Effect-Size Measures and Meta-Analytic Thinking in Counseling Psychology Research

Robin K. Henson

The Counseling Psychologist 2006; 34; 601
DOI: 10.1177/0011000005283558

The online version of this article can be found at:
http://tcp.sagepub.com/cgi/content/abstract/34/5/601
Effect sizes are critical to result interpretation and synthesis across studies. Although statistical significance testing has historically dominated the determination of result importance, modern views emphasize the role of effect sizes and confidence intervals. This article accessibly discusses how to calculate and interpret the effect sizes that counseling psychologists use most frequently. To provide context, the author presents a brief history of statistical significance tests. Second, the author discusses the difference between statistical, practical, and clinical significance. Third, the author reviews and graphically demonstrates two common types of effect sizes, commenting on multivariate and corrected effect sizes. Fourth, the author emphasizes meta-analytic thinking and the potential role of confidence intervals around effect sizes. Finally, the author gives a hypothetical example of how to report and potentially interpret some effect sizes.

For the past 50 or so years, statistical significance testing has dominated how psychologists have evaluated research outcomes. For example, Hubbard and Ryan (2000) observed that the null hypothesis significance test (NHST) was the analytic approach of choice in a historical review of 12 American Psychological Association (APA) journals. Historical use notwithstanding, researchers often misunderstand or misinterpret NHSTs (Nelson, Rosenthal, & Rosnow, 1986; Rosenthal & Gaito, 1963; Zuckerman, Hodgins, Zuckerman, & Rosenthal, 1993). As such, and due at least in part to the inherent limitations of NHSTs, increasing numbers of researchers have become aware of the value of effect sizes in result interpretation. Indeed, the editorial policies of at least 23 journals now require that authors report effect sizes.

Although researchers have debated the relative advantages and limitations of NHSTs for almost as long as the method has existed, there is growing inertia toward de-emphasizing NHSTs and emphasizing effect sizes and confidence intervals (CIs) when interpreting results. In a comprehensive review, Anderson, Burnham, and Thompson (2000) reported a dramatic rise in articles that question the use of NHSTs across a range of disciplines.
More specific to counseling psychology, several important articles have addressed the role of effect sizes (see, e.g., Trusty, Thompson, & Petrocelli, 2004; Vacha-Haase & Thompson, 2004; Wilkinson & Task Force on Statistical Inference [TFSI], 1999). This article intends (a) to build on these (and other) works by providing additional examples of effect-size calculation and use and (b) to extend these works by providing a conceptual and graphical framework around the meaning of several common effect sizes, along with a hypothetical example of how to write up some results in a journal article. Reviews of authors’ reporting practices have amply demonstrated the slow implementation of effect sizes in the literature (cf. Keselman et al., 1998; Vacha-Haase, Nilsson, Reetz, Lance, & Thompson, 2000). Continued education and explication are warranted if effect sizes are to be fruitfully employed on a regular basis.

**Purpose**

Of course, interpretation of effect sizes, unless accurate and duly informed, can lead to a whole other set of problems. Therefore, this article’s purpose is to discuss accessibly how to calculate and interpret the effect sizes that counseling psychologists use most frequently. To place this purpose in context, I first provide a brief history of NHSTs. Second, I address the difference between statistical, practical, and clinical significance. Third, I review and graphically demonstrate two common types of effect sizes, commenting on multivariate and corrected effect sizes. Fourth, I emphasize meta-analytic thinking and the potential role of CIs around effect sizes (see also Quintana & Minami, in press). Finally, I give a hypothetical example of how to report and potentially interpret some effect sizes. Although this article cannot review the complete set of issues surrounding the appropriate use of all effect sizes, I hope that it will serve as a resource for future and additional study as needed.

For the purposes of this article, I employ a fairly general definition of *effect size*. Taking the lead of Vacha-Haase and Thompson (2004), I define *effect size* as 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” (p. 473). If the null suggests that three group means are the same, for example, an effect size may quantify the degree of difference between the three means or, perhaps, the magnitude of differences between each of the three pairwise combinations of means. This general approach is important because it allows a variety of ways to quantify the divergence of sample results from the null. The appropriate effect size to use depends on the context of the study, the audience, and other factors.
A BRIEF HISTORY OF NHSTs

In 1996, the APA convened the TFSI to address the controversy surrounding the role of NHSTs in research inquiry. The TFSI expanded its charge and released a report that dealt with a range of issues critical to conducting good research (Wilkinson & TFSI, 1999). The TFSI did not support an outright ban on NHSTs but noted, “It is hard to imagine a situation in which a dichotomous accept-reject decision is better than reporting an actual $p$-value or, better still, a confidence interval. . . . Always provide some effect-size estimate when reporting a $p$-value” (Wilkinson & TFSI, 1999, p. 599, emphasis added).

Subsequently, the fifth and most recent edition of the APA Publication Manual (2001) now emphasizes the following:

For the reader to fully understand the importance of your findings, it is almost always necessary [italics added] to include some index of effect size or strength of relationship in your Results section. . . . 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)

NHSTs have been so ingrained in our psychological research history, training, and curriculum that the casual observer may not be fully aware of the controversies surrounding statistical significance tests. These tests and their corresponding $p$ values are, after all, what many learned in their doctoral training, often yearning earnestly for at least one $p < .05$ in their dissertations. The proclivity toward NHSTs has since been strongly reinforced by our research literature, which is riddled with $p$ values and biased toward publication of statistically significant results.

There has nevertheless been much debate regarding NHSTs’ relative usefulness for many decades, but the dialogue has escalated over the years (Anderson et al., 2000). For example, Jacob Cohen’s (1994) seminal American Psychologist article “The Earth Is Round ($p < .05$)” and his work in power analysis were particularly effective at bringing the misuse of NHSTs to the forefront. In addition, entire journal issues (McLean & Ernest, 1998) and books (Harlow, Mulaik, & Steiger, 1997) have been devoted to the debate.

