Journal of Modern Applied Statistical Methods

Volume 2 | Issue 1 Article 12

5-1-2003

Without Supporting Statistical Evidence, Where Would Reported Measures of Substantive Importance Lead? To No Good Effect

Anthony J. Onwuegbuzie
University of South Florida, tonyonwuegbuzie@aol.com

Joel R. Levin
University of Arizona, jrlevin@u.arizona.edu

Recommended Citation
Onwuegbuzie, Anthony J. and Levin, Joel R. (2003) "Without Supporting Statistical Evidence, Where Would Reported Measures of Substantive Importance Lead? To No Good Effect," *Journal of Modern Applied Statistical Methods*: Vol. 2 : Iss. 1 , Article 12.
DOI: 10.22237/jmasm/1051747920

Journal of Modern Applied Statistical Methods
May, 2003, Vol. 2, No. 1, 133-151

1538 – 9472/03/$30.00

Without Supporting Statistical Evidence, Where Would Reported Measures of Substantive Importance Lead? To No Good Effect

Anthony J. Onwuegbuzie
University of South Florida

Joel R. Levin
University of Arizona

Although estimating substantive importance (in the form of reporting effect sizes) has recently received widespread endorsement, its use has not been subjected to the same degree of scrutiny as has statistical hypothesis testing. As such, many researchers do not seem to be aware that certain of the same criticisms launched against the latter can also be aimed at the former. Our purpose here is to highlight major concerns about effect sizes and their estimation. In so doing, we argue that effect size measures *per se* are not the hoped-for panaceas for interpreting empirical research findings. Further, we contend that if effect sizes were the only basis for interpreting statistical data, social-science research would not be in any better position than it would if statistical hypothesis testing were the only basis. We recommend that hypothesis testing and effect-size estimation be used in tandem to establish a reported outcome's believability and magnitude, respectively, with hypothesis testing (or some other inferential statistical procedure) retained as a "gatekeeper" for determining whether or not effect sizes should be interpreted. Other methods for addressing statistical and substantive significance are advocated, particularly confidence intervals and independent replications.

Key words: Effect-size concerns, statistical inference, substantive importance

Introduction

Statistical hypothesis testing has been implemented to assess the believability, or non-"chanceness" (Levin, 1998b; Levin & Robinson, 1999), of research findings for more than 75 years, stemming from the seminal works of Fisher (1925/1941) and Neyman and Pearson (1928). Despite the widespread use of hypothesis testing during most of the last century through today, its practice has been controversial. Indeed, over the past few decades testing for statistical significance has come under close scrutiny.

Anthony J. Onwuegbuzie is Associate Professor, Department of Educational Measurement and Research, College of Education, University of South Florida, 4202 East Fowler Avenue, EDU 162, Tampa, Florida 33620-7750. Email him at tonyonwuegbuzie@aol.com). Joel R. Levin is Professor, Department of Educational Psychology, College of Education, University of Arizona. E-Mail: jrlevin@u.arizona.edu

Since 1950, for example, the number of articles published in the fields of education, psychology, ecology, and medicine criticizing hypothesis testing has been increasing at an exponential rate (Anderson, Burnham, & Thompson, 2000). Additionally:

(*a*) professional journals (e.g., *The Journal of Experimental Education* and *Research in the Schools*) have devoted special theme issues to statistical hypothesis testing; and

(*b*) symposia have been held at national annual meetings, such as the American Educational Research Association, the American Psychological Association, and the American Psychological Society. Even an edited book, *What if there were no significance tests*? (Harlow, Mulaik, & Steiger, 1997), has been devoted exclusively to the topic.

The Case Against Statistical Hypothesis Testing

Some of the staunchest critics of statistical hypothesis testing contend that this practice has been extremely harmful to scientific progress in the social sciences. For example, Meehl (1978, p.

817) stated that it "is a terrible mistake, a basically unsound, poor scientific strategy, and one of the worst things that ever happened in the history of psychology." Rozeboom (1997) continued:

> Null-hypothesis significance testing is surely the most bone-headedly misguided procedure ever institutionalized in the rote training of science students...[I]t is a sociology-of-science wonderment that this statistical practice has remained so unresponsive to criticism. (p. 335)

Similarly, Tryon (1998) complained:

> [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 undoubtedly substantial. (p. 796)

Schmidt and Hunter (1997, p. 37) claimed that "[s]tatistical significance testing retards the growth of scientific knowledge; it never makes a positive contribution," and Thompson (1992b, p. 436) added: "[Statistical significance testing] has created considerable damage as regards the cumulation of knowledge."

