Journal of Counseling Psychology
2004, Vol. 51, No. 4, 473–481

0022-0167/04/$12.00 DOI: 10.1037/0022-0167.51.4.473

How to Estimate and Interpret Various Effect Sizes

Tammi Vacha-Haase
Colorado State University

Bruce Thompson
Texas A&M University and Baylor College of Medicine

The present article presents a tutorial on how to estimate and interpret various effect sizes. The 5th edition of the *Publication Manual* of the American Psychological Association (2001) described the failure to report effect sizes as a "defect" (p. 5), and 23 journals have published author guidelines requiring effect size reporting. Although dozens of effect size statistics have been available for some time, many researchers were trained at a time when effect sizes were not emphasized, or perhaps even taught. Consequently, some readers may appreciate a review of how to estimate and interpret various effect sizes. In addition to the tutorial, the authors recommend effect size interpretations that emphasize direct and explicit comparisons of effects in a new study with those reported in the prior related literature, with a focus on evaluating result replicability.

For decades, statistical significance has been the norm for evaluating results. In fact, little change has occurred since Carver (1993) noted: "A quick perusal of research journals, educational and psychological statistic textbooks, and doctoral dissertations will confirm that tests of statistical significance continue to dominate the interpretation of quantitative data in social science research" (p. 294).

Although statistical significance "evaluates the probability or likelihood of the *sample* results, given the sample size, and assuming that the sample came from a population in which the null hypothesis is exactly true" (Thompson, 2003, p. 7), statistical testing cannot evaluate result importance. Cohen (1994) observed that the statistical significance test "does not tell us what we want to know, and we so much want to know what we want to know that, out of desperation, we nevertheless believe that it does!" (p. 997). Some unlikely events are trivial and, conversely, some likely events are nevertheless quite important (Thompson, 1996).

However, the field of psychology appears to be moving in the direction of placing more emphasis on effect sizes, although progress has tended to be incremental (Fidler et al., in press). For example, the 5th edition of the American Psychological Association's (APA, 2001) *Publication Manual* emphasized that:

> it is almost always necessary to include some index of effect size or strength of relationship. . . . The general principle to be followed . . . is to provide the reader not only with information about statistical significance but also with enough information to assess the magnitude of the observed effect or relationship. (pp. 25–26)

The *Manual* also declared the failure to report effect sizes to be a "defect" (p. 5).

Today, given evidence that the *Manual* itself has only a limited impact on reporting practices (Finch, Thomason, & Cumming, 2002; Vacha-Haase, Nilsson, Reetz, Lance, & Thompson, 2000), editors of 23 journals have supplemented the *Manual* by publishing author guidelines explicitly requiring authors to report effect sizes (cf. Harris, 2003; Snyder, 2000; Trusty, Thompson, & Petrocelli, 2004). These include the flagship journals of two professional associations (the American Counseling Association and the Council for Exceptional Children), both with circulations greater than 55,000. As Fidler (2002) recently observed, "Of the major American associations, only all the journals of the American Educational Research Association have remained silent on all these issues" (p. 754).

However, as is often the case throughout statistics, there is not universal agreement regarding the definition of terms. Some people limit the term *effect size* to refer to a specific statistic (e.g., Cohen's *d*), others to a single class of statistics (e.g., standardized differences), or still others to a complete universe of statistics (i.e., all 40+ possible effect sizes).

For the purposes of this article, the term *effect size* is used most broadly to refer to any statistic that quantifies the degree to which sample results diverge from the expectations (e.g., no difference in group medians, no relationship between two variables) specified in the null hypothesis (Cohen, 1994; Thompson, 2002b, 2003). Effect sizes can be used to inform judgment regarding the "practical significance" of study results (Kirk, 1996).

If sample results match the expectations specified by the null hypothesis, the effect size is zero. For example, if the null is that the population postintervention medians of the experimental and the control groups are equal, and the two sample medians are 60.2 and 60.2, the effect size is zero. However, if a null hypothesis specifies that the population standard deviations of depression scores in three groups will be equal, and the sample standard deviations are 5.0, 5.5, and 9.0, the effect size is not zero. Greater divergence in sample results from the expectations specified in the null hypothesis regarding population parameters results in effect

Tammi Vacha-Haase, Department of Psychology, Colorado State University; Bruce Thompson, Department of Educational Psychology, Texas A&M University, and Department of Family and Community Medicine, Baylor College of Medicine.

Correspondence concerning this article should be addressed to Tammi Vacha-Haase, Department of Psychology, Colorado State University, Fort Collins, CO 80823-1876, or to Bruce Thompson, Department of Educational Psychology, Texas A&M University, College Station, TX 77843-4225. E-mail: tvh@lamar.colostate.edu or bruce-thompson@tamu.edu

sizes larger in magnitude. Huberty (2002) provided a history of the numerous effect size choices.

Like other statistics, effect sizes are *on the average* for the data set as a whole. For example, if a multiple correlation squared coefficient (R^2) is .50, this does not necessarily mean that every participant's *Y* score was predicted equally well. Indeed, there may be subsets of participants for whom the effect size would be larger and subsets for whom the effect size may be quite small. These dynamics suggest the importance of exploring data dynamics more deeply than merely computing summary statistics.

Purpose of the Article

The purpose of the present article was to provide a tutorial on how to calculate effect sizes. First, the three classes of effect size are discussed, providing guidance for the computation of each. Second, three guidelines for reporting effect sizes are suggested. Third, suggestions for interpreting effect sizes in applied research are presented. Finally, a heuristic example of an applied interpretation problem is presented.

Space considerations preclude coverage of every available choice, and we focus on the most commonly encountered effect sizes (Huberty, 2002; Kirk, 1996). Readers seeking more detail may consult Snyder and Lawson (1993) or Rosenthal (1994). Contemporary treatments are provided by Kirk (in press), Hill and Thompson (in press), and Thompson (in press). The book, *Beyond Significance Testing*, by Kline (2004) is especially contemporary and comprehensive.