Some decade and a half ago, Harris (1991) metaphorically observed,

There has been a long and honorable tradition of blistering attacks on the role of statistical significance testing in the behavioral sciences, a tradition reminiscent of knights in shining armor bravely marching off, one by one, to slay a rather large and stubborn dragon. . . . Given the cogency, vehemence and repetition of such attacks, it is surprising to see that the dragon will not stay dead. (p. 375)
In spite of many thoughtful attempts to reveal misuses and dispel misunderstandings concerning this dragon, NHSTs have remained a feature of the research literature. Although space does not allow a historical discussion of the reasons for the entrenchment of NHSTs (for a complete discussion of NHST history, see Huberty, 1993; Huberty & Pike, 1999), two important points are worth noting. First, attempts at methodology remediation and encouragements to report effect sizes have often been ineffective. In the fourth edition of the APA Publication Manual (1994), authors are only “encouraged to provide effect-size information” (p. 18). Empirical studies have documented that this encouragement has changed the literature little (Vacha-Haase et al., 2000).

Second, there is now growing awareness regarding the need for information beyond or instead of NHSTs for result interpretation. The very fact that an article on effect size appears in this important special issue of The Counseling Psychologist is evidence of the field’s evolution. Many researchers have advanced the field by placing research focus on effect sizes and CIs and by emphasizing meta-analytic thinking (Cumming & Finch, 2005; Kline, 2004; Quintana & Minami, in press [TCP special issue, part 2]; Smithson, 2000; Thompson, 2002; Tryon, 2001).

Misunderstandings and Misuses of NHSTs

Put simply, a NHST gives “the probability (0 to 1.0) of the sample statistics, given the sample size, and assuming the sample was derived from a population in which the null hypothesis (H_o) is exactly true” (Thompson, 1996, p. 27). A statistical significance test does not speak to result importance, replicability, or even the probability that a result was due to chance (Carver, 1978), in spite of many researchers’ misconceptions to the contrary (see Kline, 2004, for an excellent review of NHST misconceptions and appropriate uses). As Cohen (1994) observed, an NHST “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). Furthermore, Kline (2004) recently suggested mistaken beliefs about what statistical tests tell us act as a collective form of cognitive distortion that has hindered research progress in the behavioral sciences. This is apparent by the failure to develop a stronger tradition of replication compared to the natural sciences, a lack of relevance of much of our research, and wasted research effort and resources. (p. 10)

In sum, the way psychologists report and interpret research is changing. More and more, they are considering effect sizes as critical for effective research interpretation, and single $p$ values are often of little value in the grand research scheme.
STATISTICAL, PRACTICAL, AND CLINICAL SIGNIFICANCE

Suppose a psychologist is studying the effects of a certain cognitive treatment on major depression. The psychologist has randomly assigned clients, all of whom had received an external diagnosis, to a treatment group ($n = 8$) and a waiting-list control group ($n = 8$). If the results on a depression measure were (as $T$ scores, where high scores reflect increased depressive symptoms) $M_{\text{treat}} = 62.00$ ($SD_{\text{treat}} = 2.24$) and $M_{\text{control}} = 64.00$ ($SD_{\text{control}} = 2.32$), then an independent samples $t$ test would yield $t(14) = -1.752$, $p = .051$ (a one-tailed test was due to the anticipation of a reduction in symptomology).

This statistically nonsignificant result (employing $\alpha = .05$) may lead the rigid NHST observer to conclude that the treatment was ineffective in reducing depression scores relative to the control group. The less rigid observer may declare that the results were approaching significance. Unfortunately, this language is ambiguous because the researcher does not know whether the results were actually “trying to avoid being statistically significant” (Thompson, 1994, p. 6)—that is, if the study were conducted again, the $p$ value would not necessarily be smaller; it very well might be larger.

If we assume the same $M$s and $SD$s and simply add one more client to each group, the obtained $p$ value would be a statistically significant .049 (again, $\alpha = .05$). In this case, it would not be uncommon to read statements about the treatment’s effectiveness in reducing depression.

The bottom line is that $p$ values actually provide limited information in and of themselves. In the hypothetical case, replicating the original example many times over, with each study yielding $p = .051$ or so, the thoughtful researcher would unlikely dismiss the results. Here, the interpretive emphasis would be on result stability and replication rather than the alpha level. In such a case, “surely, God loves the .06 nearly as much as the .05” (Rosnow & Rosenthal, 1989, p. 1277).

What is missing from the above example is some index of how big the group differences were, not just whether there were differences from a statistical significance standpoint. We need to know, that is, whether there is any practical significance to the findings (Kirk, 1996). Even miniscule group differences may be statistically significant at some sample size, but are they big enough to matter? Effect sizes provide one avenue for evaluating practical significance.

Furthermore, there also may be implications regarding clinical significance, which can be distinguished from both statistical and practical significance. For example, in the above example, some of the clients in the
treatment group may have had sufficiently reduced symptoms to warrant a diagnosis change, or their scores on the depression measure may have moved below a clinical cutoff. In this case, the results may indeed have important clinical implications, regardless of the $p$ value and even the magnitude of the mean difference.

In the $t$ test above, for example, we may observe a statistically significant mean difference and even generate an effect measure for how big this difference is. However, there very well may be some people in the treatment group who did not improve at all relative to the control, and therefore, the treatment may be clinically significant for some but not for all.

The concept of *clinical significance* is in no way new (see, e.g., Jacobson, Follette, & Revenstorf, 1984; Kazdin, 1977; Nietzel & Trull, 1988), although, similar to effect sizes, its use tends to be somewhat limited. It is nevertheless important to distinguish it from the role of effect sizes. As Campbell (2005) noted,

> Effect sizes do not directly indicate to us what proportion of individuals have improved or recovered as a result of the intervention (Jacobson, Roberts, Berns, & McGlinchey, 1999). Two studies with equal effect sizes (e.g., Cohen’s $d$) might differ in their clinical significance (i.e., the proportion of children who have improved or no longer need intervention). (p. 214)

Approaches to assessing clinical significance tend to focus on whether (a) individuals’ outcome scores undergo a meaningful change or (b) treatment participants’ outcomes are indistinguishable from so-called normal individuals on the construct or symptom of interest (Kendall, Marrs-Garcia, Nath, & Sheldrick, 1999). Jacobson et al. (1999; Jacobson & Truax, 1991) work with the reliable change index has probably been the most common approach to dealing with meaningful change in scores. Kendall et al. (1999; Kendall & Grove, 1988) have demonstrated the normative comparison approach to clinical significance.

