneous. Consequently, pooling pairs of sample variances could result in different standardized values of identical mean differences when several treatment groups were being compared with a single control group. Conversely, as Glass et al. (1981, pp. 106–107) pointed out, standardizing the differences between the mean scores obtained across several treatment conditions against the standard deviation of the control group alone would ensure that equal mean differences were associated with equal effect sizes in the face of heterogeneous within-group variances. They also cautioned that the latter problem could arise in research practice as the result of ceiling and floor effects (pp. 109–111).

However, Hedges (1981) argued that, if the assumption of homogeneity of variance were tenable, then the most precise estimate of the population variance would be obtained by pooling all the sample variances, and that in any case Glass's reservation simply did not apply to an investigation that involved merely two samples (see also Hedges & Olkin, 1985, p. 78). Hedges went on to show that the expected value of Glass's estimate of δ was also $\delta/c(m)$, where $c(m)$ was defined as above, but where m was simply the degrees of freedom for the control group $(n_2 - 1)$. Hedges and Olkin (1985, p. 79) pointed out that the bias and the variance of d were smaller than the bias and the variance of Glass's estimate, and that consequently the former was a uniformly better estimator than the latter, regardless of the value of δ. Rosenthal and Rubin (1982) and Kraemer (1983) showed how values of d obtained from several different experiments could be tested for homogeneity for the purposes of meta-analysis. Hedges (1982a, 1982b) presented an equivalent test for the homogeneity of values of the unbiased estimate d', and showed how the latter values could be combined to yield both a weighted estimator of δ and confidence intervals for δ. Hedges (1982c) developed additional procedures for analyzing whether effect size could be predicted by either continuous or discrete independent variables.

Strengths and Weaknesses of the Standardized Mean Difference

Hedges and Becker (1986) identified a number of positive features of the standardized mean difference as a measure of effect size. First, they claimed that it was easy to understand and had a consistent interpretation across different research studies. Second, it preserves information about the direction of the relevant effects (although it is possible to adapt it to measuring differences in either direction by defining $\delta = |\mu_1 - \mu_2|/\sigma$; Cohen, 1969, p. 18). Third, the sampling distributions of the uncorrected statistic d and the corrected statistic d' are simple and well understood, which facilitates the use of analytic procedures. In addition, Hedges and Becker pointed out that these quantities can be readily computed from the values of the test statistics t and F reported by other researchers in published articles. This is of course not surprising, since t is normally calculated as $(m_1 - m_2)/\sqrt{[s^2(1/n_1 + 1/n_2)]}$, and since F can be shown to be equal to t^2. If a study provides a value of t, then the value of the uncorrected statistic d can be computed as $t \cdot \sqrt{(1/n_1 + 1/n_2)}$; if a study provides a value of F from a one-way analysis of variance, then the value of d can be computed as $\sqrt{[F \cdot (1/n_1 + 1/n_2)]}$. Both computations obviously assume homogeneity of within-group variance (Glass et al., 1981, p. 108). More complicated computations are needed in the case of factorial designs, but in each case the value of the corrected statistic d' can be calculated as $d \cdot c(m)$.

Against these features, a number of criticisms have been put forward. First, Gibbons, Olkin, and Sobel (1977) suggested that because the standardized mean difference was unitless, its specification "requires a much more sophisticated acquaintance with both the details of the application as well as the statistical analysis and its implications" (p. 63). Nevertheless, although there may well be practical circumstances in which an investigator might find it more congenial to express research findings in terms of the original units of measurement, there are also many situations in which the specific scale of measurement is of no theoretical or practical interest. Second, Wilcox (1987) pointed out that the standardized mean difference assumed that the samples had been drawn from populations with the same variance, and that if this as-

sumption were violated, a unitless measure of effect size "would seem not to exist" (p. 47). However, Hedges and Olkin (1985, p. 78) remarked that there were different ways to create an estimated standardized mean difference of the form $(m_1 - m_2)/s^*$, where s^* was a standard deviation; different choices of s^* would yield different estimators, but s^* could be defined (for example) either as s_1 or as s_2 (see also Glass et al., 1981, p. 106). Finally, Kraemer and Andrews (1982) noted that the standardized mean difference reflected the choice of measuring instrument as well as the magnitude of the treatment effect in that it was sensitive to nonlinear transformations of the raw data. They put forward a nonparametric measure of effect size based on the ordinal properties of the measurement scales and therefore invariant under all monotonic transformations of the data. Nevertheless, their criticism would also be true of the nonstandardized mean difference, and it does not of course detract from the fact that the standardized mean difference is invariant over all *linear* transformations of the raw data.

However, Hedges (1981) himself identified three factors that tend to weaken the standardized mean difference as a measure of effect size. Two of these relate to the fact that the magnitude of the group difference is compared with the variability within each of the groups, with the implicit assumption that the latter results from stable differences among subjects, an assumption that might not be valid. First, the responses of different subjects to the experimental treatment may vary, even if the nature of the intervention is identical for all the subjects in the experimental group. In other words, there may be a subject-by-treatment interaction, and this will contribute to the residual term in the structural model, as will any other unmeasured "nuisance" variables. Second, if the response measure is not perfectly reliable, then measurement error will also contribute to the within-group variability. If δ is taken to refer to the standardized mean difference in the absence of errors of measurement, d' will systematically underestimate that quantity. Hedges then noted that the standardized mean difference when errors of measurement are present is $\delta' = \delta \cdot \sqrt{\rho}$, where ρ is the reliability of the response measure. Accordingly, if ρ is known, one can remove the bias resulting from measurement error by dividing d' by $\sqrt{\rho}$. The third factor is the adequacy of the response measures as valid indices of the underlying traits, abilities, or processes; to the extent that they have unique factors, they will be partially invalid. Hedges showed that if the experimental treatment affects only the common factor assumed to be shared by the tests measuring a particular trait, ability, or process, then the presence of unique factors reduces the standardized mean difference (and hence the estimated value of δ). The extent of this bias can be computed (and thus corrected) if the correlation between the invalid response scale and a valid response scale is known. However, if the intervention affects both the common and unique factors, the effect of invalidity may be either to increase or decrease the standardized mean difference.