Three Major Classes of Effect Sizes

Various frameworks for classifying the several dozen effect size statistics have been presented (cf. Kirk, 1996; Thompson, 2002a). Here, three major categories are discussed: (a) standardized differences effect sizes, (b) variance-accounted-for effect sizes, and (c) "corrected" effect sizes. Corrected effect sizes attempt to better estimate either population or future sample effects by removing from the results the estimated influences of sample idiosyncrasies.

Standardized differences. In medicine, there are natural, universally accepted metrics with which to measure intervention effects. For example, cholesterol is universally measured in milligrams per deciliter. Studies of medications given to heart attack patients routinely compare number of deaths of the placebo group versus deaths in the postintervention experimental group. Unstandardized differences (e.g., 10 fewer deaths per 100 treated patients) in such studies are reasonable effect estimates, and indeed any other choices might be deemed thoughtless.

In psychology, however, there are no universal metrics with which to measure abstract constructs such as self-concept or IQ, and the metrics of measures are arbitrarily fixed by different test authors. Furthermore, different measures of the same construct may have different metrics (e.g., one IQ measure has a standard deviation of 15, whereas another has a standard deviation of 16, and a new measure might be developed with a standard deviation of 10).

To compare effects across an entire literature in which different researchers have used different measures of the outcome variable, researchers must resort to using standardized differences so that the effect sizes may be compared "apples to apples." Standardized differences are computed by the generic formula:

$$\{M_E - M_C\}/SD, \quad (1)$$

where M_E is the posttest sample mean of the experimental group, M_C is the posttest sample mean of the control group, and *SD* is some estimate of the population standard deviation. Effect sizes may also be computed using sample medians or other sample central tendency statistics, but sample group means are used most commonly.

Standardized differences express effect size in standard deviation units. Thus, if the effect is .20, the treatment group mean is one fifth of a standard deviation higher than the *control group* mean. If the outcome scores in both groups are normally distributed, the standardized difference can be readily expressed in terms of percentage of group overlap. For example, if 1,000 participants are in both groups, and the standardized difference is .20, then 85%, or 1,700, of the participants in the two groups overlap as regards their outcome scores, whereas 300 have scores outside the overlap of the two distributions (see Howell, 2002, p. 228).

The standard deviation used in the standardization can be estimated in different ways, and of course, the choice can make a large difference. For example, as part of his articulation of meta-analytic methods, Glass (1976) proposed delta, which divides the sample mean difference by the sample standard deviation of the control group. Alternatively, Cohen's (1969) *d* invokes a standard deviation estimate that "pools" or averages the outcome variable's sample standard deviations across *both* the intervention and the control group.

In an intervention study of the effects of a year of taking an herbal supplement on measured IQ, the intervention group (n_E = 50) might enjoy an advantage on the average of 3.0 IQ points ($[M_E = 103] - [M_C = 100] = 3.0$) over the control group (n_C = 50). For this situation in which group sizes are equal (i.e., both *n*s = 50), we estimate effect using the pooled standard deviation across the two groups by pooling the variances (i.e., SD^2):

$$d = \{M_E - M_C\}/\{\mathrm{SQRT}[(SD_E^2 + SD_C^2)/2]\}. \quad (2)$$

If the sample standard deviations of the outcome variable scores for the experimental and control groups were SD_E = 19.0 and SD_C = 15.0, respectively, for these data, we obtain,

$$\begin{aligned} d &= \{103 - 100\}/\{\mathrm{SQRT}[(19.0^2 + 15.0^2)/2]\} \\ &= 3.0/\{\mathrm{SQRT}[(361.0 + 225.0)/2]\} \\ &= 3.0/\{\mathrm{SQRT}[586.0/2]\} \\ &= 3.0/\{\mathrm{SQRT}[293.0]\} \\ &= 3.0/17.117 \\ &= .175. \end{aligned} \quad (3)$$

Delta is computed as $(M_E - M_C)/SD_C$, and for the same data, delta would be 0.200 (3.0/15.0 = 0.200).

A comparison of these two from among the several standardized difference choices makes clear the importance of thoughtful selection of effect indices. There are not always definitively correct single right answers in statistics. As Huberty and Morris (1988, p.

573) emphasized: "As in all statistical inference, subjective judgment cannot be avoided. Neither can reasonableness!"

In this example, the pooled estimate (SD = 17.117) has the appeal that the estimate is based on a sample size of 100 rather than of 50. However, if one believes that the population standard deviation (σ) is 15.0 or 16.0, the finding that SD_E is 19.0 suggests the possibility that the intervention impacted both the posttest sample mean and the posttest sample standard deviation of scores in the intervention group. Therefore, on balance, the researcher may reasonably prefer the use of delta over d for these data because the control group's standard deviation could not have been impacted by the intervention.

Researchers will generally prefer delta when the sample size in the control group is very large so that relatively little precision in estimating the standard deviation is gained by pooling both samples. Researchers also may prefer delta when the intervention and control conditions themselves are hugely different, because pooling, again, makes less sense in this context (Rosenthal, 1994). For example, it may be reasonable to pool across interventions involving 10 weekly therapy sessions of 50 versus 60 min in length, but it may not be reasonable to pool across 10 weekly sessions versus 2 years of daily therapy. Of course, the selection makes less difference as the groups' standard deviations are more similar.

This discussion also reinforces the notion of why standardizing can be important. An unstandardized mean difference of 3.0 in the prior example translated into a standardized difference effect size of .175 or .200 IQ standard deviations. If the 3.0 mean difference had instead involved measurement of body temperatures in Fahrenheit, perhaps with a standard deviation of 0.2, the standardized difference effect becomes 15.0 (i.e., 3.0/0.2), reflecting the fact that an increase of 3 IQ points indicates that the intervention group in the first example is perhaps somewhat smarter, but the intervention group in the health example has become on the average very ill indeed. Therefore, researchers must attend to the nature of the outcome variable (i.e., whether higher or lower scores are desired) as well as the mean difference in relation to the standard deviation.