A full discussion of clinical significance is beyond the scope of this article, but interested readers may consult the above works for more detailed information, including some specific examples of employing the reliable change index and conducting normative comparisons (see also Kazdin, 1999). Although not always, in some respects clinical significance may relate more to individuals, whereas both NHSTs and effects sizes tend to be average-level statistics. Regardless, clinical significance necessarily requires the presence of meaningful normative information or diagnostic criteria of some sort. I turn now to some effect-size options as measures of practical significance.
EFFECT SIZE AND PRACTICAL SIGNIFICANCE

There are dozens of effect-size indices available to researchers, each with relative strengths and weaknesses for particular purposes (see, e.g., Huberty, 2002; Kirk, 1996). However, the literature actually uses much fewer effect-size measures.

Of course, limited use need not imply that those indices that do enjoy more recognition are somehow better than those that seldom appear. The use and accuracy of an effect-size statistic may depend on many factors (e.g., distributional assumptions, congruence with the research question being asked). This review will nevertheless focus on the most commonly used indices because it is perhaps best to begin at a point of familiarity.

For the most part, one can categorize the dozens of available effect-size indices into two groups: standardized mean differences and measures of association (sometimes called variance-accounted-for measures; Kirk, 1996; Maxwell & Delaney, 1990; Snyder & Lawson, 1993). In the following sections, I discuss two standardized mean difference effects, Cohen’s $d$ and Glass’s delta, and provide some guidance on interpreting their magnitude. I also review two variance-accounted-for effects, $R^2$ for nongrouped data and $\eta^2$ for grouped data, along with interpretive guidance. Finally, I present several multivariate analogs of these effect sizes and introduce two corrected effect sizes. I also refer readers to Kline’s (2004) book Beyond Significance Testing, which is a particularly excellent resource on modern approaches to research inquiry (for additional approaches to testing multivariate models using structural equation modeling, see Martens & Hasse, in press [TCP special issue, part 2]; Weston & Gore, 2006 [this issue]).

Standardized Mean Differences

Returning to our original depression treatment example, we still do not have much clarity on how big the group difference is. We, of course, know that there is a two-point difference between the means, but because we are unaware of the scaling, this is largely useless information. It is often the case in psychology that the scaling used for particular measures is somewhat arbitrary. Some other disciplines, such as medicine, do not suffer from this ambiguity to the same degree, where some things (e.g., blood pressure, cholesterol levels) are measured on commonly accepted metrics (Vacha-Haase & Thompson, 2004).

How one researcher measures the quality of the working alliance, for example, necessarily results in an arbitrary scale, which may even vary from other measures of the same construct (Hoyt, Warbasse, & Chu, in press [TCP special issue, part 2]). The Working Alliance Inventory (Horvath
Greenberg, 1989) yields scores ranging from 36 to 252, whereas the California Psychotherapy Alliance Scale (Gaston & Marmar, 1994) yields scores ranging from 24 to 168, both of which differ from the brief form of the Working Alliance Inventory (range = 12 to 84; Tracey & Kokotovic, 1989). This type of arbitrary scaling makes using standardized measures of mean differences all the more important in psychology and the social sciences (on scale-development issues, see Helms, Henze, Sass, & Mifsud, 2006 [this issue]; Worthington & Whittaker, in press [TCP special issue, part 2]).

Standardization can come in different ways, but it typically involves dividing the mean difference by some measure of score variability (much like a z score). We could, for example, divide our two-point mean difference by the standard deviation of the control group: \( \frac{62.00 - 64.00}{2.32} = -0.86 \), which would indicate that the mean of the treatment group was .86 standard deviations lower than the mean of the control. A change of almost one standard deviation seems substantially different from the null (where the change would be 0 and the means would be the same).

If, however, national T score norms were available on the depression measure (we gave our original means as T score averages and would expect that norms would be accessible), then it may be more appropriate to use the T score standard deviation of 10 in the division, as it would more accurately reflect population variability, presumably because the normative sample would be much larger than the current sample. In this case, \( \frac{62.00 - 64.00}{10} = -0.20 \), indicating a much smaller mean shift in standard deviation units, and the interpretation difference between the –.89 and –.20 would likely be substantial. The point here is heuristic—that is, when discussing effect sizes, it is critical that the researcher is clear on (a) the kind of effect used, (b) the meaning of the effect, and (c) the data elements used in computing the statistic. It is a matter of research clarity that readers and editors should expect.

Standardized mean differences also allow researchers to compare intervention or other effects across studies that use different or arbitrary variable scaling. This comparison is very important when placing one’s findings in the context of the broader literature. Such as in the earlier example, because standardized differences involve division by a measure of variability, we interpret these statistics in terms of standard deviation units.

**Cohen’s d.** Probably the most commonly employed standardized mean difference effect is Cohen’s \( d \) (Cohen, 1969). Indeed, Cohen’s (1988) now classic and highly influential work on power analysis is largely centered on his \( d \) statistic. To compute \( d \), one uses a pooled estimate of the treatment and control standard deviations in the division. This pooled estimate would be the standard deviation one would get if the scores in both groups were
combined. When sample sizes for both groups are equal, Cohen’s $d$ can be computed as

$$d = \frac{M_{\text{treat}} - M_{\text{control}}}{\sqrt{SD_{\text{treat}}^2 + SD_{\text{control}}^2 / 2}},$$

(1)

using just the $M$s and $SD$s. The denominator of this formula averages the variances (i.e., the square of the $SD$s) and then converts this average into a pooled $SD$ via the square root.

