"Statistical," "Practical," and "Clinical": How Many Kinds of Significance Do Counselors Need to Consider?

Bruce Thompson

The present article reviews and distinguishes 3 related but different types of significance: "statistical," "practical," and "clinical." A framework for conceptualizing the many "practical" effect size indices is described. Several effect size indices that counseling researchers can use, or that counselors reading the literature may encounter, are summarized. A way of estimating "corrected" intervention effects is proposed. It is suggested that readers should expect authors to report indices of "practical" or "clinical" significance, or both, within their research reports; and it is noted that indeed some journals now require such reports.

Statistical significance tests have a long history dating back at least to the 1700s. In 1710 a Scottish physician, John Arbuthnot, published his statistical analysis of 82 years of London birth rates as regards gender (Hacking, 1965). Similar applications emerged sporadically over the course of the next two centuries.

But statistical testing did not become ubiquitous until the early 1900s. In 1900, Karl Pearson developed the chi-square goodness-of-fit test. In 1908, William S. Gossett published his *t* test under the pseudonym "Student" because of the employment restrictions of the Dublin-based Guinness brewery in which he worked.

In 1918, Ronald Fisher first articulated the analysis of variance (ANOVA) logic. Snedecor (1934) subsequently proposed an ANOVA test statistic, that he named "*F*" in honor of Fisher, who of course subsequently became "Sir" Ronald Fisher. But it was with the 1925 first publication of Fisher's book *Statistical Methods for Research Workers* and the 1935 publication of his book *The Design of Experiments* that statistical testing was really popularized.

Huberty (1993; Huberty & Pike, 1999) provided authoritative details on this history. However, it is noteworthy that criticisms of statistical testing are virtually as old as the method itself (cf. Berkson, 1938). For example, in his critique of the mindless use of statistical tests titled "Mathematical vs. Scientific Significance," Boring (1919) argued some 80 years ago,

> The case is one of many where statistical ability, divorced from a scientific intimacy with the fundamental observations, leads nowhere. (p. 338)

Statistical tests have been subjected to both intermittent (e.g., Carver, 1978; Meehl, 1978) and contemporary criticisms (cf. Cohen, 1994; Schmidt, 1996). For example, Tryon (1998) recently lamented,

> [T]he fact that statistical experts and investigators publishing in the best journals cannot consistently interpret the results of these analyses is extremely disturbing. Seventy-two years of education have resulted in minuscule, if any, progress toward correcting this situation. It is difficult to estimate the handicap that widespread, incorrect, and intractable use of a primary data analytic method has on a scientific discipline, but the deleterious effects are doubtless substantial. (p. 796)

Anderson, Burnham, and Thompson (2000) provided a chart summarizing the frequencies of publications of such criticisms across both decades and diverse disciplines.

Such criticism has stimulated defenders to articulate views that are also thoughtful. Noteworthy examples include Abelson (1997), Cortina and Dunlap (1997), and Frick (1996). The most balanced and comprehensive treatment of diverse perspectives is provided by Harlow, Mulaik, and Steiger (1997; for reviews of this book, see Levin, 1998; Thompson, 1998).

PURPOSE OF THE PRESENT ARTICLE

The purpose of the present review is not to argue whether statistical significance tests should be banned (cf. Schmidt & Hunter, 1997) or not banned (cf. Abelson, 1997). These various views have been repeatedly presented in the literature.

Bruce Thompson *is a professor and a distinguished research scholar in the Department of Educational Psychology at Texas A&M University, College Station; an adjunct professor of family and community medicine at Baylor College of Medicine, Houston, Texas; and a Visiting Distinguished Fellow at the University Institute for Advanced Study at La Trobe University in Melbourne, Australia. Correspondence regarding this article should be sent to Bruce Thompson, TAMU Department of Educational Psychology, College Station, TX 77843-4225 (Web URL: http://www.coe.tamu.edu/~bthompson).*

Instead, this article has three purposes. First, the article seeks to clarify the distinction between three "kinds" of significance: "statistical," "practical," and "clinical." Second, various indices of practical and clinical significance are briefly reviewed. Finally, it is argued that counselors should not consider only statistical significance when conducting inquiries or evaluating research reports.

Practical or clinical significance, or both, will usually be relevant in most counseling research projects and should be explicitly and directly addressed. Authors should always report one or more of the indices of "practical" or "clinical" significance, or both. Readers should expect them. And it is argued in this article that editors should require them.