Variance-accounted-for indices. In statistics, the general linear model (GLM) is the idea that all commonly used analyses are part of a single analytic family (Cohen, 1968; Thompson, 1991, 2000). All GLM analyses (e.g., t tests, analysis of variance [ANOVA], analysis of covariance [ANCOVA], regression, factor analysis, descriptive discriminant analysis) are correlational and explicitly or implicitly apply weights (e.g., beta weights, factor pattern coefficients) to measured variables to obtain scores on composite or latent variables (e.g., regression or ANOVA $\hat{Y}$ scores, factor scores). One implication of the GLM is that an r^2-type effect size can be computed in all the commonly used analyses.

For example, in bivariate correlation or in regression analyses, the sum of squares (SOS) of the $\hat{Y}$ scores can be computed. This is synonymously called the "$SOS_{EXPLAINED}$," "SOS_{MODEL}," "$SOS_{REGRESSION}$," or "SOS_{WITHIN}." In fact, to confuse everybody, different procedures within a given statistics package often name these synonymous results with different labels.

The SOS of the criterion variable quantifies the amount and the origins of information available, as regards the criterion variable. For example, for a given data set, if SOS_Y (also called SOS_{TOTAL}) is 50.0, there is more information regarding individual differences on the outcome variable than if SOS_Y had been 10.0. We can partition this SOS in various ways to explore the origins of this information. First, we can partition the information on the basis of who generated the information. "Wendy" may have generated 30.0 of the 50 units of information about individual differences, "Deborah" 15.0 units, and "Susan" 5.0 units.

Second, we can also partition this information on the basis of how well other variables explain or predict these individual differences. For example, in regression, if this SOS is 50.0, and the $SOS_{EXPLAINED}$ is 10.0, the R^2 is .20 (10.0/50.0), or 20%. With knowledge of the scores on the predictor variables, we can explain 20% of the individual differences on the criterion variable.

Similarly, given the GLM, if in either an ANOVA or a t test the SOS_Y is 50.0 and the $SOS_{EXPLAINED}$ is 10.0, the variance-accounted-for effect size is 0.20 (10.0/50.0 = 0.20), or 20%. Here, with knowledge of group membership on the independent variable, we can explain 20% of the individual differences on the criterion variable. However, unlike R^2, which only measures linear relationship (unless predictor variables are raised to exponential powers), the ANOVA variance-accounted-for effect size is sensitive to various kinds of relationship, including nonlinear relationship. Therefore, to distinguish this ANOVA effect index from the R^2, the ANOVA effect is named the *correlation ratio*, or eta squared (η^2).

Given the GLM, analogous variance-accounted-for effect sizes can be computed for multivariate analyses. Multivariate eta squared can be computed for multivariate analysis of variance (MANOVA) and descriptive discriminant analysis. The squared canonical correlation coefficient, R_C^2, can be computed for canonical correlation analysis.

Analogous variance-accounted-for effect sizes can also be computed for categorical data such as contingency tables (e.g., counts of heart attacks or not across taking aspirin daily vs. a placebo). However, variance-accounted-for effect sizes for such data are difficult to interpret with respect to practical significance. For example, the r^2 effect for aspirin on heart attack is only .0011, even though people taking aspirin daily have 4% fewer heart attacks (Rosenthal, 1994). Therefore, for categorical data, effect sizes such as the binomial effect size display or odds ratios are recommended. Fliess (1994) provided a summary of the available choices.

Corrected effect sizes. All classical statistical analyses (e.g., t tests, ANOVA, descriptive discriminant analysis) are least squares methods that maximize the sample $SOS_{EXPLAINED}$ and minimize the sample $SOS_{UNEXPLAINED}$. This has the concurrent impact of maximizing the sample effect size.

The problem is that parts of the SOS_Y and of the $SOS_{EXPLAINED}$ are unique to a given sample and do not exist in the population and will not exist in the same form in other future samples. In other words, in essence, every sample has its own "personality" or uniqueness (i.e., sampling error variance), and some samples have more sampling error variance than others. And when the SOS_Y is 50.0, and the $SOS_{EXPLAINED}$ is 10.0, the classical analysis does not consider how much of the 50.0 and of the 10.0 are unique to a particular sample, and thus does not consider how much of the effect is real or replicable. Therefore, the effect size for a sample tends to overestimate the effect size both in the population and in future samples (Snyder & Lawson, 1993).

We can adjust or correct the sample effect size if we can successfully estimate the amount of sampling error variance in the sample data, and then remove this influence from the effect size.

Only if we collect data from the entire population (or the population effect size is perfect) will the sample effect size not be inflated by sampling error.

We know that three design elements cause sampling error. More sampling error variance (and thus more positively biased sample effect estimates) occurs when (a) the sample size is smaller, (b) the number of measured variables is larger, and (c) the population effect size is smaller. Because the third element is unknown, or the study would not be done in the first place, we use the sample effect size as the best estimate of the unknown population parameter. Numerous correction formulae can be used to adjust effect estimates using these three design features.

In regression, the Ezekiel (1930) formula is often used. The corrected or "adjusted" multiple correlation squared coefficient, R^2, can be computed as:

$$1 - \{[n - 1]/[n - v - 1]\} \times \{1 - R^2\}, \quad (4)$$

where n is the sample size, v is the number of predictor variables, and R^2 is the uncorrected squared multiple correlation coefficient. The formula can be equivalently expressed as:

$$R^2 - \{[1 - R^2] \times [v/n - v - 1)]\}. \quad (5)$$

The Appendix presents some corrections for three illustrative values of the sample R^2 (i.e., .95, .50, and .05), three sample sizes (i.e., 120, 30, and 5), and three numbers of predictor variables (i.e., 12, 6, and 1). Note that the difference between the R^2 and the "adjusted R^2" (R^2*), which is called *shrinkage*, is small when the R^2 is large, even when the sample size is quite small (e.g., 5) or the number of predictors is fairly large (e.g., 6 for the sample size of 30).