Generalization to $k > 2$

As originally defined above, the parameter δ does not generalize to designs involving k treatments (where $k > 2$) in a straightforward manner. This may encourage researchers to group together possible treatments into just two superordinate categories (e.g., experimental vs. control) for the purposes of meta-analysis. Presby (1978) argued, however, that this would obscure genuine differences among the treatments within these categories. Cohen (1969, p. 269) suggested that for $k \geq 2$, one could define δ to be the range of the standardized means (or the standardized range of the means), $(\mu_{\max} - \mu_{\min})/\sigma$, where $\mu_{\max}$ is the largest of the k means, $\mu_{\min}$ is the smallest of the k means, and σ, as before, is the common standard deviation within each of the k populations. Cohen suggested that when $k = 2$, the effect size index is reduced to that defined earlier, $(\mu_1 - \mu_2)/\sigma$. In fact, however, it reduces to the nondirectional effect size index, $|\mu_1 - \mu_2|/\sigma$. Moreover, when $k > 2$, this new index is not affected by the precise values of the $(k - 2)$ intermediate means, and hence it is an insensitive measure of effect size among the entire set of k treatments.

Earlier, Winer had described an alternative approach to this problem as part of the single-factor analysis of variance (1962, pp. 57–65). He defined the effect of the ith treatment, τ_i, as the difference between the population mean for the ith treatment, μ_i, and the grand mean of the population means, μ. Winer then pointed out that one parameter indicating the extent to which the treatment effects differ is $\sigma_\tau^2 = (\sum \tau_i^2)/(k - 1) = [\sum(\mu_i - \mu)^2]/(k - 1)$. He showed that if each sample contains n individuals and σ^2 is the variance due to experimental error within each of the populations, then the expected value of the mean squares across the treatments is $(n\sigma_\tau^2 + \sigma^2)$, and the expected value of the residual mean squares is σ^2. The null hypothesis (that $\sigma_\tau^2 = 0$) might therefore be tested by computing the usual F ratio between the mean squares across the treatments and the residual mean squares. Under the alternative hypothesis (that $\sigma_\tau^2 \neq 0$), Winer stated that the expected value of the latter ratio was $(n\sigma_\tau^2 + \sigma^2)/\sigma^2$, but this is incorrect. The expected value of the ratio between two variables is a biased estimate of the ratio between their individual expected values (see note 1). In particular, if s_1^2 and s_2^2 are independent unbiased estimators of σ_1^2 and σ_2^2, respectively, then the expected value of s_1^2/s_2^2 is greater than σ_1^2/σ_2^2 (Kendall & Stuart, 1977, p. 242). This error was corrected in the second edition of Winer's book (see Winer, 1971, p. 166). Otherwise, he gave no indication as to how his effect size index might be estimated from sample data.

The rationale for the use of $(k - 1)$ rather than k in the denominator of Winer's formula for σ_τ^2 is also unclear. Vaughan and Corballis (1969) noted that it was appropriate in the case of a random-effects design where the k treatments are regarded as a sample from some indefinite population of treatments. However, in this case, as mentioned above, it makes little sense to compute an ef-

fect size index by comparing the means of the k treatments that happened to be sampled. For a fixed-effects design, on the other hand, the k treatments exhaust the relevant population, and σ_τ^2 is itself a parameter of that population rather than a statistic. Vaughan and Corballis pointed out that it should therefore be defined as $(\sum\tau_i^2)/k$ (see Winer et al., 1991, p. 123). The expected value of the mean squares across the treatments is therefore $[kn \cdot \sigma_\tau^2/(k-1) + \sigma^2]$ (see also Fleiss, 1969), and it follows that an unbiased estimate of σ_τ^2 is given by $(k-1)[\text{MS(Treatments)} - \text{MS(Residual)}]/(kn) = (k-1)(F-1) \cdot \text{MS(Residual)}/(kn)$ (cf. Winer, 1971, pp. 428–429). Vaughan and Corballis showed how this approach could be generalized to two-factor and three-factor designs with interaction terms and to designs in which within-subjects comparisons are used.

The variance of the treatment means has the disadvantage that it is expressed in terms of the square of the original units of measurement, a scale that might not in itself be meaningful, and that will in any case be contingent on the specific procedure that was employed to obtain the raw data. Once again, it might be helpful to standardize this measure in some way, so that it is not tied to arbitrary technical aspects of particular research studies. Hays (1963, p. 384) pointed out that under the alternative hypothesis the ratio MS(Treatments)/MS(Residual) would be expected to follow the noncentral F distribution with a noncentrality parameter of $\sqrt{[(\sum n \cdot \tau^2)/\sigma^2]}$, which is equal to $\sqrt{(N\sigma_\tau^2/\sigma^2)}$ or $(\sigma_\tau/\sigma) \cdot \sqrt{N}$. Consequently, the variance (or the standard deviation) of the treatment means might be conveniently standardized against the variance (or the standard deviation) of the constituent populations.

Cohen (1969, pp. 267–269) accordingly proposed an alternative effect size index, f, defined as the ratio between the standard deviation of the treatment means and the standard deviation within the populations. Thus, $f = \sigma_\tau/\sigma$, where $\sigma_\tau = \sqrt{\{[\sum(\mu_i - \mu)^2]/k\}}$. As Cohen noted, this is equal to the standard deviation of the standardized population means and is a dimensionless quantity. Cohen claimed that when $k = 2$, $f = \frac{1}{2}\delta$, which is strictly speaking incorrect: f is nonnegative and nondirectional and thus is equal to $\frac{1}{2}|\delta|$. Cohen went on to suggest that small, medium, and large effects could be defined in terms of values of f equal to 0.1, 0.25, and 0.4 (pp. 277–281). He also discussed how f could be applied to factorial designs (pp. 277–281), and in later writings he described how it could be generalized to multiple regression (Cohen, 1977, p. 410; 1988, p. 473). This index is in itself of limited relevance to research practice, however, because Cohen did not show how it could be reasonably estimated from sample data. Nevertheless, the square of f is equal to the ratio between the component of variance that is explained by the treatment variable and the component that is not so explained. The alternative approach to deriving measures of effect size is based on the estimation of these variance components.