THREE KINDS OF SIGNIFICANCE

"Statistical" Significance

What "statistical" significance tests do. Statistical significance estimates the probability ($p_{CALCULATED}$) of sample results deviating as much or more than do the actual sample results from those specified by the null hypothesis for the population, given the sample size (Cohen, 1994). In other words, these tests do *not* evaluate the probability that sample results describe the population; if these statistical tests did that, they would bear on whether the sample results are replicable. Instead, the tests assume that the null exactly describes the population and then test the sample's probability (Thompson, 1996).

Of course, this logic is a bit convoluted and does not tell us what we want to know regarding population values and the likelihood of result replication for future samples drawn from the same population. Thus Cohen (1994) concluded 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).

This logic is sufficiently convoluted that, as empirical studies confirm, many users of statistical tests indeed do not understand what these tests actually do (Mittag & Thompson, 2000; Nelson, Rosenthal, & Rosnow, 1986; Oakes, 1986; Rosenthal & Gaito, 1963; Zuckerman, Hodgins, Zuckerman, & Rosenthal, 1993). Some test users are reduced to merely saying at a superficial level, these tests evaluate whether my results were "due to chance."

What "statistical" tests do not do. Clearly, however, statistical significance does *not* evaluate whether results are important. Some likely events are very important even if they are not unusual or atypical. For example, it is unlikely that an asteroid will destroy our planet in 10 minutes. Although this outcome is expected (i.e., that the planet will not be destroyed in the next 10 minutes), the outcome nevertheless is noteworthy, because our continued existence seems important.

By the same token, very unlikely events (i.e., p is small) may still be very important. In his classic hypothetical dialogue between two graduate students, Shaver (1985) illustrated the folly of equating result improbability with result importance:

> Chris: . . . I set the level of significance at .05, as my advisor suggested. So a difference that large would occur by chance less than five times in a hundred if the groups weren't really different. An unlikely occurrence like that *surely* must be important.
>
> Jean: Wait a minute, Chris. Remember the other day when you went into the office to call home? Just as you completed dialing the number, your little boy picked up the phone to call someone. So you were connected and talking to one another without the phone ever ringing. . . . Well, that must have been a truly important occurrence then? (p. 58)

Furthermore, because the premises of statistical significance tests do not invoke human values, and in valid deductive argument conclusions cannot under any circumstances contain information not present in deductive premises, "If the computer package did not ask you your values prior to its analysis, it could not have considered your value system in calculating p's, and so p's cannot be blithely used to infer the value of research results" (Thompson, 1993, p. 365).

"Practical" Significance

Given considerations such as these, Roger Kirk titled his Southwestern Psychological Association presidential address "Practical Significance: A Concept Whose Time Has Come." Kirk (1996) emphasized that statistical significance tests only evaluate "ordinal relationships" (e.g., whether two group standard deviations are different or one is larger than the other). He argued,

> Is this any way to develop psychological theory? I think not. How far would physics have progressed if their researchers had focused on discovering [only] ordinal relationships [such as those tested by conventional null hypothesis tests]? What we want to know is the size of the difference between *A* and *B* and the error associated with our estimate; knowing *A* is greater than *B* is not enough. (p. 754)

This emphasis on quantifying findings in service of evaluating the practical noteworthiness of results also is not new. For example, long ago Fisher (1925) advocated the calculation in ANOVA of the index called eta squared (or the correlation ratio). Similarly, Kelley (1935) proposed another ANOVA index of practical significance: epsilon squared. These indices have generically come to be called "effect sizes." There are literally dozens of available choices. Various syntheses of these choices are available (cf. Kirk, 1996; Olejnik & Algina, 2000; Snyder & Lawson, 1993).

Effect sizes are particularly important because statistical tests are so heavily influenced by sample sizes. This is one reason why the use of "what if" analyses have been promoted as an adjunct to the use of conventional statistical tests (Snyder & Lawson, 1993; Thompson & Kieffer, 2000).

Thompson (1993) provided a heuristic example that dramaticizes the distinction between statistical and practical significance. The example presumes a researcher was working with test scores from 200,000 students in a large school district.

> If the researcher decided to compare the mean IQ scores ($\bar{X}$ = 100.15, SD = 15) of 12,000 hypothetical students randomly as-

> signed at birth to live in one zip code with the mean IQ ($\bar{X}$ = 99.85, *SD* = 15) of the 188,000 remaining hypothetical students randomly assigned to reside in other zip codes, it would be decided that the two means differ to a statistically significant degree ($Z_{CALC} = 2.12 > Z_{CRIT} = 1.96$, $p \leq .05$). The less thoughtful researcher might suggest to school board members that special schools for gifted students should be erected in the zip code of the 12,000 students, since they are "significantly" brighter than their compatriots. (p. 362)

Obviously, however, a difference of less than a single IQ point is not practically or educationally significant. The result is particularly trivial in relation to the standard error of the measurement of most IQ tests (e.g., SEM = 4.2 or more). That is, we normally would expect scores or means to be *several* SEM's different if we wanted to be certain that differences were not merely an artifact of measurement error.