Note also that R^2* can be computed to be negative even though this statistic is a squared variance-accounted-for result. Of course, such a result, although mathematically possible, indicates serious design problems. This result is analogous to obtaining an estimated reliability coefficient, such as Cronbach's alpha, that is negative (e.g., $\alpha = -.7$, or even $\alpha = -7.5$), even though alpha is also inherently in a squared variance-accounted-for metric (Thompson, 2003).

The negative R^2* suggests fundamental problems. If R^2* equals a negative number, this means that with knowledge of the predictor variable scores, one believes he or she can explain less than zero of the variability of the outcome variable scores, or less than the zero variability that he or she could explain using no predictors.

In the ANOVA case, the analogous omega squared can be computed using the formula from Hays (1981, p. 349):

$$\{SOS_{BETWEEN} - [(k - 1) \times MS_{WITHIN}]\}/\{SOS_{TOTAL} + MS_{WITHIN}\}, \quad (6)$$

where SOS_{TOTAL} is the SOS of the dependent variable, $SOS_{BETWEEN}$ is the SOS between, MS_{WITHIN} is the mean square within, and k is the number of groups. For example, if n was 31, $SOS_Y = 100.0$, there were $k = 7$ groups, and the correlation ratio (η^2) was .50, or 50%, the ANOVA summary table would be:

SOS	*df*	*MS*	*F*	η^2
50.0	6	8.33	4.00	0.50
50.0	24	2.08		
100.0	30			

The omega squared would equal:

$$\{50.0 - [6 \times 2.08]\}/\{100.0 + 2.08\}$$

$$\{50.0 - 12.5\}/102.08$$

$$37.5/102.08 = 0.367. \quad (7)$$

So, from the original η^2 of 0.50, or 50%, the shrunken effect size (ω^2) would be 0.367, or 36.7%, with a shrinkage of 0.133 (i.e., 50.0% − 36.7% = 13.3%), or a decrease of 26.5% (i.e., 13.3%/50.0% = 26.5%) in the original, uncorrected estimate.

Correction formulae also are available for multivariate analyses. For example, in MANOVA and descriptive discriminant analysis, a multivariate omega squared from Tatsuoka (1973) can be used. And some simulation research suggests that the Ezekiel (1930) correction formula also may be used with the squared canonical correlation coefficient (Thompson, 1990).

Should corrected effect sizes be preferred over uncorrected estimates? Because corrected estimates are generally more accurate estimates of population effects or effects likely to be encountered in replication (Snyder & Lawson, 1993), a reasonable argument can be made that corrected estimates should be used more often than their uncorrected counterparts. But it will make less difference which class of estimates is used if (a) sample size is very large, (b) the number of variables is very small, and (c) the unknown population effect size is in reality quite large. Because it is not entirely clear when a sample becomes large, or the number of variables becomes small, the prudent researcher probably does draw more often from the well of corrected effects than from the well of uncorrected effect sizes.

Effect Sizes in SPSS

Some brief comment on obtaining effect sizes in each of the three classes (i.e., standardized differences, variance-accounted-for, and corrected or adjusted) using computer software is warranted. Discussion is limited to the SPSS package. Of course, the features of SPSS are continually being modified, so these comments may be time bound.

SPSS does not yield standardized difference effect sizes (e.g., *d*, delta). However, the computations are fairly simple and can be implemented using either a spreadsheet program or a calculator, using the formulas presented earlier.

In regression, SPSS always outputs both the uncorrected and the corrected R^2. If one has a bivariate problem (i.e., Pearson r), the corrected value can be obtained by running the analysis using the REGRESSION procedure rather than the CORRELATION procedure. In ANOVA, SPSS will provide eta squared upon request.

Table 1 presents some common analytic methods and displays how effect sizes can be obtained for each. Again, for more detail on these different estimates, the reader is referred to Snyder and Lawson (1993), Rosenthal (1994), Hill and Thompson (in press), Kirk (in press), Thompson (in press), or Kline's (2004) book, *Beyond Significance Testing*.

Three Reporting Rules

Here we recommend three guidelines for reporting effect sizes. First, *explicitly say exactly what effect sizes are being reported*. As

Table 1
Strategies for Obtaining Effect Sizes for Selected SPSS Analyses

Analysis	Possible strategy
Contingency table (r or odds ratio)	Run the CROSSTABS procedure and select the desired effect from the STATISTICS submenu.
Independent t test (d, η^2, or ω^2)	Compute a Cohen's d by hand. Or, run the analysis as a one-way ANOVA using the GLM program; click on the OPTION requesting an effect size to obtain η^2. Use the Hay's correction formula (ω^2) if an adjusted estimate is desired.
ANOVA (η^2 or ω^2)	Run the analysis as an ANOVA using the GLM program; click on the OPTION requesting an effect size to obtain η^2. Use the Hay's correction formula by hand if an adjusted estimate is desired.
Regression (R^2 or R^2*)	Run the REGRESSION procedure. Both the uncorrected R^2 and the corrected variance accounted for (R^2*) estimates are displayed, by default.
MANOVA (multivariate η^2 or ω^2)	Run the analysis as a MANOVA using the GLM program; click on the OPTION requesting an effect size to obtain η^2. A corrected estimate, multivariate ω^2, (Tatsuoka, 1973), can be computed by hand.
Descriptive discriminant analysis (multivariate η^2 or ω^2)	Run the analysis as a MANOVA using the GLM program; click on the OPTION requesting an effect size to obtain η^2. A corrected estimate, multivariate ω^2 (Tatsuoka, 1973), can be computed by hand.
Canonical correlation analysis (R_c^2 or R_c^2*)	Run the analysis in the MANOVA procedure using the syntax suggested by Thompson (2000). The R_c^2 is reported. Apply the Ezekiel correction by hand if a corrected value (R_c^2*) is desired.

Note. ANOVA = analysis of variance; GLM = general linear model; MANOVA = multivariate analysis of variance.

noted previously, there are dozens of effect size choices. The reader must know specifically which effect is being reported in order to evaluate the effect. However, empirical studies of reporting practices indicate that in a surprising number of studies, authors report effects without clearly indicating exactly which effect is being presented (Kieffer, Reese, & Thompson, 2001; Vacha-Haase et al., 2000).