As a result of the purported flaws that statistical hypothesis testing has been accused of, several researchers have recommended that it be banned completely (e.g., Bakan, 1966; Cahan, 2000; Carver, 1978, 1993; Cohen, 1994; Guttman, 1985; Loftus, 1996; Meehl, 1967, 1978; Nix & Barnette, 1998; Rozeboom, 1960; Schmidt, 1992; 1996; Schmidt & Hunter, 1997). Although we: (*a*) agree that statistical hypothesis testing has been misused, and (*b*) concur with many of the criticisms of it that have been offered, it is quite a leap to charge that hypothesis testing by itself has stunted "the cumulation of knowledge" (Thompson, 1992b, p. 436), is "one of the worst things that ever happened in the history of psychology" (Meehl, 1978, p. 817), or "retards the growth of scientific knowledge... [and]... never makes a positive contribution" (Schmidt & Hunter, 1997, p. 37).

Furthermore, some of the assertions made in an attempt to invalidate the hypothesis-testing practice either have been accompanied by unsubstantiated claims or represent flawed logic. As noted by Krantz (1999):

> It is one thing to accuse scientists of showing their ignorance of statistical reasoning in the course of their science, but this does not imply that their ultimate conclusions will be incorrect, nor even that their efficiency in reaching correct conclusions will be impaired. A causal attribution of this sort needs to be supported by careful empirical arguments. (p. 1378)

The foregoing concerns aside, valid criticisms of statistical hypothesis testing have nonetheless been made. Fan (2001) provided a summary of some of these criticisms:

> Thompson (1993) discussed three relevant criticisms for (*sic*.) statistical significance testing: (*a*) overdependency on sample size, (*b*) some nonsensical comparisons, and (*c*) some inescapable dilemmas created by statistical significance testing (e.g., testing for assumption vs. testing for the research hypothesis). In a similar vein, Kirk (1996) discussed three major criticisms of statistical significance testing: (*a*) Significance testing does not tell researchers what they want to know, but rather, it creates the illusion of probabilistic proof by contradiction (Falk & Greenbaum,

> 1995). (*b*) Statistical significance testing is often a trivial exercise because it simply indicates the power of the design (which primarily depends on the sample size) to reject the false null hypothesis. (*c*) Significance testing "turns a continuum of uncertainty into a dichotomous reject-do-not-reject decision," and this dichotomous decision process may "lead to the anomalous situation in which two researchers obtain identical treatment effects but draw different conclusions" (Kirk, p. 748) because of the slight differences in their design (e.g., sample sizes). (p. 276)

Because of these and other concerns, many researchers have called for the reporting of measures of practical significance (or substantive importance, as reflected by effect size or strength of relationship indices), either in addition to or instead of testing for statistical significance. Indeed, the most recent edition of the influential *Publication Manual of the American Psychological Association* (2001) states:

> 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. (p. 26)

Certain anti-hypothesis-testers (e.g., Carver, 1993) even go so far as to endorse effect-size estimates as *replacements* for statistical significance testing – that is, they contend that effect sizes are all that are needed to make inferences about empirical research outcomes. As is argued throughout the remainder of this manuscript, however, we believe that such practice would only lead to no good effect!

Debates about the value and warrants of statistical hypothesis testing can be traced back to Boring (1919) and Berkson (1938, 1942). Over the last decade, many researchers have seemingly jumped on the effect-size bandwagon without scrutinizing its use to the same degree as has occurred for hypothesis testing. Moreover, what appears to have been lost in all this fervor for effect-size provision – and as we illustrate later – is that many of the same criticisms launched against statistical hypothesis testing can also be aimed at effect sizes. As one salient illustration, cautions concerning hypothesis testing and its interpretation can be found in such sources as the aforementioned APA *Publication Manual* (2001) – namely, that *p*-values (statistical significance probabilities) do not directly reflect "the magnitude of an effect or the strength of a relationship" (p. 25). Yet, no such cautions about effect-size measures are found in that pivotal reference source.

Concerns and Cautions About Effect Sizes

In what follows we highlight several major concerns about effect sizes and their estimation, in what might be called *nine effect-size nuisances and no-no's*. In doing so, we consider several rarely acknowledged limitations of effect-size measures. We (as others before us) argue that effect-size measures are influenced by, and therefore must be interpreted with respect to, a number of critical factors. As a preliminary comment, we regard certain of these considerations as being especially relevant when effect sizes are reported as *sole* indicators of an empirical study's significance (i. e., as reflected in Carver's, 1993, "effect-size only" recommendation). We return to this fundamental issue in a later section.