"Clinical" Significance

In clinical work, practitioners must often make categorical decisions. For example, a psychiatrist must decide whether a patient does or does not require medication for depression. Or a counselor must decide whether or not an acutely depressed patient should or should not be involuntarily hospitalized.

These decisions may be guided by diagnostic criteria or score cutoffs. For example, a physician may invoke a rule that total blood cholesterol greater than 200 milligrams per deciliter requires medication. Or a counselor may render a given diagnosis if four out of six possible symptoms are deemed present.

Effect size indices of "practical" significance may be only partially relevant for applications of research evidence to these sorts of clinical situations. As Jacobson, Roberts, Berns, and McGlinchey (1999) noted, "Group means, for example, do not in and of themselves indicate the proportion of participants who have improved or recovered as a result of treatment" (p. 300). The standard of ultimate clinical significance addresses the question "are treated individuals as a group indistinguishable from normals with respect to the primary complaints following treatment?" (Kendall, 1999, p. 283).

For example, two studies might both involve mean decreases on flagged MMPI-2 scales of 10 T-score points. However, in the first study, all participants might nevertheless still require hospitalization following the intervention, whereas in the second study, with the same effect size, many or even all of the participants might no longer require hospitalization.

This is not to say that the first intervention is not noteworthy. Nevertheless, from a clinical perspective the two studies with identical indices of "practical" effect do still clearly differ as regards what the intervention results are for the patients in a psychiatric hospital.

Kazdin (1999) defined "*clinical" significance* as referring "to the practical or applied value or importance of the effect of the intervention—that is, whether the intervention makes a real (e.g., genuine, palpable, practical, noticeable) difference in everyday life to the clients or to others with whom the client interacts" (p. 332). He distinguished practical and clinical significance still further by noting that even interventions that yield no effects may be clinically significant. For example, an intervention for depression may have no discernable impact in regard to making participants indistinguishable from control group members who were not depressed (i.e., change in symptoms), but it still may do a lot to help people cope with their symptoms or to improve quality of life.

HOW MANY KINDS OF SIGNIFICANCE ARE NEEDED?

Statistical significance is not sufficiently useful to be invoked as the sole criterion for evaluating the noteworthiness of counseling research. Indeed, even statistically nonsignificant studies may yield effects that are still both noteworthy and replicable.

For example, conceivably 200 studies of a new antitumor drug could each yield practically noteworthy effects for which $p_{CALCULATED}$ values were all .06 (Thompson, 1999b). The effects are noteworthy in that they involve human longevity and, in this scenario, are demonstrably replicable. In this sequence of events, "surely, God loves the .06 nearly as much as the .05" (Rosnow & Rosenthal, 1989, p. 1277) level of statistical significance.

At a minimum, counselors should expect a research literature that says more than that therapy *A* is better than therapy *B*. Within the context of a single study, the question must be "by how much is therapy *A* better?" Too much of "What we see [today] is a reject-nonreject decision strategy that does not tell us what we want to know and a preoccupation with *p* values that are several steps removed from examining the data" (Kirk, 1996, pp. 754–755).

Furthermore, counselors should expect a literature in which the results of a single study are explicitly interpreted using effect sizes in direct comparisons with the typical effect sizes in previous studies and the ranges of those effect sizes. This focuses attention on evaluating how consistent the intervention is across settings or situations. Practice will improve once researchers formally consider the replicability of results when they evaluate results.

Why Effect Size Interpretation Should Be Required

As readers know, the 1994 American Psychological Association (APA) *Publication Manual* incorporated an important revision "encouraging" (p. 18) authors to report effect sizes. However, there are now 11 empirical studies of either 1 or 2 volumes of 23 different journals demonstrating that this encouragement has been ineffective (Vacha-Haase, Nilsson, Reetz, Lance, & Thompson, 2000).