COMPARISONS BETWEEN VARIANCE COMPONENTS

The Correlation Coefficient

The alternative approach to deriving measures of effect size is based on quantifying the proportion of variance in the dependent variable that is explained by the independent variable. As Hedges and Olkin (1985, p. 100) noted, the explained "variance" is often not formally a variance at all, but the difference between the overall variance in the dependent variable and the conditional variance in the dependent variable, taking into account the effect of the independent variable. On this approach, one tackles the problem of quantifying the magnitude of treatment effects by measuring the strength of association between the independent variable and the dependent variable, and the latter is expressed in terms of some kind of correlation coefficient (Winer et al., 1991, p. 121).

Cohen (1965) remarked that the possibility for confusion between the levels of statistical significance associated with particular empirical findings and the magnitude and hence the importance of the relevant effects could be reduced if the outcomes are expressed as correlation coefficients. It is fairly well known that the linear correlation coefficient, Pearson r, has a straightforward interpretation as a measure of effect size, in that r^2, which is often termed the "coefficient of determination," is equal to the proportion of the total variation in the dependent variable that can be predicted or explained on the basis of its regression on the independent variable within the sample being studied (see, e.g., Hays, 1963, p. 505). Similarly, the square of a population correlation coefficient, ρ, can be interpreted as the proportion of the variance in the dependent variable that is explained by its regression on the independent variable within the population in question (see, e.g., Hays, 1963, p. 512). Elsewhere, Cohen (1969, pp. 76–77) suggested that in correlational research "small," "medium," and "large" effects could be characterized as values of ρ equal to .1, .3, and .5, corresponding to values of ρ^2 equal to .01, .09, and .25, respectively. In addition, Glass (1976) noted that r could be employed as an index of effect size in meta-analytic investigations, and Kraemer (1979) described procedures for evaluating the homogeneity of the correlation coefficients obtained from several different studies.

Suppose that the number of pairs of observations within a sample is N, that the independent and dependent variables are X and Y, respectively, and that the total variation (in other words, the total sum of squares) in Y is SS(Total). The mean square that is associated with the linear regression of Y on X will be $\text{SS(Total)} \cdot r^2$ with one degree of freedom, and the mean square that is associated with the residual (i.e., unexplained) variation in Y will be $\text{SS(Total)} \cdot (1 - r^2)/(N-2)$ with $(N-2)$ degrees of freedom (cf. Hays, 1963, pp. 517–521). Under the null hypothesis of no correlation between X and Y (i.e., $\rho = 0$), these are independent estimates of the population

variance in Y, and hence the statistic $r^2 \cdot (N - 2)/(1 - r^2)$ is distributed as F with 1 and $(N - 2)$ degrees of freedom. Equivalently, the square root of this quantity, $r \cdot \sqrt{[(N - 2)/(1 - r^2)]}$, is distributed as t with $(N - 2)$ degrees of freedom.

Under the alternative hypothesis (i.e., $\rho \neq 0$), however, the total population variance on Y (σ_Y^2, say) is to be divided into two parts: the explained variance, $\rho^2\sigma_Y^2$, and the residual variance ($\sigma_{Y\ X}^2$, say). Here, the expected value of the mean square associated with the total variance in Y is σ_Y^2, but the expected value of the mean square associated with the residual variance is $\sigma_{Y|X}^2$. The ratio between the latter mean square and the former mean square is thus a reasonable estimate of the proportion of variance in the dependent variable that is not explained by its regression on the independent variable, and hence the following would be a reasonable estimate of ρ^2:

$$\text{est.}\,\rho^2 = 1 - \frac{\text{MS(Residual)}}{\text{MS(Total)}}.$$

The latter quantity is equal to $(Nr^2 - r^2 - 1)/(N - 2)$, which is less than r^2 itself except when $r = \pm 1$.

A different approach to the same problem can be taken if one notes that the expected value of the mean square associated with the regression of Y on X in the sample is $(\sigma_{Y|X}^2 + N\rho^2\sigma_Y^2)$, and the expected value of the mean square associated with the residual variance in the sample is $\sigma_{Y\ X}^2$. It then follows that the difference between these mean squares is an unbiased estimate of the quantity $N\rho^2\sigma_Y^2$, whereas the sum of the former and $(N - 1)$ times the latter is an unbiased estimate of $N\sigma_Y^2$. Thus, the ratio between these quantities would be an alternative estimate of ρ^2:

$$\text{est.}\,\rho^2 = \frac{\text{MS(Regression)} - \text{MS(Residual)}}{\text{MS(Regression)} + (N - 1) \cdot \text{MS(Residual)}}.$$

This suggestion was made by Hays (1963, pp. 523–524). The latter quantity is equal to $(Nr^2 - r^2 - 1)/(N - r^2 - 1)$, which is once again less than r^2 except when $r = \pm 1$. The ratio between the first and second estimates of ρ^2 equals 1 + [MS(Residual)/SS(Total)], which is at most $[1 + 1/(N - 2)]$.

If the independent variable is dichotomous, the situation is formally equivalent to the comparison of two treatment means, as discussed earlier in this article. In other words, as Cohen (1965) pointed out, an index of effect size for the comparison of two treatment means can be obtained if one defines a dichotomous dummy variable to represent membership of one or the other of the two populations and computes the point-biserial correlation coefficient between the continuous dependent variable and the dichotomous dummy variable. This can be calculated from reported values of t or F by the formulae $r_{pb} = \sqrt{[t^2/(t^2 + N - 2)]}$ and $r_{pb} = \sqrt{[F/(F + N - 2)]}$. In this situation, r_{pb}^2 measures the proportion of the total variation in the dependent variable that is associated with membership of the two treatment groups. Cohen (1969, p. 22) pointed out that there was a straightforward relationship between the population point-biserial correlation coefficient ρ_{pb} and the effect size index δ described previously. If p and q are the proportions of cases in the two populations, then $\rho_{pb} = \delta/\sqrt{[\delta^2 + (1/pq)]}$; more specifically, if $p = q = 1/2$, then $\rho_{pb} = \delta/\sqrt{(\delta^2 + 4)}$. In the case of the sample statistics r_{pb} and d, however, simple algebraic manipulation of the formulae already given shows that $r_{pb} = d/\sqrt{[d^2 + N(N - 2)/n_1n_2]}$.