Only reporting that an effect size is .5 without indicating which effect size is being presented is inherently ambiguous. A Cohen's d of .5 is very different than an r^2 of .5, just as an r of .50 is in a different metric than an r^2 of .50. The reader cannot intelligently evaluate the effect size if the estimate being reported is not clearly stated.

Specificity also allows readers to use conversion formulae to reexpress effects in new metrics. If some previous authors reported Pearson r values, and other authors reported Cohen's d values, a researcher could reexpress all the effect sizes as rs, or as ds, so that the summary is completely apples-to-apples.

For example, an r can be converted to a d using Friedman's (1968, p. 246) Formula 6:

$$d = \{2[r]\}/\{[1 - r^2]^{.5}\}. \quad (8)$$

Conversely, a Cohen's d can be converted to an r using Cohen's (1988, p. 23) approximation Formula 2.2.6:

$$r = d/\{[d^2 + 4]^{.5}\}. \quad (9)$$

For example, if $d = .5$, r would equal approximately:

$$\begin{aligned} &= .5/\{[.5^2 + 4]^{.5}\} \\ &= .5/\{[.25 + 4]^{.5}\} \\ &= .5/\{4.25^{.5}\} \\ &= .5/2.061552 \\ &= 0.242. \end{aligned} \quad (10)$$

When total sample size is small or group sizes are quite disparate, it is advisable to use a slightly more complicated but more precise conversion formula provided by Aaron, Kromrey, and Ferron (1998).

For heuristic purposes, let us also convert d to r using the more precise formula, and assuming both groups have a sample size of 50. Now r is estimated to be:

$$\begin{aligned} &= .5/\{.5^2 + [(100^2 - 2(100)/(50 \times 50)]^{.5}\} \\ &= .5/\{.25 + [(100^2 - 2(100)/(50 \times 50)]^{.5}\} \\ &= .5/\{.25 + [(10{,}000 - 200)/2{,}500]^{.5}\} \\ &= .5/\{.25 + [9{,}800/2{,}500]^{.5}\} \\ &= .5/\{.25 + [3.92]^{.5}\} \\ &= .5/\{.25 + 1.980\} \\ &= .5/2.230 \\ &= .224. \end{aligned} \quad (11)$$

Note that the approximation yields a result (.242) close to the more precise conversion (.224). Of course, once the r is derived, a variance-accounted-for effect size, if desired, can be obtained simply by squaring r.

Second, *interpret effect sizes by taking into consideration both their assumptions and their limitations*. As with other statistics, when analytic assumptions are violated, results are compromised. If distribution or homogeneity assumptions are severely violated, F and p calculated values may be compromised, but so too will be the effect estimates. However, some Monte Carlo research suggests that some newer effect indices, such as group overlap I indices, may be more robust to the violations of methodological assumptions (cf. Hess, Olejnik, & Huberty, 2001; Huberty & Holmes, 1983; Huberty & Lowman, 2000). Thus, when methodological assumptions are not well met, the I effect sizes may be preferred.

It is also important to compare effect sizes across studies, taking into account design differences. For example, as Olejnik and Algina (2000) pointed out, the eta squared in a fixed-effects

ANOVA study involving five therapy sessions of intervention targeting depression should not be expected to equal the eta squared in a similar intervention involving 20 therapy sessions. Effects across such disparate studies should not be compared apples-to-apples without taking into account intervention differences. In other words, effect sizes are not magically independent of the designs that created them.

Third, for various reasons summarized by Wilkinson and Task Force on Statistical Inference (1999), *report confidence intervals for effect sizes and other study results*. One reason for recommending confidence intervals is that these intervals are readily amenable to graphic presentation, thus allowing a large number of effects across studies to be compared in an economical manner (Thompson, 2002b). Figures graphing confidence intervals can be easily prepared in Excel using the Stock Chart menu and inputting the two interval endpoints and the point estimate (e.g., the mean or the effect size) as "High," "Low," and "Close," respectively.

A second reason for recommending confidence intervals is that the widths of intervals can be compared to evaluate the precision of the estimates in a given study or a given literature. A third reason for preferring intervals is that the consultation of intervals across studies will eventually lead to an accurate estimate of parameters even if a priori expectations are wildly wrong (Schmidt, 1996).

Confidence intervals for parameter estimates such as means, medians, or standard deviations can be computed using formulae. However, confidence intervals for effect sizes cannot be computed using formulae and instead must be estimated using iterative statistical methods normally requiring specialized software. Fleishman (1980) presented the basic theory underlying these estimates.

Cumming and Finch (2001) provided an excellent tutorial on estimating confidence intervals for effect sizes. Free software is available to generate the estimates (cf. Algina & Keselman, 2003; Smithson, 2001; Steiger & Fouladi, 1992). Cumming, Williams, and Fidler (in press) explain how confidence intervals can be used to inform judgments about result replicability.

Suggestions for Interpreting Effect Sizes

In his various books on power analysis (cf. Cohen, 1968), although he hesitated to do so, Jacob Cohen proposed some tentative benchmarks for what might be deemed small (e.g., $d = |.2|$, $\eta^2 \cong 1\%$), medium (e.g., $d = |.5|$, $\eta^2 \cong 10\%$), and large (e.g., $d = |.8|$, $\eta^2 \cong 25\%$) effects, as regards the typicality of results across the entirety of the social science literature. He felt that people would be less likely to focus on effect sizes absent such benchmarks. Indeed, Kirk (1996) argued that one reason why effect sizes have gained increasing popularity is that Cohen provided this interpretive framework.

But as Thompson (2001) noted, "if people interpreted effect sizes [using fixed benchmarks] with the same rigidity that $\alpha = .05$ has been used in statistical testing, we would merely be being stupid in another metric" (pp. 82–83). One-size-fits-all rules of thumb are not always very helpful in interpreting effect sizes, as Prentice and Miller (1992) pointed out in their article, "When small effects are impressive." The context of the study (e.g., whether the outcome is life or death, whether other effective interventions already exist) necessarily impacts the evaluation of effect sizes.