The reasons why the "encouragement" has been so ineffective include the fact that only "encouraging" effect size reporting

> presents a self-canceling mixed- message. To present an "encouragement" in the context of strict absolute standards regarding the esoterics of author note placement, pagination, and margins is to send the message, "these myriad requirements count, this encouragement doesn't." (Thompson, 1999b, p. 162)

Consequently, various journals have now adopted editorial policies "requiring" that effect sizes be reported. These include

- *The Career Development Quarterly*
- *Contemporary Educational Psychology*
- *Educational and Psychological Measurement*
- *Exceptional Children*
- *Journal of Agricultural Education*
- *Journal of Applied Psychology*
- *Journal of Community Psychology*
- *Journal of Consulting and Clinical Psychology*
- *Journal of Counseling & Development*
- *Journal of Early Intervention*
- *Journal of Educational and Psychological Consultation*
- *Journal of Experimental Education*
- *Journal of Learning Disabilities*
- *Language Learning*
- *Measurement and Evaluation in Counseling and Development*
- *The Professional Educator*
- *Research in the Schools*

Such policies are consistent with the recent recommendations of the APA Task Force on Statistical Inference, which was appointed by the APA Board of Scientific Affairs in 1996.

In its August 1999 article in the *American Psychologist*, the Task Force noted, "*Always* [italics added] provide some effect-size estimate when reporting a *p* value" (Wilkinson & APA Task Force on Statistical Inference, 1999, p. 599). Later the Task Force also wrote, "We must stress again that reporting and interpreting effect sizes in the context of previously reported effects is *essential* [italics added] to good research" (p. 599).

More recently, the fifth edition of the APA (2001) *Publication Manual* was published. The new manual emphasized

> that 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)

The view that editorial policies should require effect size reporting places the burden for change at the doorstep from which contemporary practices originated. Years ago Glantz (1980) noted that "The journals are the major force for quality control in scientific work" (p. 3). And as Sedlmeier and Gigerenzer (1989) argued, "there is only one force that can effect a change, and that is the same force that helped institutionalize null hypothesis testing as the *sine qua non* for publication, namely, the editors of the major journals" (p. 315).

Brief Review of Effect Size Choices

Ideally, counselors would have access to a literature describing both the "practical" and the "clinical" significance of their intervention choices. However, evaluating clinical significance is both methodologically and philosophically more complicated than evaluating practical significance (cf. Kendall, 1999). Thus, the field would move forward if at least effect size reporting finally became routine. And, although large practical effects do not assure clinically significant effects, nevertheless, "large effects are more likely to be clinically significant than small ones" (Jacobson et al., 1999, p. 300).

Although the 1994 APA *Publication Manual* "encouraged" effect size reporting, "unfortunately, . . . the effect size of this encouragement has been negligible" (Wilkinson & APA Task Force on Statistical Inference, 1999, p. 599) for various reasons including the self-canceling nature of the "encouragement." However, progress has been slow also because, until recently, effect sizes computations were not widely available within statistical packages.

A host of effect sizes are available to counseling researchers. There is no one definitely correct choice, and, in any case, the reporting and interpretation of any effect size indices would represent improvement over common contemporary practice.

However, some counseling researchers and research consumers may have had only limited exposure to effect size concepts during their graduate training. Thus, a brief review of a framework for conceptualizing effect sizes may be useful. The framework and some illustrative choices are presented in Figure 1.

Standardized differences versus variance-accounted-for indices. Although Kirk (1996) acknowledged the existence of a third, "miscellaneous" class of effects, Kirk and others generally recognized two major classes of effect sizes: standardized differences and variance-accounted-for indices. There are numerous effect size choices available within each category.

A simple difference in group means, for example, can be computed to evaluate counseling intervention effects. Thus, the mean in one treatment group might equal 101, whereas the mean in the control group might equal 100. Here the mean difference equals 1.0.

However, this mean difference is not a suitable index of intervention effect. This is because the meaning of a 1.0 difference depends entirely on the scale of the measurement. If, on the one hand, the experiment dealt with an intervention to raise IQ, and the standard deviation equaled 15.0, this one unit difference in the means would be fairly small. On the other hand, if the measurement involved the

	Standardized Differences	Relationship Variance-Accounted-For
Uncorrected	Glass's g' Cohen's d	eta^2 (η^2; also called correlation ratio [not the correlation coefficient!])
"Corrected"	Thompson's "Corrected" d^*	Hays's omega^2 (ω^2) Adjusted R^2

FIGURE 1

A Framework for Conceptualizing the Most Common Effect Sizes

Note. The standardized differences indices are in an unsquared, standardized score metric. The relationship variance-accounted-for indices are in a squared metric (e.g., r^2).

temperature of children in degrees Fahrenheit, for which standard deviation is roughly .2, this one unit mean difference would be quite large. This dynamic is the reason why researchers presenting means should *always* provide the standard deviations of the scores about every mean.

The scaling problem can be addressed in interpreting mean differences by "standardizing" the difference. This is accomplished by dividing the mean difference by some estimate of standard deviation. Several choices are reasonable, only two of which are summarized here.