The Correlation Ratio

The same procedure can be used in situations in which there are more than two treatment groups, provided that they can be assigned meaningful numerical values. Of course, as Hedges and Olkin (1985, p. 101) pointed out, in this case the squared correlation coefficient reflects the degree of linear relationship between the independent variable and the dependent variable, and does not necessarily reflect nonlinear components of their association. Equivalently, in comparing more than two treatment samples, the computation of a linear correlation coefficient will systematically underestimate the effect size. The appropriate generalization of the correlation coefficient is the correlation ratio, η (eta), which was first developed by Pearson (1905) to measure the degree of association between two variables, X and Y, when the different values of X are categorized into various classes or arrays. The square of the correlation ratio is referred to as the differentiation ratio, and measures the proportion of the variability in Y that is associated with membership of the different classes or arrays defined by X. It can be calculated conveniently with the formula $\eta^2 =$ SS(Treatment)/SS(Total) = 1 − SS(Residual)/SS(Total).

The correlation ratio thus subsumes both the linear and the nonlinear components of the association between X and Y. If the number of groups is greater than two (k, say) and they have been assigned numerical values in an arbitrary way, it does not make sense to talk about the "direction" of such an association, and hence η is conventionally taken to be a positive quantity (Peters & Van Voorhis, 1940, pp. 313, 318). Pearson noted that $\eta \geq r$, with equality only when there is a linear relationship between the dependent variable and the numerical values assigned to the various groups defining the independent variable; equivalently, the difference between the differentiation ratio and the coefficient of determination is an index of the deviation of the obtained regression curve from the least-squares regression line (p. 11; cf. Fisher, 1922). The differentiation ratio is also equal to the squared multiple correlation coefficient obtained when the single X variable is recoded as $(k - 1)$ independent dichotomous "dummy" variables (Cohen, 1969, p. 275; Winer et al., 1991, p. 124).

If the total variation in Y is referred to as SS(Total), the mean square between the different groups defined by the X variable is SS(Total) $\cdot\ \eta^2/(k - 1)$ and the mean square within the different groups is SS(Total) $\cdot (1 - \eta^2)/(N - k)$. Under the null hypothesis of no difference among the latter groups, these two quantities are independent estimates of the population variance in Y, and hence the sta-

tistic $\eta^2(N-k)/[(1-\eta^2)(k-1)]$ is distributed as F with $(k-1)$ and $(N-k)$ degrees of freedom (Diamond, 1959, p. 186; Hays, 1963, p. 548; McNemar, 1962, pp. 270–271). Cohen (1965) pointed out that the corresponding values of η can be calculated from reported values of F by means of the following formula: $\eta^2 = F(k-1)/[F(k-1) + (N-k)]$. When $k = 2$, η is equivalent to the point-biserial correlation coefficient and can be calculated from reported values of t by means of the following formula: $\eta^2 = t^2/(t^2 + N - 2)$ (cf. Hays, 1981, p. 294).

For modern readers, Pearson's (1905) use of the Greek letter η is a trifle confusing, because it obscures the fact that the correlation ratio measures the degree of association between the X and Y variables within a particular sample. Subsequent commentators recognized this explicitly or implicitly in their own writings on this subject (see, e.g., Cohen, 1965; Diamond, 1959, pp. 54–55; McNemar, 1962, pp. 202–203, 270–271; Peters & Van Voorhis, 1940, pp. 312–319). Hays (1981, p. 349) suggested that the correlation ratio was a perfectly satisfactory descriptive statistic for evaluating the extent to which the experimental treatments accounted for variance in the dependent variable. Nevertheless, it is not satisfactory for most research purposes because it is not an unbiased estimate of the corresponding parameter of the underlying population.

Sample Estimates of the Population Correlation Ratio

This problem had been suspected by a number of researchers, including Pearson (1923) himself. However, it was first properly analyzed by Kelley (1935), who defined the true or population value of the correlation ratio, $\tilde{\eta}$, in terms of the proportion of the total population variance in Y that was explained by membership of the various classes or arrays defined by X. In this case, the residual variance in Y (i.e., $\sigma^2_{Y|X}$) is equal to the variance due to experimental error within each of the treatment populations (i.e., σ^2). Consequently, $\tilde{\eta}^2 = 1 - \sigma^2/\sigma^2_Y$. An unbiased estimate of the residual variance in Y is SS(Residual)/$(N-k)$, whereas an unbiased estimate of the total variance in Y is SS(Total)/$(N-1)$. Kelley then argued that an unbiased estimate of $\tilde{\eta}^2$, which he called ε^2, is given by the formula

$$\varepsilon^2 = 1 - \frac{(N-1)\cdot \text{SS(Residual)}}{(N-k)\cdot \text{SS(Total)}}.$$

An informal derivation of this was offered by Diamond (1959, p. 130). Since $\eta^2 = 1 - \text{SS(Residual)}/\text{SS(Total)}$, $\varepsilon^2 = (\eta^2 N - k + 1 - \eta^2)/(N-k) = \eta^2 - (1-\eta^2)(k-1)/(N-k)$. Thus, $\varepsilon^2 \le \eta^2$, with equality only when $\eta^2 = \varepsilon^2 = 1$. Kelley also noted that when $\varepsilon^2 = 0$, $\eta^2 = (k-1)/(N-k)$, which he concluded was the expected value of η^2 under the null hypothesis. It may be noted that when $k = 2$, $\eta^2 \equiv r^2$ and ε^2 reduces to the first of the two estimates of ρ^2 that were derived earlier. Peters and Van Voorhis (1940, pp. 421–422) observed that corresponding values of ε^2 could be calculated from reported values of F by means of the formula $\varepsilon^2 = (F-1)(k-1)/[F(k-1) + (N-k)]$. First Cohen (1965) and then Winer et al. (1991, p. 124) pointed out that the statistic ε^2 is exactly equivalent to the "shrunken" estimate of the multiple correlation coefficient originally proposed by Wherry (1931).