For example, if a psychologist investigating self-concept had $M_{\text{treat}} = 24$ ($SD_{\text{treat}} = 8$) and $M_{\text{control}} = 21$ ($SD_{\text{control}} = 6$), we find,

$$d = (24 - 21) / \{\text{SQRT}[(8^2 + 6^2)/2]\}$$
$$= 3.0 / \{\text{SQRT}[(64 + 36)/2]\}$$
$$= 3.0 / \{\text{SQRT}[100/2]\}$$
$$= 3.0 / \{\text{SQRT}[50]\}$$
$$= 3.0 / 7.071$$
$$= .424.$$

This indicates that the treatment group, on average, scored just over two fifths (.424 to be exact) of a standard deviation higher than the control group. In the context of a given study, this may or may not be meaningful, and it is up to the researcher to determine this in light of other literature in the area.

Glass’s delta. Glass (1976) proposed delta ($\Delta$) as part of his seminal work in meta-analysis. Delta, such as $d$, involves division by a standard deviation but with using only the standard deviation of the control group (as with our original depression example). The logic is simple. There may be some cases when the treatment itself may affect not only the treatment mean but also the treatment standard deviation. For example, a particular assertiveness intervention may be useful for increasing assertiveness in part, but not all, of the treatment group. The treatment group mean would increase but so would the standard deviation due to the differential treatment effect within the group. Using the same data for the self-concept example above, delta would be

$$\text{delta} = (24 - 21) / 6$$
$$= .50,$$

(2)

which would indicate that the self-concept mean in the treatment group was half a standard deviation above the control-group mean.
A Couple of Caveats

Two caveats are worth noting here. First, selecting what effect size to use is important because the results can vary as $d$ and delta demonstrated above. Therefore, the researcher should be clear not only regarding exactly what effect size he or she used but also on the data features that affect the statistic. Reporting a mean difference effect of .30 simply does not give the reader enough information. The researcher must define this effect (e.g., Cohen’s $d$, Glass’s delta, etc.), maybe explain it (e.g., “divides by the control group $SD$”), and definitely interpret it within the context of the study (e.g., “The .30 effect indicates almost a one third $SD$ improvement for the treatment group, which is consistent with the average effect observed in prior related literature”). Note as well that other standardized mean difference effects are available, such as Hedges’s $g$, but $d$ and delta are the most common.

Choosing between $d$ and delta depends on the nature of the data and the purpose of the research. In the original depression example, we computed dramatically different effects (−.89 and −.20, respectively) when using the control group $SD = 2.32$ versus the known population $SD = 10$. In this case, the control group $SD$ varied considerably from what one might expect in the population, even though there was no intervention. Once we consider, however, that the sample consisted of clients with diagnosed depression (a relatively homogeneous subgroup of the population), this difference is perhaps less surprising. If the researcher wished to speak to the impact of the treatment as regards the specific control group used, then perhaps the obtained $SD$s would be appropriate. If the researcher wished to place the results in the context of the national norm scaling, then maybe he or she could employ the population $SD$. Similarly, because the treatment and control $SD$s (2.24 and 2.32, respectively) were similar, the researcher may opt to pool the $SD$s and employ Cohen’s $d$ because the effect would benefit from an $SD$ estimate from a larger sample size.

Second, the assertiveness example above, in which the treatment differentially affected the intervention group, highlights the fact that most effect sizes are summary statistics and are “on average”. They therefore do not speak directly to individual outcomes. Some clients may have improved many standard deviation units; others may have experienced little change. The astute researcher needs to keep this in mind when interpreting effect sizes. If we fail to adequately understand what our effect sizes do and do not tell us, then we may fall victim to new misconceptions about our research methods.
Interpreting Magnitude of Standardized Mean Difference Effects

In his work on power analysis, Cohen (1988) hesitantly presented several rules of thumb for interpreting a $d$ effect as small (.20), medium (.50), and large (.80). Many readers are probably familiar with these benchmarks via observing authors justify their effect sizes with such criteria. There is nevertheless considerable risk in using Cohen’s benchmarks within the context of any given study. As Thompson (2001) poignantly explained, “If people interpreted effect sizes 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). Similarly, Kline (2004) noted that unthoughtfully using these guidelines can be “fraught with . . . potential problems” (p. 132).

Others have discussed reasons that these benchmarks should not be rigidly used (e.g., Fern & Monroe, 1996; Kirk, 1996; Vacha-Haase & Thompson, 2004). Essentially, the “demonstration of an effect’s significance, whether theoretical, practical or clinical, calls for more discipline-specific expertise than the estimation of its magnitude (Kirk, 1996)” (Kline, 2004, p. 134). Effect sizes are best interpreted relative to (a) the context of the current study and (b) the size and nature of effects in prior studies. I will consider (b) again later.

It is perhaps not surprising that in the absence of other guidance, some researchers would employ Cohen’s (1988) benchmarks. This nevertheless is a black-box process, possibly due in part to a lack of conceptual understanding about the meaning of the effect sizes. However, one can also conceptualize $d$ as a percentage of group overlap (or nonoverlap), which may be more meaningful for some. Figures 1 through 3 provide a frame of reference for interpreting $d$ effects of .2, .5, and .8, respectively. The figures make some assumptions, including the facts that the two distributions are normal with homogeneous variances ($SD_{\text{treat}} = SD_{\text{control}} = 10$) and are of equal sample sizes. The effect sizes are based on the control group’s having a mean of 50 in each case.

Conceptually, larger effect sizes would result in less group overlap and, thus, more distinction between the groups (Huberty & Lowman, 2000). Under the above assumptions, if $d$ were 0.0, the two distributions would be identical. Figure 1 results when $d = .2$, and about 85.3% of the scores in the two distributions overlap, leaving 14.7% of nonoverlap (Cohen, 1988). Graphically, these two distributions clearly have more in common than otherwise, and it is “approximately the size of the difference in mean height between 15- and 16-year-old girls” (Cohen, 1988, p. 26).

A so-called medium effect (again, under the same assumptions) of .5 is demonstrated in Figure 2, where 67.0% of the area is overlapping. Because the distributions overlap less, we expect a larger effect. Cohen (1988) suggested
FIGURE 1. Illustration of \( d = .2 \times \frac{(52 - 50)}{10} \) for Two Groups (\( SD_{\text{exper}} = SD_{\text{control}} = 10 \))
NOTE: The shaded portion represents 85.3% of the total area for both distributions.