In 1976, as part of his articulation of meta-analytic methods, Glass proposed g' (or Δ), which divides the mean difference by the standard deviation of the control group. Glass reasoned that the standard deviation in this group was the best estimate of the population standard deviation, in that the intervention, which may affect the standard deviation in addition to affecting the mean, would not have done so in the control group. This reasoning is most tenable when the control group has received no treatment or an irrelevant placebo treatment.

Cohen's (1969) *d*, on the other hand, invokes a standard deviation estimate that is "pooled" or averaged across both the intervention and the control groups. Cohen reasoned that both groups provide information about score scaling and that a more stable estimate would be achieved by using a larger sample size derived from both groups.

In articulating *d* (and other indices), Cohen provided general suggestions for interpreting these indices regarding their typicality in the literature throughout the behavioral sciences. He suggested that a standardized difference of about |.5| is "medium," whereas values of |.2| and |.8| are "small" and "large," respectively. As Kirk (1996) noted, these guidelines helped make the indices more appealing to researchers. And various meta-analyses have suggested that Cohen's intuitions regarding effect typicality were fairly accurate.

However, Cohen did not want researchers to invoke these guidelines blindly. Indeed, as Zwick (1997) suggested, if we used these guidelines with the same rigidity that the $\alpha = .05$ criterion has been used, we would merely be being stupid in a new metric.

Of course, counseling researchers are not only interested in means. For example, an intervention may not affect the means of treatment recipients (e.g., the average depression score may remain unaltered after intervention) but may make the scores more variable (i.e., the *SD* of the depression scores might be increased by the intervention, as might happen if the intervention made highly depressed participants even more depressed but made less depressed participants still less depressed). In addition, not all research is experimental. Thus, variance-accounted-for relationship effect sizes may also be relevant.

Because all statistical analyses (e.g., ANOVA, *t* test, *R*, R_C) are correlational, even though designs may be experimental or nonexperimental, a variance-accounted-for relationship effect size analogous to r^2 can be computed in all parametric analyses (see Cohen, 1968; Knapp, 1978; and Bagozzi, Fornell, & Larcker, 1981, respectively, for their seminal explications of the univariate, the multivariate, and the structural equation modeling General Linear Model [GLM]). For example, in multiple regression, the R^2 can be computed by dividing the sum of squares explained by the sum of squares total. This effect tells the researcher what percentage of the variability in individual differences of the participants on the outcome variable can be explained or predicted with knowledge of the scores on the predictor variables.

In ANOVA, the analogous effect size eta squared (η^2; Fisher, 1925) can be computed by dividing the sum of squares between (also called "model" or "regression") by the sum of squares total. This effect tells the researcher what percentage of the variability in individual differences of the participants on the outcome variable can be explained or predicted with knowledge of the group or cell membership of the participants.

Uncorrected versus "corrected" effect sizes. One problem that researchers inherently confront is probably underrecognized in contemporary research: Every sample, like every person, has its own unique and irreproducible character. These individual differences in persons make people interesting; these sampling differences in samples, however, are less appealing because they make results difficult to replicate. The problem is that every sample from the population will contain some "flukiness" even if the sample is randomly drawn.

The difficulty is that all GLM analyses (e.g., ANOVA, regression, descriptive discriminant analysis) capitalize on all the variances in our data, including variance that is unique to a particular given sample. This capitalization results in an inflated variance-accounted-for effect size that is positively biased (i.e., overestimates the true population effect or the effect in future samples).

Happily, we know what design features cause more or less sampling error variance. Consequently, we can "correct" our effect sizes for these influences. When we invoke these corrections, the "shrunken" estimates will always be equal to or less than our original uncorrected (and positively biased) estimates.

Three design features create more sampling error variance and thus positive bias in effect estimation. First, as would be expected, studies with smaller sample sizes involve more sampling error. Second, studies involving more measured variables have more sampling error; this is because there are more opportunities to create sample "flukiness" as we measure more variables. Third, there is more sampling error variance in studies conducted when the population effect size is smaller.

This third influence is more difficult to understand. As an extreme heuristic example, pretend that one was conducting a bivariate r^2 study in a situation in which the population r^2 value was 1.0. In this population scattergram, every person's asterisk is exactly on a single regression line. In this instance, even if the researcher draws ridiculously small samples, such as $n = 2$ or $n = 3$, and no matter which participants are drawn, the researcher simply cannot incorrectly estimate the variance-accounted-for effect size. That is, *any* two or three or four people will always define a straight line in the sample scattergram, and thus r^2 will always be 1.0.