According to Wilkinson and the Task Force on Statistical Inference (1999, p. 599) "[R]eporting and interpreting effect sizes...is essential to good research." Unfortunately, this statement might suggest to some that the provision of effect sizes necessarily improves the quality of empirical studies. Yet, the uncritical acceptance of effect size measures is problematic because, as is now discussed, such measures are sensitive to a number of factors, such as: the research objective; sampling design (including the levels of the independent variable, choice of treatment alternatives, and statistical analysis employed); sample size and variability; type and range of the measures used; and score reliability (see, for example, Fern & Monroe, 1996; Frick, 1995;

O'Grady, 1982; Olejnik & Algina, 2000; and Sechrest & Yeaton, 1982).

1. The research objective. According to Fern and Monroe (1996), one's interpretation of an effect size should vary, depending on whether the objective of the study is what they call theory application or effects application. In theory-application research (or explanatory studies) the goal is to identify theories that increase our understanding of phenomena. Studies involving theory application, which consist primarily of theory generation and theory testing, typically focus on generalizing theories beyond the underlying sample and/or context. More specifically, in explanatory studies, the goal is to determine the "shape or functional nature of a relationship" (O'Grady, 1982, p. 770).

In such investigations, a large effect size is not necessarily of interest. Indeed, a large effect may be viewed as a negative outcome if it was not predicted by theory. That is, in theory-application research, a small effect may be more informative and useful than a large effect (Calder, Phillips, & Tybout, 1981). In fact, using "large" effect-size guidelines (e.g., Cohen, 1988) as the criterion for choosing among several independent variables in explanatory studies may culminate in misleading final theoretical models being selected. Conversely, in effects-application research (or predictive studies), researchers usually are not interested in generalizing the results beyond the levels of the variables selected. That is, in effects-application studies, the interest is more on the size of the effect than on determining the generalizability of a particular theory. This suggests that effect sizes should not be interpreted without taking into account whether one's research objective is essentially explanatory or predictive in nature.

2. Choice of a specific research design and experimental conditions. The selected research design also affects interpretation of effect sizes. Specifically, because within-subject sampling designs typically are more efficient than are between-subject sampling designs – inasmuch as they tend to minimize error variance (Maxwell & Delaney, 1990) – they tend to yield larger effect sizes (Keppel, 1991; O'Grady, 1982). Therefore, in interpreting effect sizes, consideration should be given to the sampling design used.

Although experimental studies allow the strongest causal inferences to be made and typically result in relatively smaller error variance in comparison to correlational studies, experimental designs also tend to yield smaller effect sizes than do correlational designs. This is because in experimental research the independent variable is artificially created specifically for the study and thus is weaker than it is in the population (Kerlinger, 1973). As such, comparing effect sizes stemming from experimental studies and those generated from correlational studies easily can be the equivalent of comparing apples and oranges. Moreover, in fixed-effects models, the magnitude of the omnibus effect size depends on the specific levels of the variables of interest. If different levels of the independent variable are studied, the effect sizes are not comparable (Oljenik & Algina, 2000).

Further, the number of experimental conditions (or levels of the independent variable) used in a study can either increase or decrease the effect size. O'Grady (1982, p. 773) provides a striking example of a two-conditions study (yielding $M_1 = 10$ and $M_2 = 18$, with common *SD*s of 2 and *n*s of 10) in which the proportion of variance accounted for by the treatment factor (sample η^2) is .82. Yet, had the same two conditions been part of a study that also included three additional experimental conditions, whose resulting means ranged in equal increments between the two original means (i.e., $M_3 = 12$, $M_4 = 14$, and $M_5 = 16$), with the same *SD*s and *n*s as before, the proportion of variance accounted for by the treatment factor is reduced to .69. Of course, had the proportion of variance associated with *just* the two focal conditions been calculated and reported (i.e., the sample η^2 associated with the Treatment 1 vs. Treatment 2 *contrast*), it would be equal to the original .82.