FIGURE 2. Illustration of \( d = .5 \times \frac{(55 - 50)}{10} \) for Two Groups (\( SD_{\text{exper}} = SD_{\text{control}} = 10 \))
NOTE: The shaded portion represents 67.0% of the total area for both distributions.
that a medium effect should “be visible to the naked eye” (p. 26) and compared it to about the “average difference in mean IQ between clerical and semiskilled workers or between members of professional and managerial occupational groups” (p. 26), citing the now somewhat dated Super (1949). Finally, Figure 3 illustrates $d = .8$. Here, a little more than half (52.6%) of the entire area overlaps. Cohen (1988) likened this effect to the “mean IQ difference estimated between holders of the Ph.D. degree and typical college freshmen” (p. 27). As a PhD holder, I personally am comfortable with labeling this effect in this context as large.

Of course, in other contexts, these varied effects may or may not be practically or clinically significant. Furthermore, in the real world, our data seldom perfectly meet the assumptions that are the bases for these graphics (and the percentages of overlap), thereby potentially affecting interpretations. In sum, there simply is no substitute for discipline-specific evaluation of effects in the context of those found in prior literature (Kirk, 1996; Thompson, 2002).

### Variance-Accounted-For Measures

Some readers may be aware that most of our classical parametric analyses are part of the general linear model (GLM) family. Being familial, all

![Figure 3. Illustration of $d = .8 [(58 - 50) / 10]$ for Two Groups ($SD_{\text{exper}} = SD_{\text{control}} = 10$)]
GLM analyses (a) are correlational in nature, (b) yield $r^2$-type effect sizes, and (c) maximize shared variance between variables or sets of variables (Bagozzi, Fornell, & Larcker, 1981; Cohen, 1968; Henson, 2000; Knapp, 1978; Thompson, 1991). As Sherry and Henson (2005) noted, “The GLM provides a framework for understanding all classical analyses in terms of the simple Pearson $r$ correlation” (p. 37).

To be more specific, we typically are interested in predicting or explaining some dependent/criterion/outcome variable (assuming for now a univariate analysis, such as ANOVA or regression). As such, our GLM analyses partition the dependent variable’s variance into that which can be explained and that which cannot. Often, we use sum of squares ($SS$) to represent the differing amounts of our variance partitions.

Figure 4 uses Venn diagrams to illustrate a multiple regression with two uncorrelated predictors. In this example, let us assume that the $SS_Y$ (criterion variable) is 100 and that both the $SS_{X_1}$ and the $SS_{X_2}$ are 50. Notice that the Venn diagram can exactly represent the proportionate amount of variance in each variable, such that the areas for both predictors are the same size and $Y$ is twice as large.

From this diagram, it is clear that 50% of the total criterion variable variance can be accounted for by the two predictors (25% by each)—more...
specifically, $SS_{explained} = 50$ (often called $SS_{regression}$ or $SS_{model}$) and $SS_{unexplained} = 50$ (often called $SS_{error}$ or $SS_{within}$). With knowledge of the variance explained in a criterion variable, it is rather straightforward to compute a variance-accounted-for effect size, which here would be $R^2 = (SS_{explained} / SS_Y) = (50/100) = .50$. As the diagram demonstrated, this effect would inform us that the two predictors can explain 50% of the variance in the dependent variable, which is directly analogous to an $r^2 = .50$ from a squared bivariate correlation.

The general formula of $(SS_{explained} / SS_Y)$ holds throughout the GLM in various forms. For example, in ANOVA, the $SS$ due to differences between the group means ($SS_{between}$) can be divided by the $SS_Y$ to yield the $\eta^2$ (eta squared) effect size (see later on).

**Nongrouped data.** Of the variance-accounted-for effects for nongrouped data (i.e., samples where focus is on relationships between variables rather than differences between groups), readers are probably most familiar with $R^2$, due to its association with multiple regression. Indeed, Kirk (1996) observed that this effect size is one of the most often reported but also noted that this is likely because $R^2$ is routinely given in the output of statistical software. It is interesting to note that although analogous effect sizes can be reported throughout the GLM, reporting rates for other types of effects are much lower (Keselman et al., 1998; Kirk, 1996), perhaps at least in part due to the fact that statistical software output often does not give them.

Researchers can readily compute $R^2$ as explained above, but as an additional example, consider a study in which the researcher is investigating whether three family history variables (number of immediate family members, empathy of mother, and age of mother at birth) were predictive of secure adult attachment (all variables are continuously scaled). The three predictor variables would unlikely be perfectly uncorrelated, and therefore, each may be able to explain some of the same portions of the secure-attachment variance. Regardless of the combination, the sum of the unique and shared portions of dependent variable that can be explained would yield the $R^2$ effect (excluding possible suppressor variables, which are outside the scope of this article). Because the predictor variables can explain some of the same variance, the sum of the three individual $r^2$s between each predictor and secure attachment would be larger than the $R^2$. In such a case, it would be important to examine both standardized weights and structure coefficients to evaluate each variable’s contribution to the effect size (Courville & Thompson, 2001; Henson, 2002).

**Grouped data.** Researchers can also compute variance-accounted-for effects when evaluating group differences (most often mean differences, but
other comparisons are possible, such as median differences). If comparing only two means, then a standardized mean difference is probably most relevant. When comparing three or more means, the most common variance-accounted-for effect is $\eta^2$ (eta squared), which represents the proportion of dependent-variable variance due to differences among the group means ($SS_{between}/SS_{total}$). Holding all else constant (e.g., sample size, group variances), as group means become more different, $\eta^2$ will increase. If $\eta^2$ equals 0, then the group means are equal. Unfortunately, researchers report $\eta^2$ much less often than its $R^2$ analog in regression, at least partially because statistical software packages tend to not automatically include $\eta^2$ in output. However, one can easily compute it by hand using the ANOVA sum of squares and can request it in SPSS.