Because we do not actually know the true population variance-accounted-for effect size, we typically use the

actual sample value (e.g., η^2, R^2) as the estimated population effect in our corrections. Examples of "corrected" variance-accounted-for effect size indices, each of which invoke these three (and only these three) design features, are the regression "adjusted R^2" (Ezekiel, 1930) or the ANOVA omega squared (ω^2; Hays, 1981).

Conversion of effects into each other's metrics. As noted previously, standardized differences are in an unsquared standardized metric, whereas variance-accounted-for relationship effect sizes are in a squared metric. These metric differences can be surmounted to convert these effects into each others' metrics.

For example, if the previous school district example somehow involved an experiment, the Cohen's *d* would be

$$\begin{aligned} d &= (\bar{X}_E - \bar{X}_C) / SD_{pooled} \\ &= (100.15 - 99.85) / 15 \\ &= .3 / 15 \\ &= .02 . \end{aligned}$$

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

$$\begin{aligned} r &= d / [(d^2 + 4)^{.5}] \\ &= .02 / [(.02^2 + 4)^{.5}] \\ &= .02 / [(.0004 + 4)^{.5}] \\ &= .02 / [(4.0004)^{.5}] \\ &= .02 / 2.000099 \\ &= .00999 . \end{aligned}$$

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

$$\begin{aligned} d &= [2 (r)] / [(1 - r^2)^{.5}] \\ &= [2 (.00999)] / [(1 - .00999^2)^{.5}] \\ &= .019999 / [(1 - .00999^2)^{.5}] \\ &= .019999 / [(1 - .000099)^{.5}] \\ &= .019999 / [(.999900)^{.5}] \\ &= .019999 / .999950 \\ &= .02 . \end{aligned}$$

Proposed "corrected" standardized difference. The facts (a) that variance-accounted-for effect sizes can be computed in all parametric analyses, given the general linear model, and (b) that effects can be converted into squared or unsquared metrics, suggest the possibility proposed here of computing a "corrected" standardized-difference effect size. The mandate for such an effect index is straightforward: Under the GLM, the same three study features that create sampling error variance and bias the "uncorrected" variance-accounted-for effect size also introduces biases in the standardized differences.

A "corrected" standardized difference (d^*) can be computed by (a) converting a standardized difference (e.g., *d*) into an *r*, (b) converting the *r* into an r^2 by squaring *r*, (c) invoking a sampling error variance correction formula (e.g., Ezekiel, 1930) to estimate the "corrected" effect r^{2*}, (d) converting this corrected r^{2*} back into r^*, and then (e) converting the "corrected" r^* back into d^*. For example, let's assume that an intervention study involving two groups both with sample sizes of 30 yielded a Cohen's *d* of +.5. Using the formula presented previously, the equivalent *r* would be

$$\begin{aligned} r &= d / [(d^2 + 4)^{.5}] \\ &= .5 / [(.5^2 + 4)^{.5}] \\ &= .5 / [(.25 + 4)^{.5}] \\ &= .5 / [(4.25)^{.5}] \\ &= .5 / 2.061552 \\ &= .242535 . \end{aligned}$$

The squared *r* would then be $.242535^2 = .058823$.

One of the several equivalent formulas for the Ezekiel (1930) correction can be expressed as

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

In our study we effectively have one predictor variable (i.e., group membership), so the correction would be

$$\begin{aligned} r^{2*} &= .058823 - ((1 - .058823) \times (1 / (60 - 1 - 1))) \\ &= .058823 - ((.941176) \times (1 / (60 - 1 - 1))) \\ &= .058823 - (.941176 \times (1 / 58)) \\ &= .058823 - (.941176 \times .017241) \\ &= .058823 - .016227 \\ &= .042596 . \end{aligned}$$

Converting the squared value back to an unsquared value yields $.042596^{.5} = .206388$.

Finally, applying the conversion of *r* to *d* yields the "corrected" standardized difference, d^*:

$$\begin{aligned} d^* &= [2 (r)] / [(1 - r^2)^{.5}] \\ &= [2 (.206388)] / [(1 - .206388^2)^{.5}] \\ &= [.412777] / [(1 - .206388^2)^{.5}] \\ &= .412777 / [(1 - .042596)^{.5}] \\ &= .412777 / [(.957403)^{.5}] \\ &= .412777 / .978470 \\ &= .421860 . \end{aligned}$$

This standardized difference has "shrunken" (from +.50 to +.42) once the sampling error influence has been removed for the original effect size estimate. The shrunken value is more conservative, but most importantly is more likely to be accurate and replicable.