Hays (1963, p. 381–385) took an alternative approach based on the deviation of the mean of the ith population from the overall mean, $\tau_i = \mu_i - \mu$. Assuming a fixed-effects design, $\sigma^2_\tau = (\sum \tau_i^2)/k$, as noted earlier. In this case, $\sigma^2_Y = \sigma^2 + \sigma^2_\tau$. Hays introduced the symbol ω^2 to refer to the population value of the squared correlation ratio, and noted that (in the present notation) $\omega^2 = (\sigma^2_Y - \sigma^2)/\sigma^2_Y = \sigma^2_\tau/(\sigma^2 + \sigma^2_\tau)$ (see also Cohen, 1969, pp. 273–274). The expected value of the mean square across the treatments is $[kn \cdot \sigma^2_\tau/(k-1) + \sigma^2]$, and the expected value of the residual mean square is σ^2. Under the null hypothesis (i.e., that $\sigma^2_\tau = 0$), the ratio MS(Treatments)/MS(Residual) would be expected to follow the F distribution with $(k-1)$ and $(N-k)$ degrees of freedom. Under the alternative hypothesis (i.e., that $\sigma^2_\tau \neq 0$), that ratio would be expected to follow the noncentral F distribution with a noncentrality parameter of $\sqrt{(N\sigma^2_\tau/\sigma^2)} = \sqrt{[N\omega^2/(1-\omega^2)]}$. It then follows that the expected value of $(k-1)$[MS(Treatments) − MS(Residual)] is equal to $kn \cdot \sigma^2_\tau$, and that the expected value of $(k-1) \cdot$ MS(Treatments) plus $(N-k+1) \cdot$ MS(Residual) is equal to $kn(\sigma^2 + \sigma^2_\tau)$. Hays concluded that the following was a reasonable estimate of the squared population correlation ratio:

$$\text{est. } \omega^2 = \frac{\text{SS(Treatments)} - (k-1)\cdot \text{MS(Residual)}}{\text{SS(Total)} + \text{MS(Residual)}}.$$

Fleiss (1969) and Winer et al. (1991, pp. 123–125) subsequently provided similar estimates of ω^2. It can readily be shown that est. $\omega^2 \le \eta^2$, with equality only when est. $\omega^2 = \eta^2 = 1$.

Glass and Hakstian (1969) subsequently noted that

$$\varepsilon^2 = \frac{\text{SS(Treatments)} - (k-1)\cdot \text{MS(Residual)}}{\text{SS(Total)}}.$$

and hence that $\varepsilon^2/(\text{est. } \omega^2) = 1 + [\text{MS(Residual)}/\text{SS(Total)}]$. They then commented that this latter quantity has an upper bound when SS(Residual) = SS(Total) of $[1 + 1/(N-k)]$ and tends toward 1 as N increases, and they concluded that in practice the two statistics would probably not differ by more than 0.01 or 0.02. Fleiss (1969) observed that corresponding values of est. ω^2 could be calculated from reported values of F by the formula est. $\omega^2 = (k-1)(F-1)/[(k-1)(F-1) + N]$, and Craig et al. (1976) tabulated values of est. ω^2 that corresponded to commonly used threshold probability (alpha) levels for different values of $(N-2)$. Hays (1963, pp. 326–327) himself noted that when $k = 2$, $\omega^2 = (\mu_1 - \mu_2)^2/4\sigma^2_Y$, and that values of est. ω^2 could be calculated from re-

ported values of t by the formula $(t^2 - 1)/(t^2 + N - 1)$. However, in this case, $\eta^2 \equiv r^2$ and est. ω^2 reduces to the second of the two estimates of ρ^2 derived earlier.

The Intraclass Correlation Coefficient

It should be noted that Hays's derivation of est. ω^2 assumed that the X variable was a fixed factor: That is, the particular groups included in the study exhausted all the treatments of interest and were not obtained by sampling from some wider set of treatments or factor levels. When X is a random factor, however, it is possible to define an analogous measure of effect size, the population intraclass correlation coefficient, ρ_I. This expresses the proportion of the total variance that is attributable to the membership of different categories within this wider set. (Note that this definition is more akin to that of the coefficient of determination, r^2, than to that of the coefficient of correlation, r.) Hays (1963, p. 424) commented that this index was identical to ω^2 in its general form and its meaning, but he claimed that different estimation methods applied in this situation.

In fact, it is possible to derive two different estimates of ρ_I that parallel the two different estimates of the squared population correlation ratio described earlier. In the first place, Kelley's (1935) account did not make any assumption about whether the treatments factor was fixed or random. Even with a random-effects design, it remains the case that SS(Total)/$(N - 1)$ is an unbiased estimate of the total variance in Y and that SS(Residual)/$(N - k)$ is an unbiased estimate of the residual variance in Y. It thus follows that the ratio between the latter estimate and the former estimate provides a reasonable estimate of the proportion of the total variance in the dependent variable that is not explained by membership of the set of treatment categories defined by the independent variable, and that the complement of this ratio, which Kelley denoted by ε^2, yields a reasonable estimate of the population intraclass correlation coefficient.

The second estimate of ρ_I is derived from the account that had been presented incorrectly by Winer (1962, pp. 57–65) in the case of a fixed factor. With a random factor, the variance of the treatment means, σ_τ^2, is equal to $(\sum \tau_i^2)/(k - 1)$, and the expected value of the mean squares across the treatments is $(n \cdot \sigma_\tau^2 + \sigma^2)$. Vaughan and Corballis (1969) noted that an unbiased estimate of σ_τ^2 was therefore given by the expression [MS(Treatments) − MS(Residual)]/n. Since F = MS(Treatments)/MS(Residual), this is equal to $(F - 1) \cdot$ MS(Residual)/n. Moreover, an unbiased estimate of $(\sigma_\tau^2 + \sigma^2)$ is given by [MS(Treatments) + $(n - 1) \cdot$ MS(Residual)]/n. It follows that the ratio between these two quantities will be a reasonable estimate of the population intraclass correlation coefficient:

$$\text{est. } \rho_I = \frac{\text{MS(Treatments)} - \text{MS(Residual)}}{\text{MS(Treatments)} + (n-1) \cdot \text{MS(Residual)}}.$$

Vaughan and Corballis pointed out that this was a consistent estimate of ρ_I, but also a biased one. They went on to show how this approach could be generalized to two-factor and three-factor designs including estimates of interaction effects and to designs using within-subject comparisons. Fleiss (1969), Dodd and Schultz (1973), and Shrout and Fleiss (1979) made further contributions to this discussion.