It is important to recognize, though, that just as an omnibus ANOVA $F$ test does not tell the researcher about where group differences may exist, the $\eta^2$ also fails to characterize where the effect magnitude originates (assuming more than two groups). A good strategy in this case is to report the overall $\eta^2$ for the ANOVA and a standardized mean difference effect such as Cohen’s $d$ alongside specific mean comparison post hoc tests. I demonstrate this strategy in a hypothetical example later in this article.

A Repeat of a Prior Caveat

Similar to standardized mean differences, variance-accounted-for effects are on average statistics. An $R^2$ informs us of the amount of individual differences, on average, that can be explained, but it does not tell us the level of prediction for any given case. Of course, researchers can consult other statistics, such as error scores ($Y - \hat{Y}$), to determine the accuracy for individual cases.

Interpreting Magnitude of Variance-Accounted-For Effects

So, exactly how much variance needs to be accounted for before one can declare a result noteworthy? As before, the answer is “It depends.” Again, research context and the magnitude of effects in prior literature are critical in evaluating result importance.

Nevertheless, because of the GLM, even Cohen’s $d$ (standardized mean difference) and $r^2$ (variance-accounted-for) are related and can be converted into one another using either Cohen’s approximate conversion formula (Cohen, 1988):

$$ r = \frac{d}{\sqrt{d^2 + 4}}, \quad (3) $$

\[ \text{Downloaded from } \text{http://tcp.sagepub.com at Karolinska Institutets Universitetsbibliotek on March 27, 2010} \]
or Aaron, Kromrey, and Ferron’s (1998) more precise formula, which accounts for sample sizes:

\[ r = \frac{d}{\sqrt{d^2 + \frac{N^2 - 2N}{n_1n_2}}} \]  

where \( N \) is the total sample size for both groups and \( n_1 \) and \( n_2 \) are the sample sizes for each group. Of course, squaring the resulting \( r \) would yield the \( r^2 \) variance-accounted-for effect. Equation 4 tends to differ from Equation 3 as sample sizes are smaller and more unequal.

Table 1 presents Equation 3 conversions for the \( d \) benchmarks illustrated above (i.e., .2, .5, and .8), and therefore provides equivalent \( r^2 \)-type effect metrics. Note that the \( r_{pb} \) (point biserial \( r \)) column represents the Equation 3 conversion. However, because the continuously scaled nature of many variables, which are often used as predictors (rather than a dichotomous grouping variable, such as in a \( t \) test) affects the size of the Pearson’s \( r \) correlation (Cohen & Cohen, 1983), Cohen also offered Pearson’s \( r \) conventions (i.e., not direct formulaic conversions) as noted in Table 1.

Although Cohen’s benchmarks for \( d \) are relatively well known, fewer are aware of comparable variance-accounted-for magnitudes. However, knowing that an \( r^2 = .25 \) falls within Cohen’s large effect designation is not particularly informative outside some informed judgment of effect magnitude.

### Multivariate Analogs

To this point, discussion has focused on univariate effect sizes. Similar effects are available for multivariate analyses when simultaneously examining two or more dependent variables.
In MANOVA, for example, researchers can obtain an $\eta^2$ analog by
$(1 – \text{Wilks’s lambda})$ because Wilks’s lambda has the interesting feature of representing the variance not explained in the model. This variance-accounted-for effect would indicate the amount of variance in the combined dependent variables (via a linear equation that creates a synthetic dependent variable, akin to $\hat{Y}$ in regression) that is due to differences between the groups. In canonical correlation analysis, the squared canonical correlation ($R_c^2$) is analogous to regression’s multiple $R^2$ and represents the shared variance between the predictor and the criterion variable sets (Henson, 2002; Sherry & Henson, 2005). In factor analysis, researchers can obtain the total variance reproduced by the factor solution and each factor (Henson, Capraro, & Capraro, 2004; Henson & Roberts, 2006; Kahn, 2006 [this issue]). All of these effects are direct extensions of their univariate variance-accounted-for counterparts.

It follows as well that within the GLM, there should be a multivariate analog to the standardized mean difference. Indeed, Mahalanobis distance ($D^2$) is a squared multivariate version of Cohen’s $d$. Researchers can use $D^2$ to represent the squared standardized statistical distance (difference) between groups on several means simultaneously. Detailed demonstration of the connection between $d$ and $D^2$ is beyond the scope of this article, but Henson (2005) provides an accessible treatment.

Corrected Effect Sizes

Because GLM analyses minimize the error in the analysis, they yield the maximum possible effect size for the obtained data. This would be fine if the sample’s data were perfectly representative of the population to which the researcher hoped to generalize. Of course, this is never the case, and our samples always have some degree of sampling error (nonrepresentativeness) unique to our data. Because some of the variance in our data are unique and because GLM analyses will use this variance to give the largest possible effect, effect sizes such as those discussed above tend to overestimate the effects that would be observed in the population or future samples (Roberts & Henson, 2002; Snyder & Lawson, 1993).

Therefore, researchers can use another set of effect indices, called corrected effects, to provide better estimates of what we might expect in the population or in future samples. This is, of course, the very thing that interests most researchers. Each of these corrections shrinks the uncorrected effect based on the degree of sampling error anticipated in the sample. Because sampling error theoretically increases as (a) the sample size decreases, (b) the number of variables (or groups) in the analysis increases,
and (c) the unknown population effect decreases, more adjusting is necessary for small samples when analyzing many variables or groups and when the obtained effect is small (which we use as our estimate of the unknown population effect).

For univariate group comparisons, $\omega^2$ (omega squared; Hays, 1981) is the most common adjustment of the uncorrected $\eta^2$:

$$\omega^2 = \frac{SS_{between} - (k - 1)MS_{within}}{SS_{total} + MS_{within}},$$  \hspace{1cm} (5)$$

where $SS_{between}$ is the SS between group means, $SS_{total}$ is the SS of the dependent variable, $MS_{within}$ is the mean square within (error), and $k$ is the number of groups (see also Hinkle, Wiersma, & Jurs, 2003). All this information is readily obtainable from one’s ANOVA summary table.