For example, the eta squared effect size for smoking versus not smoking on longevity is around 2%. We deem the result quite noteworthy, first because the outcome is so precious and, second, because related study after study has replicated this approximate effect.

In general, we believe that effect sizes should be interpreted by (a) considering what outcome is being studied and (b) *directly* and *explicitly* comparing effects with those in related prior studies, and not by rigidly invoking Cohen's benchmarks for small, medium, and large effects. It is exactly by directly and explicitly comparing the effects in the study with those in the related prior studies that the replicability of results can be evaluated. As Cohen (1994) explained in some detail, statistical significance tests do not evaluate result replicability. That is why direct comparisons of effects across studies are so critical. It is exactly through these comparisons, and not by statistical testing, that the serendipitous or anomalous result is detected.

One caveat must be emphasized, however. When researchers are conducting groundbreaking research in areas of inquiry involving no or few previous studies, effect sizes cannot be evaluated in the context of related prior effects. In such situations, the use of Cohen's benchmarks is then more appropriate.

In Howell's (2002) words: "We should not make too much of Cohen's levels, but they are helpful as a rough guide" (p. 206), at least in new areas of inquiry. But in more established areas of research, "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, McGaw, & Smith, 1981, p. 104).

An Applied Example

Consider a hypothetical literature addressing whether counseling psychologists, on average, are happier than the general population. The hypothetical measurement has a normative sample mean of 50 and standard deviation of 10. Let us presume that, diagnostically, people are flagged as having clinically noteworthy happiness if $X_i > 60.0$.

Table 2 presents the hypothetical universe of prior studies. We compare interpretations from (a) a published literature selected from available studies only if statistical significance is achieved, (b) "vote counting" of statistically significant results, (c) consult-

Table 2
Hypothetical Literature Consisting of Nine Studies of the Happiness of Counseling Psychologists

Study	*M*	*SD*	*n*	*p*	*d*
1	64.9	11.4	21	.063	.430
2	68.0	10.9	9	.059	.734
3	63.8	8.7	22	.053	.437
4	62.3	11.2	88	.057	.205
5	67.5	9.9	8	.069	.758
6	64.0	9.1	21	.058	.440
7	65.3	11.1	21	.041	.477[a]
8	63.6	13.2	63	.034	.273[a]
9	58.0	9.7	96	.046	−.206[a]

[a] In a literature plagued by the "file drawer" problem (Rosenthal, 1979), only these studies would be published.

ing effect sizes (here Cohen's *d*) for all studies, and (d) confidence intervals graphically reported for the literature as a whole.

First, if only statistically significant studies were admitted to the published literature, Studies 1–6 would be subjected to the "file drawer" problem (i.e., nonsignificant results are filed and not even submitted for publication) and would perish (Greenwald, 1975; Rosenthal, 1979). Such a literature has not been informed by Rosnow and Rosenthal's (1989) admonition that "surely, God loves the .06 [level of statistical significance] nearly as much as the .05" (p. 1277).

The problem with a literature driven solely by statistical significance tests is that although researchers use small alpha levels, some Type I errors will inevitably occur across a large literature. These are then afforded priority for publication. As Thompson (1996, p. 28) explained: "This is problematic in the context of a bias against reporting results that are not statistically significant, 'because investigators generally cannot get their failures to replicate published, [and so] Type I errors, once made, are very difficult to correct'" (Clark, 1976, p. 258).

Second, if vote counting (i.e., counts of significant vs. nonsignificant studies) was used to interpret the literature, the vote would be mixed, 3 to 6. Such mixed results are not uncommon and occur partly because, as a discipline, we conduct our studies with stunningly low power. As Schmidt and Hunter (1997) noted, average power

> in typical studies and research literatures is in the .40 to .60 range. . . . That is, in a research area in which there really is a difference or relation, when the significance test is used to determine whether findings are real or just chance events, the null hypothesis significance test will provide an erroneous answer about 50% of the time. This level of accuracy is so low that it could be achieved just by flipping a (unbiased) coin! (p. 40)

One problem with vote counting, even if power were not such a serious problem, is that vote counting does not tell all that is of importance to know. As Roger Kirk (1996) explained:

> . . . a rejection means that the researcher is pretty sure of the direction of the difference. Is that any way to develop psychological theory? I think not. How far would physics have progressed if their researchers had focused on discovering ordinal relationships? What we want to know is the size of the difference between *A* and *B* and the error associated with our estimate; knowing that *A* is greater than *B* is not enough. (p. 754)

Third, if effect sizes were reported for all hypotheses, we would know that *d* ranged from −.206 to .758. We would know that eight of the nine studies had positive *d* values, which bears directly on result replicability. We would know that the weighted average *d* across the literature was .189, or roughly 0.2 standard deviations. This average (.189) was computed by weighting each sample *d* by study sample size; a sophisticated alternative is to create a pooled estimate by weighting by effect variance (see Cohn & Becker, 2004).

If we believe that happiness is normally distributed, about 84.13% of the population has happiness less than our gold standard person with a happiness score of 60.0 (i.e., 1 standard deviation above the normative sample mean of 50.0). The 349 counseling psychologists in the hypothetical nine prior studies are roughly 0.2 standard deviations (i.e., the weighted average *d*) above 60.0.

Whether this average effect size (or any other) is noteworthy is a value judgment. The difference may be noteworthy if we care a lot about counseling psychologists and how happy they are. Such value judgments are inescapable in research. And remember that statistical significance testing also cannot make these value judgments for us.

The tabled results also emphasize the importance of reporting effect sizes for all results, including those that are statistically nonsignificant. Note that the weighted average *d* for all nine studies was .189, whereas *d* for the three statistically significant studies was .041.