Further Ramifications

It should also be noted that, although the different estimators of ρ, $\tilde{\eta}^2$, and ρ_I described above are prima facie reasonable, none of them could be regarded as intrinsically unbiased (cf. Hedges & Olkin, 1985, p. 102). Each is based on estimating the value of a fraction by means of inserting unbiased estimates of its numerator and denominator. Winer et al. (1991, p. 125) justified this as a "heuristic approach," and yet it is well known that the expected value of the ratio between two variables is a biased estimate of the ratio between their expected values (see note 1). Glass and Hakstian (1969) noted that ε^2 was not an unbiased estimate of $\tilde{\eta}^2$, contrary to Kelley's (1935) original claim, while Winkler and Hays (1975) were themselves quite explicit that Hays's estimate of ω^2 "is biased, and it may not be a good estimator in some other respects as well" (p. 766). It would perhaps be reasonable to think that ε^2 was more satisfactory than η^2 as an estimate of $\tilde{\eta}^2$, and Winer (1971, p. 124) indeed stated without elaboration that the former tended to be less biased than the latter. At present, however, there is no principled means of differentiating between ε^2 and est. ω^2 or est. ρ_I as estimates of $\tilde{\eta}^2$.

Hays (1963, pp. 325, 547) introduced the expression ω^2 as opposed to η^2 to make it explicit that the former was a measure of the strength of the association between the independent and dependent variables within the underlying population, while the latter was a descriptive statistic based on the comparison of two or more samples. Nevertheless, Hays incorrectly referred to η^2 itself as the correlation ratio rather than as the squared correlation ratio or differentiation ratio. This usage was also adopted more recently by Hedges and Olkin (1985, pp. 101–102).

Moreover, contemporary commentators have come to use the symbol η^2 as a parameter of a population (in other words, the proportion of the total variance of the k populations that is accounted for by membership of a particular population) that itself has to be estimated from statistics calculated from a sample. This practice was employed by Wishart (1932), who introduced the symbol E^2 to denote the square of the correlation ratio calculated from a sample, but it has also been picked up by a number of modern authors (see Cohen, 1969, pp. 274–281; Hedges & Olkin, 1985, pp. 101–102; Winer et al., 1991, pp. 123–124). Cohen (1969) noted that the correlation ratio was related to his effect size index, f (the standard deviation of the standardized population means), by the formula $\eta^2 = f^2/(1 + f^2)$ or, equivalently, $f^2 = \eta^2/(1 - \eta^2)$. This is analogous to the association between the point-biserial correlation coefficient ρ_{pb} and the effect size index δ (Winer et al., 1991, p. 124).

Variance is a quantity that by definition cannot be negative, and it follows a fortiori that measures of explained variance must be nonnegative too. However, Peters and Van Voorhis (1940, p. 355) pointed out that the estimate ε^2 will be negative whenever MS(Treatments) is less than MS(Residual), and it can easily be shown that the same is also true of the other estimates of the proportion of population variance that is explained by the independent variable in question. Equivalently, these estimates of explained population variance will be negative whenever the corresponding values of t or F are less than 1. Hays (1963, pp. 327, 383) recommended that in this case the researcher should set the estimate of the proportion of explained population variance equal to zero. Vaughan and Corballis (1969) pointed out that this strategy imposes a positive bias on these variables, and hence it invalidates them as estimators of the relevant population variance components. This consequence would not be important if the researcher's concern were merely to determine whether the observed estimate of explained population variance exceeded some critical level for the sole purpose of rejecting the null hypothesis. Vaughan and Corballis argued that the original negative value should be reported if it is to be compared with estimates obtained in other experiments.

Limitations of Measures of Explained Variance

O'Grady (1982) identified three somewhat more fundamental limitations on estimates of explained variance as measures of effect size. First, if the dependent variable is not perfectly reliable, then measurement error will contribute to the within-group variability and reduce the proportion of variance that can in principle be explained. In general, an estimate of explained variance will have an upper bound equal to the product of the reliabilities of the independent and dependent variables. O'Grady argued that, since many studies that try to measure explained variance use only a single manipulation of the supposed causal factor and a single criterion to evaluate the effects of that manipulation, the reliabilities of these variables might be quite low, even if they are sound from a theoretical point of view. Consequently, much psychological research would appear to be destined to generate relatively small measures of explained population variance.

Second, O'Grady pointed out a number of methodological issues. Possibly the most important of these is the observation that measures of the proportion of explained population variance depend on the choice and number of levels of the independent variable. Fisher (1925, p. 219) pointed out that when the latter variable is theoretically continuous, the value of the differentiation ratio (and hence of the correlation ratio) obtained from a particular sample would depend not only on the range of values that is explored, but also on the number of values employed within that range. Similarly, Lindquist (1953) argued that "in most applications of analysis of variance to experimental designs, the value of either F or ε^2 depends upon the arbitrary choice of categories in the treatment classifications, and hence is not meaningful as an index of strength of relationship" (p. 63; see also Glass & Hakstian, 1969; Hedges & Olkin, 1985, p. 104; Norton & Lindquist, 1951). Levin (1967) noted in particular that the percentage of explained variance could be artificially inflated by the inclusion of a treatment group that was known to produce a substantially different level of performance. Levin suggested that in this situation, SS(Treatments) should be partitioned into $(k - 1)$ orthogonal components and a value of ω^2 calculated for each one. O'Grady suggested that as a general rule the more diverse a population is in terms of the factor of interest, the higher will be the estimates of explained variance in the dependent variable. As Hedges and Olkin (1985) concluded, "Indices of variance accounted for depend on functions of arbitrary design decisions as well as the underlying relationship between theoretical constructs" (p. 104).

Finally, O'Grady pointed out that if either or both of two theoretical constructs are determined by more than one causal agent, any estimates of explained variance will be limited to the maximum amount of variance that is actually shared between the two constructs. Since most psychological constructs are considered to be multiply determined, it follows that any measures of explained criticisms of measures of explained variance are similar to the limitations of the standardized mean difference that were identified by Hedges (1981). Essentially they amount to the point that measures of effect size depend upon the population of measurements.

APPLICATIONS TO META-ANALYSIS

As noted, one motivation for seeking to derive measures of effect size is to evaluate the results obtained across different studies in the research literature by means of the techniques of meta-analysis (Glass et al., 1981). Investigations of this sort have used measures based on comparisons of treatment means as well as estimates of the explained population variance. Rosenthal (1984, p. 23) noted, however, that most meta-analytic studies compare just two treatments at a time; thus measures of explained variance are rarely used (though see Hyde, 1981).