Additional Choices

In addition to arguing that effect size reporting is essential to good research practice, the Task Force (Wilkinson & APA Task Force on Statistical Inference, 1999) also strongly encouraged the reporting of confidence intervals. Indeed, the new APA (2001) *Publication Manual* noted that confidence intervals "are, in general, *the best* reporting strategy. The use of confidence intervals is therefore *strongly recommended* [italics added]" (p. 22).

A logical combination of these recommendations is to report confidence intervals about effect sizes themselves. This is an appealing strategy, but estimating these intervals can be very complicated. Fortunately, Cumming and Finch (2001) and Smithson (2001) explain how to estimate confidence intervals about effect sizes and provide software with which to do so.

DISCUSSION

When we evaluate counseling studies we cannot use $p_{CALCULATED}$ values as a satisfactory index of study effect sizes.

This is because sample size directly affects *p* values, and thus "virtually any study can be made to show significant results if one uses enough subjects" (Hays, 1981, p. 293). The problem is that when different studies involve different sample sizes, *p* values will differ in each study, even if every study had exactly the same effect size. As noted elsewhere,

> The calculated *p* values in a given study are a function of several study features, but are particularly influenced by the confounded, joint influence of study sample size and study effect sizes. Because *p* values are confounded indices, in theory 100 studies with varying sample sizes and 100 different effect sizes could each have the same single $p_{CALCULATED}$, and 100 studies with the same single effect size could each have 100 different values for $p_{CALCULATED}$. (Thompson, 1999a, pp. 169–170)

What we seek in evaluating clinical interventions are indices characterizing (a) the typical effect and (b) the range of clinical effects across studies. Calculated *p* values are not sufficient for this purpose.

Effect sizes, on the other hand, are useful quantifications of intervention impacts in a single study. And effect sizes are particularly valuable when we (a) formulate anticipated study effects prior to the intervention by consulting effects from previous related studies and (b) interpret actual study effects once the study has been conducted in the context of prior effects.

As the APA Task Force on Statistical Inference noted,

> It helps to add brief comments that place these effect sizes in a practical and theoretical context. . . . We must stress again that reporting and interpreting effect sizes in the context of previously reported effects is *essential* [italics added] to good research. (Wilkinson & APA Task Force on Statistical Inference, 1999, p. 599)

The context of previous effects in related studies helps inform judgment regarding the likely replicability of results. Authors should provide this contextual information as part of result interpretation. Editors and readers should expect this information. In some contexts information about "clinical" significance should also be expected. However, as noted previously, "practical" and "clinical" significance are related. Interventions with large effect sizes are disproportionately likely to be "clinically" significant as well.

In summary, disciplines move slowly. However, they do move inexorably. In the past decade, in particular, more people have recognized that "statistical" significance may not be sufficient to the task of serving as the sole criterion for evaluating result import. Thus, effect size reporting has moved from not being mentioned, to being "encouraged" (APA, 1994, p. 18), and finally to being characterized as "essential" (Wilkinson & APA Task Force on Statistical Inference, 1999, p. 599). At the same time, more journals have adopted editorial policies requiring effect size reporting and interpretation.

Some may object that evaluating result "practical" significance makes the process of inquiry feel less scientific. As Kirk (1996) acknowledged,

> it is true that an element of subjectivity is introduced into the decision process when researchers make this kind of judgment. And the judgment inevitably involves a variety of considerations, including the researcher's [personal] value system. . . . However, I believe that researchers have an obligation to make this kind of judgment. (p. 755)

The clients that our profession serves probably do not want to know how unlikely intervention effects are in relation to what Cohen (1994) called the "nil" null hypothesis. Instead, what they may want to know is how much difference treatments will make in their lives. So perhaps this should be the metric that guides both our conduct and our consumption of counseling scholarship.

REFERENCES

Abelson, R. P. (1997). A retrospective on the significance test ban of 1999 (If there were no significance tests, they would be invented). In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 117–141). Mahwah, NJ: Erlbaum.

American Psychological Association. (1994). *Publication manual of the American Psychological Association* (4th ed.). Washington, DC: Author.

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

Anderson, D. R., Burnham, K. P., & Thompson, W. (2000). Null hypothesis testing: Problems, prevelance, and an alternative. *Journal of Wildlife Management, 64*, 912–923.

Bagozzi, R. P., Fornell, C., & Larcker, D. F. (1981). Canonical correlation analysis as a special case of a structural relations model. *Multivariate Behavioral Research, 16*, 437–454.