Not reporting effect sizes for nonsignificant results is the same as treating these effects as zero, which amounts to the fallacy of "accepting" the null hypothesis. As Wilkinson and Task Force on Statistical Inference (1999) emphasized: "Never use the unfortunate expression 'accept the null hypothesis.' *Always* [italics added] provide some effect-size estimate when reporting a *p* value" (p. 599).

Fourth, confidence intervals might be reported for the parameter of interest (here, *M*) or for the related effect size (here *d*). Figure 1 graphically presents the 95% confidence intervals about the nine sample means.

What do we gain from such a presentation? Confidence intervals tell us about the precision of our estimate or of the literature. Every value within a given interval is considered a plausible estimate in a given context. Such intervals help us to see that our studies are being conducted with too little precision.

Such confidence intervals also help us to interpret our study in direct and explicit comparison with results in prior studies. If the ninth study was our research, we would not be sanguine about overinterpreting our results. Our mean is anomalously low. The nine intervals have an average lower bound of about 59 and an average upper bound of around 69. The comparison would force us to reflect on what may have been different in our study versus the prior eight studies.

Such comparisons do evaluate result replicability. Statistical significance tests do *not* evaluate result replicability (Cohen, 1994; Thompson, 1996). Such comparisons facilitate the "meta-analytic thinking" so important in good science (Cumming & Finch, 2001). For example, the Figure 1 results suggest that (a) precisions vary considerably across studies, as reflected in the variable widths of the intervals; (b) the means themselves are somewhat variable; but (c) the means tend to fall within the range of about 60 to about 70 (i.e., 1–2 standard deviations above the normative mean).

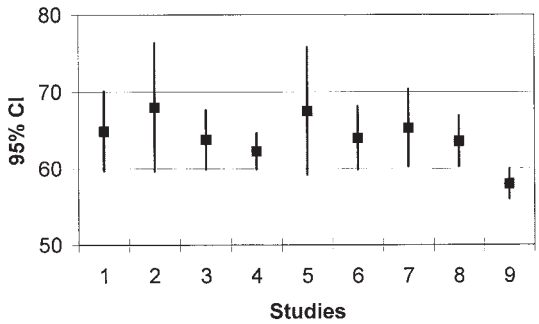

Figure 1. Confidence intervals (CI; 95%) of means from nine studies. Solid squares = mean.

Conclusions

Although it is important to report effect sizes for all primary results, regardless of the outcome of statistical tests, there may not be one particularly preferred effect index across all studies. The field is still learning the pluses and minuses of different choices, and new effect indices are being developed (cf. Hess et al., 2001).

Tracey (2000) suggested that presenting effect size is "easier for most people to grasp" and the "presentation of effect sizes and confidence intervals would make for easy translation into meta-analytic studies" (p. 183). Of most help will be interpretations that emphasize *direct and explicit* comparisons of effects in a new study with those reported in the prior related literature, with a focus on evaluating result replicability.

References

Aaron, B., Kromrey, J. D., & Ferron, J. M. (1998, November). *Equating r-based and d-based effect size indices: Problems with a commonly recommended formula*. Paper presented at the annual meeting of the Florida Educational Research Association, Orlando, FL.

Algina, J., & Keselman, H. J. (2003). Approximate confidence intervals for effect sizes. *Educational and Psychological Measurement, 63,* 537–553.

American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: Author.

Carver, R. (1993). The case against statistical significance testing, revisited. *Journal of Experimental Education, 61,* 287–292.

Clark, H. H. (1976). Reply to Wike and Church. *Journal of Verbal Learning and Verbal Behavior, 15,* 257–261.

Cohen, J. (1968). Multiple regression as a general data-analytic system. *Psychological Bulletin, 70,* 426–443.

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillside, NJ: Erlbaum.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49,* 997–1003.

Cohn, L. D., & Becker, B. J. (2004). How meta-analysis increases statistical power. *Psychological Methods, 8,* 243–253.

Cumming, G., & Finch, S. (2001). A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement, 61,* 532–574.

Cumming, G., Williams, J., & Fidler, F. (in press). Replication, and researchers' understanding of confidence intervals and standard error bars. *Understanding Statistics*.

Ezekiel, M. (1930). *Methods of correlational analysis*. New York: Wiley.

Fidler, F. (2002). The fifth edition of the APA publication manual: Why its statistics recommendations are so controversial. *Educational and Psychological Measurement, 62,* 749–770.

Fidler, F., Cumming, G., Thomason, N., Pannuzzo, D., Smith, J., Fyffe, P., et al. (in press). Toward improved statistical reporting in the Journal of Consulting and Clinical Psychology. *Journal of Consulting and Clinical Psychology*.

Finch, S., Thomason, N., & Cumming, G. (2002). Past and future American Psychological Association guidelines for statistical practice. *Theory and Psychology, 12,* 825–853.

Fleishman, A. I. (1980). Confidence intervals for correlation ratios. *Educational and Psychological Measurement, 40,* 659–670.

Fliess, J. L. (1994). Measures of effect size for categorical data. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 245–260). New York: Russell Sage Foundation.

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

Glass, G. V, McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research*. Beverly Hills, CA: Sage.

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin, 82,* 1–20.

Harris, K. (2003). Journal of Educational Psychology: Instructions for authors. *Journal of Educational Psychology, 95,* 201.

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

Hess, B., Olejnik, S., & Huberty, C. J. (2001). The efficacy of two improvement-over-chance effect sizes for two-group univariate comparisons under variance heterogeneity and nonnormality. *Educational and Psychological Measurement, 61,* 909–936.

Hill, C. R., & Thompson, B. (in press). Computing and interpreting effect sizes. In J. C. Smart (Ed.), *Higher education: Handbook of theory and research* (Vol. 19). New York: Kluwer Academic.

Howell, D. C. (2002). *Statistical methods for psychology* (5th ed.). Pacific Grove, CA: Duxbury.

Huberty, C. J (2002). A history of effect size indices. *Educational and Psychological Measurement, 62,* 227–240.

Huberty, C. J, & Holmes, S. E. (1983). Two-group comparisons and univariate classification. *Educational and Psychological Measurement, 43,* 15–26.