For regression, there are many corrections available, but the Ezekiel (1930) formula is probably most common, mainly because popular statistical software output routinely reports the formula (e.g., labeled “adjusted $R^2$” in SPSS). Vacha-Haase and Thompson (2004) provided an accessible treatment of this formula. In a simulation study of many correction formulae, however, Yin and Fan (2001) demonstrated that other formula options may lead to more accurate corrections. Brevity precludes a review of each of these approaches here, but it is worth noting that some correction is probably better than no correction at all, which is an argument for at least considering the Ezekiel (1930) adjustment. For more information regarding the various options of corrected effects, I refer readers to Leach and Henson (2003), Snyder and Lawson (1993), Vacha-Haase and Thompson (2004), and Yin and Fan (2001).

**Effect-Size Limitations and Cautions**

As noted above, empirical research has documented the many misconceptions and misinterpretations surrounding NHSTs. However, unless effect sizes are appropriately applied and their interpretation and use duly informed, then another set of misinterpretations may arise (Olejnik & Algina, 2000). Effect sizes are not a panacea for all our result-interpretation ills, and therefore I note a few cautions here.

We must first recognize that effect sizes from our samples are simply point estimates of the population effect. Similar to means and other statistics, effects can and will vary from sample to sample. Researchers should therefore exercise caution when interpreting any single effect size. There are ways, however, to help overcome this issue within a study (i.e., using CIs
around effect sizes) and across a given literature (i.e., meta-analytically considering effects across studies). I detail both of these points further below.

It should also be apparent from the above examples that score variability affects effect sizes. For standardized mean difference effects, there is always division by a variability estimate (e.g., pooled SD). It stands to reason that as this variability changes, so would the effect’s magnitude, and so researchers should be aware of this when interpreting results. This dynamic was explicitly demonstrated in the depression treatment example above, where dividing by the control group SD versus the larger T score SD yielded substantially different results. Jacobson and Truax (1991) presented a hypothetical weight-loss example of this possibility, in which a 2-pound average weight loss for a treatment group is only slightly better than an average of 0 pounds in the control. However, if the variability of the two groups were low (i.e., the treatment participants lost about the same amount, and the control participants all tended to lose no weight), then the resulting effect size could be quite large because of division by a smaller variance estimate. In this case, the concept of clinical significance would be critical for result interpretation—that is, was the observed weight loss a meaningful outcome? Furthermore, variance heterogeneity, or unequal population variances, can also negatively affect mean difference effect sizes. Olejnik and Algina (2000) provided a particularly thorough discussion of variability’s role in group-comparison designs.

Although effect sizes can inform practical significance, they are not inherent indices of result meaningfulness (Campbell, 2005). This point has been made throughout the current discussion. An effect’s meaning depends on many factors, possibly including the (a) context of the study, (b) importance of outcomes, and (c) size of effects obtained in prior studies. As such, when interpreting effects, “an element of subjectivity is introduced into the decision process” (Kirk, 1996, p. 755). Although some may be uncomfortable with such interpretive subjectivity, as Henson and Smith (2000) noted, “to assume that science is entirely objective is to suffer from psychological denial” (p. 292). The key is to employ reflective and thoughtful judgment, which involves the defensible interpretation of effect sizes. Although effect sizes cannot cure all our ills, they can provide significant symptomatic relief if employed thoughtfully.

**Summary and Computation of Effect Sizes**

Table 2 summarizes various effect sizes for various analyses. This list is in no way exhaustive but does cover many common analyses. I have presented here some information on computing these effects. For additional
detail, including specific information on obtaining effects with SPSS, see Trusty et al. (2004) and Vacha-Haase and Thompson (2004).

### META-ANALYTIC THINKING AND ADVANCING PSYCHOLOGICAL RESEARCH

Perhaps the single most critical weakness in much of social science research is the lack of emphasis on replication. One-shot studies and, therefore, one-shot $p$ values (particularly in the absence of an effect size) are often of limited value. When we consider the infrequency of random sampling, common threats to study validity, and possible violation of statistical assumptions, the limited generalizability of much of our research is even more apparent. As Henson and Smith (2000) noted,

> By and large, science is about discovering theory that holds true, to some degree at least, in multiple applications. . . . No thoughtful researcher wants to announce his or her “groundbreaking” discovery too loudly until some evidence that the finding was not a fluke emerges. (p. 293)

Nevertheless, outside direct replication, meta-analytic thinking would considerably advance psychological research (Cumming & Finch, 2001,

---

**TABLE 2:** Summary of Some Effect Sizes for Various Analyses

<table>
<thead>
<tr>
<th>Analysis</th>
<th>Effect Sizes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson’s $r$</td>
<td>$r^2$</td>
</tr>
<tr>
<td><em>t</em> test</td>
<td>Cohen’s $d$ or Glass’$s$ delta</td>
</tr>
<tr>
<td>ANOVA (ANCOVA)</td>
<td>$\eta^2$ or $\omega^2$ (corrected effect) for main effects and interactions; Cohen’s $d$ or Glass’s delta for post hoc mean comparisons</td>
</tr>
<tr>
<td>Regression</td>
<td>$R^2$ or corrected $R^2$ (for $R^2$ corrections, see Yin &amp; Fan, 2001)</td>
</tr>
<tr>
<td>Hotelling’s $T^2$</td>
<td>$D^2$ (Mahalanobis distance)</td>
</tr>
<tr>
<td>MANOVA (MANCOVA)</td>
<td>Multivariate $\eta^2$ (1 – Wilks’s lambda); $D^2$ for multivariate mean comparisons</td>
</tr>
<tr>
<td>Descriptive discriminant</td>
<td>Multivariate $\eta^2$ or $R^2$ (squared canonical correlation); $D^2$ for multivariate mean comparisons or Cohen’s $d$ for centroid (mean) comparisons on discriminant function scores (Sherry, 2006 [this issue])</td>
</tr>
<tr>
<td>analysis</td>
<td></td>
</tr>
<tr>
<td>Predictive discriminant</td>
<td>Correctly classified hit rate; Huberty’s $I$ index (Huberty &amp; Lowman, 2000; Sherry, 2006)</td>
</tr>
<tr>
<td>analysis</td>
<td></td>
</tr>
<tr>
<td>Canonical correlation</td>
<td>$R_c^2$</td>
</tr>
<tr>
<td>analysis</td>
<td></td>
</tr>
<tr>
<td>Factor analysis</td>
<td>Variance explained for entire solution and for each factor (Kahn, 2006)</td>
</tr>
</tbody>
</table>
2005; Quintana & Minami, in press [TCP special issue, part 2]). Thompson (2002) defined meta-analytic thinking as

(a) the prospective formulation of study expectations and design by explicitly invoking prior effect sizes and (b) the retrospective interpretation of new results, once they are in hand, via explicit, direct comparison with the prior effect sizes in the related literature. (p. 28)

In other words, we need to explicitly design and place our studies in the context of the effects of prior literature. It is only when we see comparable effects across the literature that we can place increased faith in the results.