Berkson, J. (1938). Some difficulties of interpretation encountered in the application of the chi-square test. *Journal of the American Statistical Association, 33*, 526–536.

Boring, E. G. (1919). Mathematical vs. scientific importance. *Psychological Bulletin, 16*, 335–338.

Carver, R. (1978). The case against statistical significance testing. *Harvard Educational Review, 48*, 378–399.

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.). Hillsdale, NJ: Erlbaum.

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

Cortina, J. M., & Dunlap, W. P. (1997). Logic and purpose of significance testing. *Psychological Methods, 2*, 161–172.

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*, 532–575.

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

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

Fisher, R. A. (1935). *The design of experiments*. Edinburgh, Scotland: Oliver & Boyd.

Frick, R. W. (1996). The appropriate use of null hypothesis testing. *Psychological Methods, 1*, 379–390.

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

Glantz, S. A. (1980). Biostatistics: How to detect, correct and prevent errors in the medical literature. *Circulation, 61*, 1–7.

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

Hacking, I. (1965). *Logic of statistical inference*. New York: Cambridge.

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (Eds.). (1997). *What if there were no significance tests?* Mahwah, NJ: Erlbaum.

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

Huberty, C. J (1993). Historical origins of statistical testing practices: The treatment of Fisher versus Neyman-Pearson views in textbooks. *Journal of Experimental Education, 61*, 317–333.

Huberty, C. J, & Pike, C. J. (1999). On some history regarding statistical testing. In B. Thompson (Ed.), *Advances in social science methodology* (Vol. 5, pp. 1–23). Stamford, CT: JAI Press.

Jacobson, N. S., Roberts, L. J., Berns, S. B., & McGlinchey, J. B. (1999). Methods for defining and determining the clinical significance of treatment effects: Description, application, and alternatives. *Journal of Consulting and Clinical Psychology, 67*, 300–307.

Kazdin, A. E. (1999). The meanings and measurement of clinical significance. *Journal of Consulting and Clinical Psychology, 67*, 332–339.

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

Kendall, P. C. (1999). Clinical significance. *Journal of Consulting and Clinical Psychology, 67*, 283–284.

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

Knapp, T. R. (1978). Canonical correlation analysis: A general parametric significance testing system. *Psychological Bulletin, 85*, 410–416.

Levin, J. R. (1998). To test or not to test H_0? *Educational and Psychological Measurement, 58*, 311–331.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46*, 806–834.

Mittag, K. C., & Thompson, B. (2000). A national survey of AERA members' perceptions of statistical significance tests and other statistical issues. *Educational Researcher, 29*(4), 21–27.

Nelson, N., Rosenthal, R., & Rosnow, R. L. (1986). Interpretation of significance levels and effect sizes by psychological researchers. *American Psychologist, 41*, 1299–1301.

Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioral sciences*. New York: Wiley.

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

Rosenthal, R., & Gaito, J. (1963). The interpretation of level of significance by psychological researchers. *Journal of Psychology, 55*, 33–38.

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.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin, 105*, 309–316.

Shaver, J. (1985). Chance and nonsense. *Phi Delta Kappan, 67*(1), 57–60.

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.

Snedecor, G. W. (1934). *Calculation and interpretation of analysis of variance and covariance*. Ames, IA: Collegiate Press.

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

Thompson, B. (1993). The use of statistical significance tests in research: Bootstrap and other alternatives. *Journal of Experimental Education, 61*, 361–377.

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

Thompson, B. (1998). Review of *What if there were no significance tests? Educational and Psychological Measurement, 58*, 332–344.

Thompson, B. (1999a). If statistical significance tests are broken/misused, what practices should supplement or replace them? *Theory & Psychology, 9*, 167–183.

Thompson, B. (1999b). Journal editorial policies regarding statistical significance tests: Heat is to fire as *p* is to importance. *Educational Psychology Review, 11*, 157–169.

Thompson, B., & Kieffer, K. M. (2000). Interpreting statistical significance test results: A proposed new "What if" method. *Research in the Schools, 7*(2), 3–10.

Tryon, W. W. (1998). The inscrutable null hypothesis. *American Psychologist, 53*, 796.

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., & APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*, 594–604. (Reprint available through the APA Home Page: http://www.apa.org/journals/amp/amp548594.html)

Zuckerman, M., Hodgins, H. S., Zuckerman, A., & Rosenthal, R. (1993). Contemporary issues in the analysis of data: A survey of 551 psychologists. *Psychological Science, 4*, 49–53.

Zwick, R. (1997, March). *Would the abolition of significance testing lead to better science?* Paper presented at the annual meeting of the American Educational Research Association, Chicago.