Huberty, C. J, & Lowman, L. L. (2000). Group overlap as a basis for effect size. *Educational and Psychological Measurement, 60,* 543–563.

Huberty, C. J, & Morris, J. D. (1988). A single contrast test procedure. *Educational and Psychological Measurement, 48,* 567–578.

Kieffer, K. M., Reese, R. J., & Thompson, B. (2001). Statistical techniques employed in *AERJ* and *JCP* articles from 1988 to 1997: A methodological review. *Journal of Experimental Education, 69,* 280–309.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56,* 746–759.

Kirk, R. E. (in press). The importance of effect magnitude. In S. F. Davis (Ed.), *Handbook of research methods in experimental psychology*. Oxford, England: Blackwell.

Kline, R. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research*. Washington, DC: American Psychological Association.

Olejnik, S., & Algina, J. (2000). Measures of effect size for comparative studies: Applications, interpretations, and limitations. *Contemporary Educational Psychology, 25,* 241–286.

Prentice, D. A., & Miller, D. T. (1992). When small effects are impressive. *Psychological Bulletin, 112,* 160–164.

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin, 86,* 638–641.

Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 231–244). New York: Russell Sage Foundation.

Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist, 44,* 1276–1284.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for the training of researchers. *Psychological Methods, 1,* 115–129.

Schmidt, F. L., & Hunter, J. E. (1997). Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 37–64). Mahwah, NJ: Erlbaum.

Smithson, M. (2001). Correct confidence intervals for various regression effect sizes and parameters: The importance of noncentral distributions in computing intervals. *Educational and Psychological Measurement, 61,* 605–632.

Snyder, P. (2000). Guidelines for reporting results of group quantitative investigations. *Journal of Early Intervention, 23,* 145–150.

Snyder, P., & Lawson, S. (1993). Evaluating results using corrected and uncorrected effect size estimates. *Journal of Experimental Education, 61,* 334–349.

Steiger, J. H., & Fouladi, R. T. (1992). R^2: A computer program for interval estimation, power calculation, and hypothesis testing for the squared multiple correlation. *Behavior Research Methods, Instruments, and Computers, 4,* 581–582.

Tatsuoka, M. M. (1973). *An examination of the statistical properties of a multivariate measure of strength of relationship*. Urbana: University of Illinois. (ERIC Document Reproduction Service No. ED099406)

Thompson, B. (1990). Finding a correction for the sampling error in multivariate measures of relationship: A Monte Carlo study. *Educational and Psychological Measurement, 50,* 15–31.

Thompson, B. (1991). A primer on the logic and use of canonical correlation analysis. *Measurement and Evaluation in Counseling and Development, 24,* 80–95.

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher, 25*(2), 26–30.

Thompson, B. (2000). Canonical correlation analysis. In L. Grimm & P. Yarnold (Eds.), *Reading and understanding more multivariate statistics* (pp. 285–316). Washington, DC: American Psychological Association.

Thompson, B. (2001). Significance, effect sizes, stepwise methods, and other issues: Strong arguments move the field. *Journal of Experimental Education, 70,* 80–93.

Thompson, B. (2002a). "Statistical," "practical," and "clinical": How many kinds of significance do counselors need to consider? *Journal of Counseling & Development, 80,* 64–71.

Thompson, B. (2002b). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher, 31*(3), 24–31.

Thompson, B. (Ed.). (2003). *Score reliability: Contemporary thinking on reliability issues*. Newbury Park, CA: Sage.

Thompson, B. (in press). Research synthesis: Effect sizes. In J. Green, G. Camilli, & P. B. Elmore (Eds.), *Complementary methods for research in education*. Washington, DC: American Educational Research Association.

Tracey, T. J. G. (2000). Issues in the analysis and interpretation of quantitative data: Deinstitutionalization of the null hypothesis test. In S. D. Brown & R. Lent (Eds.), *Handbook of counseling psychology* (3rd ed., pp. 177–198). New York: Wiley.

Trusty, J., Thompson, B., & Petrocelli, J. V. (2004). Practical guide to implementing the requirement of reporting effect size in quantitative research in the *Journal of Counseling & Development*. *Journal of Counseling & Development, 82,* 107–110.

Vacha-Haase, T., Nilsson, J. E., Reetz, D. R., Lance, T. S., & Thompson, B. (2000). Reporting practices and APA editorial policies regarding statistical significance and effect size. *Theory & Psychology, 10,* 413–425.

Wilkinson, L., & Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594–604.

Appendix

Computation of Illustrative R^2 and Adjusted R^2 (R^2) Values*

$1 - \{[n - 1] / [n - v - 1] * [1 - R^2]\} = R^2$*	Shrinkage	% Shrinkage
$1 - \{[120 - 1] / [120 - 12 - 1] * [1 - .95]\} = .944$	.006	.59
$1 - \{[30 - 1] / [30 - 6 - 1] * [1 - .95]\} = .937$	.013	1.37
$1 - \{[5 - 1] / [5 - 1 - 1] * [1 - .95]\} = .933$	.017	1.75
$1 - \{[120 - 1] / [120 - 12 - 1] * [1 - .50]\} = .444$	.056	11.21
$1 - \{[30 - 1] / [30 - 6 - 1] * [1 - .50]\} = .370$	.130	26.09
$1 - \{[5 - 1] / [5 - 1 - 1] * [1 - .50]\} = .333$	.167	33.33
$1 - \{[120 - 1] / [120 - 12 - 1] * [1 - .05]\} = -.057$	.107	213.08
$1 - \{[30 - 1] / [30 - 6 - 1] * [1 - .05]\} = -.198$	.248	495.65
$1 - \{[5 - 1] / [5 - 1 - 1] * [1 - .05]\} = -.267$	.317	633.33

Note. n = the sample size; v = the number of predictor variables; R^2 = the uncorrected multiple correlation squared coefficient; and R^2* = the "corrected" or "adjusted" multiple correlation squared coefficient.

Received August 18, 2003
Revision received March 8, 2004
Accepted April 26, 2004 ■