The assumption here is that researchers will be less inclined to overemphasize a single study if they are simultaneously considering several. Conversely, researchers may experience increased confidence in their results if they are comparable with prior research. The reporting and referencing of effect sizes from the literature provide a vehicle to make these comparisons more explicit. Within this paradigm, theory development is analytically rather than p-value driven.

**CIs for Effect Sizes**

One relatively new emphasis has been on using CIs for effect sizes (see, e.g., Bird, 2002; Cumming & Finch, 2001; Smithson, 2001; Thompson, 2002). In a recent American Psychologist article, Cumming and Finch (2005) argued for increased use of CIs at least in part because they “support meta-analysis and meta-analytic thinking focused on estimation” (p. 171). The APA TFSI recommended increased use of CIs (Wilkinson & TFSI, 1999), and the APA Publication Manual states that CIs “are, in general, the best reporting strategy” (APA, 2001, p. 22). Thompson (2002) argued that using CIs around effect sizes can facilitate meta-analytic thinking and can advance scientific inquiry. I refer readers to these authors for methods of computing intervals around standardized mean difference and variance-accounted-for effect sizes.

**AN APPLIED EXAMPLE**

An example may best demonstrate the operationalization of effect-size reporting, meta-analytic thinking, and CIs around effect sizes. The following section represents the hypothetical literature and results of a research study that compared college students, adults, and senior adults (n = 30 in each group) on a measure of manifest anxiety. To maximize the heuristic
A Hypothetical Journal Write-Up

This write-up compares manifest anxiety scores for college, adult, and senior adult groups. Figure 5 displays the 95% CI for the means of the three groups. Examining this figure clearly indicates that the adults scored well above the college students and senior adults, on average, as well as the population mean ($T$ score $M = 50$). Furthermore, little difference seemed to exist between the college and the senior adult groups, although the senior adult group mean was a somewhat less precise estimate. Nevertheless, the variances of the groups were reasonably homogeneous ($Levene = 1.489, p = .231$), and we conducted a one-way ANOVA.

The ANOVA results were statistically significant, $F(2, 87) = 9.34, p < .001$. The observed effect was moderate ($\eta^2 = .1768$), indicating that group
membership accounted for about 18% of the anxiety score variance. We observed minimal shrinkage (.0111) due to sampling error based on the $\omega^2$ (.1657) corrected effect (Hays, 1981). Overall, these results suggest noteworthy group differences at the omnibus ANOVA level.

Tukey’s Honestly Significant Difference post hoc tests revealed what we would expect from Figure 5—that is, the adult mean was statistically significantly higher than the college and the senior adult means—$p < .001$ and $p = .002$, respectively—and the college and senior adult groups did not differ statistically ($p = .854$). Because the group variances were reasonably homogeneous, we computed Cohen’s $d$ effects for the post hoc mean comparisons using the pooled standard deviations of the current sample (Cohen, 1988). These effects were $d = 1.18$, .824, and .133, for the adult versus college, adult versus senior adult, and senior adult versus college mean differences, respectively. Consistent with Figure 5, the adult group displayed, on average, substantially more anxiety than the other groups. The small effect for the senior adult and college mean difference seems inconsequential.

These mean differences shed light on the effects observed in prior research. Based on our review of the literature, we found no other studies that simultaneously compared college students, adults, and senior adults on manifest anxiety. However, several researchers have reported comparisons of adults with college students and adults with senior adults. We found no comparisons of college students with seniors. We computed Cohen’s $d$ effect sizes for these prior comparisons based on the information reported by the authors. Figure 6 illustrates the 95% CI for the five $d$ effects obtained across

FIGURE 6. 95% Confidence Intervals for Cohen’s $d$

NOTE: The first four intervals are for adult versus college student comparisons, including the current study. The last three intervals are for adult versus senior adult comparisons, including the current study. CI = confidence interval.
prior studies (3 comparing adults with college students, 2 comparing adults with seniors), as well as the current effects.

Reviewing the intervals for adult versus college comparisons indicates some consistency regarding large differences in anxiety for these groups, in the order of roughly $d = .80$. The literature is less clear regarding the degree of anxiety reported by adults and senior adults. The two prior studies comparing adults with seniors yielded divergent results in terms of statistical significance (i.e., the Study 1 $d$ interval overlaps the $d = 0$ null). However, with the inclusion of the current study, it would seem that we may expect some differences between these groups, on average, in the order of roughly $d = .50$.

CONCLUSION

This article has attempted to address some practical issues surrounding effect-size use and interpretation. It seems apparent that the current statistical reform movement away from $p$-value dependence and toward effect estimation will continue. Of course, this process is indeed gradual given the dominance of NHSTs in our methodology curriculum and literature. Most current counseling psychologists were trained under this paradigm, and therefore, considering effect sizes may require an educational shift for some. As suggested elsewhere, “researchers will not change their ways until journal editors require them to do so, suggesting that old habits die hard and perhaps not without some extrinsic motivation” (Henson & Smith, 2000, p. 285). As noted, at least 23 journals have made such requirements. Of course, more comprehensive change will come when our methodology curriculum integrates modern views of statistical analysis.

For now, authors of The Counseling Psychologist would do well to include and interpret measures of effect in their Results section. They would do even better to meta-analytically think and comment throughout their articles on how to place observed effects in the context of the broader literature (Quintana & Minami, in press).

REFERENCES