As Hedges and Becker (1986, p. 16) remarked, the estimate d' is well suited to this purpose because it is a directional measure whose sampling properties are fairly well understood. However, Hedges and Olkin (1985, pp. 101, 103) argued that estimates of explained population variance are inappropriate for combining the results of different studies because they are inherently nondirectional and hence can take on similar values for conflicting patterns of results. They cited a hypothetical situation in which two identical studies generated a difference between two treatment groups of 1 standard deviation in magnitude but in opposite directions. Clearly, all the measures of explained variance discussed earlier in this paper would yield identical values in the two experiments, suggesting the erroneous conclusion that the experiments had obtained the same results.

Whether such indices should in fact be used to average and to compare findings across different studies is quite another matter. Eysenck (1978) criticized techniques of meta-analysis on the grounds that they ignore the methodological adequacy of individual studies. As a result, pooled effect sizes may be influenced by design flaws as well as by treatment effects. Glass (1976) suggested, however, that "it is an empirical question whether relatively poorly designed studies give results significantly at variance with those of the best designed studies" (p. 4). On the basis of his own experience, he claimed that the difference is typically so small that to eliminate studies of poor quality would be to discard unnecessarily a large amount of important data. Hedges (1982c) similarly claimed that Eysenck's criticism can be resisted (although not decisively rebutted) within any particular application of meta-analysis via a demonstration that the obtained estimates of effect size are homogeneous across the set of studies available in the research literature.

Nonetheless, Linn and Petersen (1986) made the more subtle comment that "the research perspectives in a field influence what researchers study and constrain the possible outcomes from meta-analysis" (p. 69). Certainly, statistical techniques of whatever sophistication will not compensate for the preoccupations and biases of previous researchers. Indeed, computing average measures of effect size across the available research literature if anything tends to legitimate those preoccupations and biases. Be that as it may, meta-analysis represents merely one application of measures of effect size in psychological research, and it has not been the aim of this paper to argue whether or not it constitutes a useful research tool.

CONCLUSION

As Winer et al. (1991, p. 121) pointed out, an experimental design that achieves a numerically high level of statistical power can lead to the rejection of the null hypothesis even though the treatment effects are quite trivial from a practical or theoretical point of view. The measures of effect size described in this paper represent different attempts to evaluate the importance of the observed effects in a way that is independent of the level of statistical significance that they attain.

In designs with just two levels of a fixed factor, it is quite clear that the statistic d' defined by Hedges (1981) is the preferred measure of effect size. This measure represents the standardized mean difference between the two treatments, corrected for sampling bias. In the case of designs that contrast more than two levels of a fixed factor, there is no satisfactory analogous index of effect size. Instead, it is necessary to use an index of explained variance derived from the correlation ratio, such as Kelley's (1935) ε^2 or Hays's (1963, pp. 381–385) est. ω^2. Both of these indices incorporate a correction for sampling bias, and there is currently no principled basis for preferring one over the other. In the case of designs that contrast more than two levels of a random factor, the same conclusion holds for Kelley's (1935) ε^2 and Vaughan and Corballis's (1969) estimate of the population intraclass correlation coefficient.

Measures of effect size were developed partly to compare and evaluate results obtained across different studies in the research literature, but criticisms have been expressed by various authors regarding the weaknesses and limitations of meta-analytic techniques. However, these criticisms do not in themselves call into question the usefulness of measures of effect size in reporting or interpreting the findings obtained in single studies. Cohen (1965, p. 106) and Hays (1963, p. 328) recommended that researchers routinely report measures of effect size as well as test statistics and significance levels as a matter of good practice, but this is not of course to imply that such measures should be used uncritically.

Indeed, O'Grady (1982) commented that in research that is primarily concerned with understanding rather than with prediction, the theoretical importance of an effect may have more to do with its existence than with its magnitude. Chow (1988) argued more forcefully that in the context of theory corroboration, estimates of effect size may be largely irrelevant. Nevertheless, as Craig et al. (1976) observed, the important point is that measures of effect size are simply another part of the composite picture that a researcher builds when reporting data that indicate that one or more variables are helpful in understanding a particular behavior.

REFERENCES

Chow, S. L. (1988). Significance test or effect size? *Psychological Bulletin*, **103**, 105-110.

Cohen, J. (1965). Some statistical issues in psychological research. In B. B. Wolman (Ed.), *Handbook of clinical psychology* (pp. 95-121). New York: McGraw-Hill.

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences*. New York. Academic Press.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (rev. ed.). New York: Academic Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York: Academic Press.

Craig, J. R., Eison, C. L., & Metze, L. P. (1976). Significance tests and their interpretation: An example utilizing published research and ω^2. *Bulletin of the Psychonomic Society*, **7**, 280-282.

Cramér, H. (1946). *Mathematical methods of statistics*. Princeton, NJ: Princeton University Press.

Diamond, S. (1959). *Information and error: An introduction to statistical analysis*. New York. Basic Books.

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

Eysenck, H. J. (1978). An exercise in mega-silliness. *American Psychologist*, **33**, 517.

Fisher, R. A. (1922). The goodness of fit of regression formulae, and the distribution of regression coefficients. *Journal of the Royal Statistical Society*, **85**, 597-612.

Fisher, R. A. (1925). *Statistical methods for research workers*. Edinburgh. Oliver & Boyd.

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

Gibbons, J. D., Olkin, I., & Sobel, M. (1977). *Selecting and ordering populations. A new statistical methodology*. New York. Wiley.

GLASS, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, **5**(10), 3-8.

GLASS, G. V., & HAKSTIAN, A. R. (1969). Measures of association in comparative experiments: Their development and interpretation. *American Educational Research Journal*, **6**, 403-414.

GLASS, G. V., MCGAW, B., & SMITH, M. L. (1981). *Meta-analysis in social research*. Beverly Hills, CA: Sage.

HAYS, W. L. (1963). *Statistics*. New York: Holt, Rinehart & Winston.

HAYS, W. L. (1981). *Statistics* (3rd ed.). New York: Holt, Rinehart & Winston.

HEDGES, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, **6**, 107-128.

HEDGES, L. V. (1982a). Estimation of effect size from a series of independent experiments. *Psychological Bulletin*, **92**, 490-499.

HEDGES, L. V. (1982b). Fitting categorical models to effect sizes from a series of experiments. *Journal of Educational Statistics*, **7**, 119-137.

HEDGES, L. V. (1982c). Fitting continuous models to effect size data. *Journal of Educational Statistics*, **7**, 245-270.

HEDGES, L. V., & BECKER, B. J. (1986). Statistical methods in the meta-analysis of research on gender differences. In J. S. Hyde & M. C. Linn (Eds.), *The psychology of gender: Advances through meta-analysis* (pp. 14-50). Baltimore: Johns Hopkins University Press.

HEDGES, L. V., & OLKIN, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.

HYDE, J. S. (1981). How large are cognitive gender differences? A meta-analysis using ω^2 and *d*. *American Psychologist*, **36**, 892-901.

HYDE, J. S., FENNEMA, E., & LAMON, S. J. (1990). Gender differences in mathematics performance: A meta-analysis. *Psychological Bulletin*, **107**, 139-155.

HYDE, J. S., & LINN, M. C. (Eds.). (1986). *The psychology of gender: Advances through meta-analysis*. Baltimore: Johns Hopkins University Press.

KELLEY, T. L. (1935). An unbiased correlation ratio measure. *Proceedings of the National Academy of Sciences*, **21**, 554-559.

KENDALL, M., & STUART, A. (1977). *The advanced theory of statistics: Vol. 1. Distribution theory* (4th ed.). London: Charles Griffin.

KRAEMER, H. C. (1979). Tests of homogeneity of independent correlation coefficients. *Psychometrika*, **44**, 329-335.

KRAEMER, H. C. (1983). Theory of estimation and testing of effect sizes: Use in meta-analysis. *Journal of Educational Statistics*, **8**, 93-101.

KRAEMER, H. C., & ANDREWS, G. (1982). A nonparametric technique for meta-analysis effect size calculation. *Psychological Bulletin*, **91**, 404-412.

LEVIN, J. R. (1967). Misinterpreting the significance of "explained variation." *American Psychologist*, **22**, 675-676.

LINDQUIST, E. F. (1953). *Design and analysis of experiments in psychology and education*. Boston: Houghton Mifflin.

LINN, M. C., & PETERSEN, A. C. (1986). A meta-analysis of gender differences in spatial ability: Implications for mathematics and science achievement. In J. S. Hyde & M. C. Linn (Eds.), *The psychology of gender: Advances through meta-analysis* (pp. 67-101). Baltimore: Johns Hopkins University Press.

MCNEMAR, Q. (1962). *Psychological statistics* (3rd ed.). New York: Wiley.

NORTON, D. W., & LINDQUIST, E. F. (1951). Applications of experimental design and analysis. *Review of Educational Research*, **21**, 350-367.

O'GRADY, K. E. (1982). Measures of explained variance: Cautions and limitations. *Psychological Bulletin*, **92**, 766-777.

PEARSON, K. (1905). *Mathematical contributions to the theory of evolution. XIV. On the general theory of skew correlation and non-linear regression* (Drapers' Company Research Memoirs, Biometric Series II). London: Dulau.

PEARSON, K. (1923). On the correction necessary for the correlation ratio, η. *Biometrika*, **14**, 412-417.

PETERS, C. C., & VAN VOORHIS, W. R. (1940). *Statistical procedures and their mathematical bases*. New York: McGraw-Hill.

PRESBY, S. (1978). Overly broad categories obscure important differences between therapies. *American Psychologist*, **33**, 514-515.

ROSENTHAL, R. (1984). *Meta-analytic procedures for social research*. Beverly Hills, CA: Sage.

ROSENTHAL, R., & RUBIN, D. B. (1982). Comparing effect sizes of independent studies. *Psychological Bulletin*, **92**, 500-504.

SHROUT, P. E., & FLEISS, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, **86**, 420-428.

SMITH, M. L., & GLASS, G. V. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist*, **32**, 752-760.

VAUGHAN, G. M., & CORBALLIS, M. C. (1969). Beyond tests of significance: Estimating strength of effects in selected ANOVA designs. *Psychological Bulletin*, **72**, 204-213.

WHERRY, R. J. (1931). A new formula for predicting the shrinkage of the coefficient of multiple correlation. *Annals of Mathematical Statistics*, **2**, 440-457.

WILCOX, R. R. (1987). New designs in analysis of variance. *Annual Review of Psychology*, **38**, 29-60.

WINER, B. J. (1962). *Statistical principles in experimental design*. New York: McGraw-Hill.

WINER, B. J. (1971). *Statistical principles in experimental design* (2nd ed.). New York: McGraw-Hill.

WINER, B. J., BROWN, D. R., & MICHELS, K. M. (1991). *Statistical principles in experimental design* (3rd ed.). New York: McGraw-Hill.

WINKLER, R. L., & HAYS, W. L. (1975). *Statistics: Probability, inference, and decision* (2nd ed.). New York: Holt, Rinehart & Winston.

WISHART, J. (1932). A note on the distribution of the correlation ratio. *Biometrika*, **24**, 441-456.

NOTES

1. An estimate is *consistent* if it converges to the estimated value as the size of the sample increases. An estimate is *biased* if it tends to be either systematically larger than the estimated value or systematically smaller than the estimated value. Cramér (1946, pp. 254–255) showed that the ratio between two consistent, unbiased estimates was itself a consistent estimate of the ratio between the two estimated values. It is not an unbiased estimate of the latter quantity, however. In particular, if x and y are independent variables such that $x > 0$, then the expected value of the ratio y/x is greater than or equal to the ratio between their individual expected values. The latter inequality becomes an equality only when the distribution of the denominator is wholly concentrated at a single value or, in other words, when the denominator is actually a constant (Kendall & Stuart, 1977, p. 242).

2. Strictly speaking, this depends upon the usual assumptions that the sample means are normally distributed and that the sample variances are homogeneous. As will be discussed, the use of d' assumes homogeneity of variance, but it is a consistent and unbiased estimator of δ regardless of whether the assumption of normality is satisfied. More generally, issues concerning the robustness of statistical tests have little bearing on the value of particular estimates of effect size.

(Manuscript received April 18, 1994;
revision accepted for publication October 21, 1994.)