Interpretive problems resulting from omnibus, as opposed to contrast, strength-of-relationship reporting were pointed out by Levin (1967). Such problems can be further illustrated by another hypothetical example, which represents the "flip side" of the one just presented. Suppose that a researcher compares two different experimental treatments and finds that $M_1 = 16$ and $M_2 = 17$, with common *SD*s of 2.5 and *n*s of 8. Here, the sample η^2 can be found to be a fairly "small" .04. However, had these two treatments

been part of a study that included a low-scoring "control" group (M_3 = 6) with the same *SD* and *n* as in the other two conditions, now the sample η^2 would be found to leap to an "impressive" .81. As long as the researcher focused on the Treatment 1 vs. Treatment 2 contrast (for which η^2 = .04), the same conclusion about a "small" treatment difference would have been reached as before. Unfortunately, however, many researchers routinely report and interpret the omnibus measure (here, η^2 = .81), to the detriment of the unquestioning consumer. In multifactor designs a similar opportunity arises for misleading the consumer – namely, by not recognizing Kirk's (1995, p. 261) distinction between omnibus and partial strength-of-relationship measures.

The design of an experimental study also refers to the manner in which participants are assigned to experimental conditions and treatments administered (generally characterized as between-subjects designs, within-subjects designs, mixed designs, blocking designs, and hierarchical designs), whether or not concomitant variables (covariates) are included, and the statistical analyses employed. Effect-size measures are affected by all such factors in a design, compromising comparisons of effect sizes across studies that differ in their specifics (Oljenik & Algina, 2000).

In particular, when one or more factors in a comparison-of-means analysis represents an individual difference factor (e.g., a covariate or blocking variable), problems arise with respect to what to use as the standardizer in an effect-size index. For example, in a two-factor design in which one factor is a manipulated factor and the other an individual difference factor, it is often a matter of debate whether the standardizer should be computed by ignoring or controlling for the individual difference factor (Oljenik & Algina, 2000). Whichever approach is taken leads to a different effect size being computed and, therefore, effect sizes using these two different standardizers are not comparable. In fact, as noted by Oljenik and Algina (2000): "depending on the sample size and effect sizes associated with the individual difference and interaction factors in a two-factor design, the effect size estimated for the manipulated factor can vary from trivial to quite large" (p. 250).

The difference in effect sizes is even greater if the individual difference factors vary across studies. Because varying standardizers for computing effect sizes are used in different studies, researchers should compare effect sizes only if they are completely aware of the standardizer that was used in each study of interest. Unfortunately, most researchers do not specify which standardizer was used in their effect-size computation. This discussion should make it clear that a researcher can make an effect size look larger or smaller by defining an effect size in terms of the specific design and control-variable characteristics just mentioned – basically, by incorporating (or not) any design features that serve to affect the error variance – and which may have ethical implications as well.

3. Selection of an effect-size measure. We now turn our attention to another potentially ethically sensitive effect-size issue. Although there is general agreement that the provision of effect-size information is valuable, recommendations concerning the specific measure that should be reported for a particular study are typically absent. In our view, such recommendations are critical, for as one of us noted previously:

> Which of, say, half a dozen different effect-size measures that could be summoned up for a given problem should a researcher report? The one that is most informative, the one that is most conservative, or the one that enhances the researcher's case and misleads the unsuspecting reader? For example, researchers might report percent agreement measures or percentages of variance accounted for that have not been corrected for chance, or researchers might seek out a goodness-of-fit measure that places their data in the most favorable light. For dependent measures where a frame of reference is needed or helpful, providing scale-free (relative) effect sizes (e.g., Cohen's *d* or percentages of variance accounted for) is not nearly as substantively

> interpretable as is providing the scale-dependent (absolute) measures in addition or instead...In many domains, not even knowledgeable statisticians agree on what the "best" or "most informative" effect-size measure actually is. (Levin & Robinson, 1999, p. 151)

Levin (1998b, pp. 45-46) similarly provided the following hypothetical example of the perplexing situation that effect sizes can create for researchers, readers, and other interpreters of the importance of an empirical finding:

> Suppose that an investigator wants to help older adults remember an ordered set of ten important daily tasks that must be performed (insert and turn on a hearing aid, take certain pills, make a telephone call to a caregiver, etc.). In a sample of six elderly adults, three are randomly assigned to each of two experimental conditions. In one condition (A), no special task instruction is given; and in the other (B_1), participants are instructed in the use of self-monitoring strategies. Following training, the participants are observed with respect to their success in performing the ten tasks...[T]he average number of tasks the participants correctly remembered to perform was 1.33 [*SD* = .577, raw scores = 1, 1, and 2] and 3.33 [*SD* = .577, raw scores = 3, 3, and 4] for the no-instruction (A) and self-monitoring (B_1) conditions, respectively. For [these data], it can be determined that the "conditions" factor accounts for a hefty 82% of the total variation in task performance (i.e., the squared point-biserial correlation is .82, which for the two-sample case, is equivalent to the sample η^2). Alternatively, the self-monitoring mean is 3-½ within-group standard deviations higher than the no-instruction mean (i.e., Cohen's *d* is 3.5). From either effect-size perspective (η^2 or *d*), certainly this represents an impressive treatment effect, doesn't it? Or does it?
>
> Suppose that instead of self-monitoring training, participants were taught how to employ "mnemonic" (systematic memory-enhancing) techniques (B_2) ...with the results [yielding a mean number correct of 7.67 *(SD* = 2.517, raw scores = 5, 8, and 10)]...[A] comparison with no-instruction Condition A surprisingly reveals that once again, the conditions factor accounts for 82% of the total variation in task performance (equivalently, *d* again equals 3.5). Thus, when expressed in standardized/relative terms (either η^2 or *d*), the effect sizes associated with the two instructional conditions (B_1 and B_2) are exactly the same, and substantial in magnitude. Yet, when expressed in absolute terms and with respect to the task's maximum, there are important differences in the "effects" of B_1 and B_2: Increasing participants' average performance from 1.33 to 3.33 tasks remembered seems much less impressive than does increasing it from 1.33 to 7.67. Helping these adults remember an average of only 3 of their 10 critical tasks might be regarded as a dismal failure, whereas helping them remember an average of almost 8 out of 10 tasks would be a stunning accomplishment. Yet, the conventional effect-size measures are the same in each case.

To help shed light on this seeming paradoxical situation, Levin (1998b) pointed out:

> The major problem in this example arises from the conditions' differing variabilities. That problem could be accounted for by defining alternative *d*-like effect-size measures based on just the control condition's (Condition A's) standard deviation... Interpreting effect sizes, in the absence of raw data, remains a problem for η^2 and Cohen's *d*, however. (p. 53)

Insofar as different effect-size measures are suitable for different types of data (e.g., Hogarty & Kromrey, 2001), it is surprising that some researchers do not even indicate the index to which they are referring when reporting effect sizes (Kirk, 1996). Neither do researchers appear to indicate whether the effect-size measure interpreted represents an adjusted or unadjusted index. The lack of information provided is disturbing because meta-analyses involve aggregating and comparing effect sizes across studies. How can effect sizes be aggregated if it is not clear whether they are based on the same type of index? Unfortunately, the practice of some meta-analysts to omit unlabeled effect sizes from the aggregate index introduces bias.

4. Varying, and generally arbitrary, guidelines for interpreting effect-size magnitudes. As was noted earlier, a way in which statistical hypothesis testing is abused occurs when a dichotomous decision (i.e., reject vs. do not reject) comprises the sole determinant of the significance (read importance) of an observed outcome. This is done by comparing the outcome's significance probability (*p*-value) to some predetermined standard significance level (α level), such as .05. Yet, many researchers who interpret effect sizes appear to use equally rigid categorical criteria such as those provided by Cohen (1988), who popularized the use of effect-size reporting. This occurs even though recommendations vary with respect to how effect sizes should be interpreted (McLean, O'Neal, & Barnette, 2000) and despite Cohen's (1988) admonishment that effect-size values are dependent on the specific content and methods that prevail in a given research context.

For example, in interpreting effect sizes associated with differences between two groups (i.e., Cohen's *d*), Cohen (1988) recommended demarcations of .20 for small effects, .50 for medium effects, and .80 for large effects. In stark contrast, McLean (1995) suggested the following criteria: .50 for small effects, between .50 and 1.00 for moderate effects, and above 1.00 for large effects. Regardless of which criteria are used, it is clear that adherence to such cutpoints has the effect of trichotomizing interpretations in much the same way as *p*-values dichotomize statistical decision making. As noted by Shaver (1993): "There already is a tendency to use criteria, such as Cohen's (1988) standards for small, medium, and large effect sizes, as mindlessly as has been the practice with the .05 criterion in statistical significance testing" (p. 311). Similarly, Thompson (2001) stated: "If people interpreted effect sizes [using fixed benchmarks] with the same rigidity that $\alpha = .05$ has been used in statistical testing, we would merely be being stupid in another metric" (p. 82-83).

In addition, blending the previous concern (different effect-size measures may lead to different conclusions) with the present one (effect-size descriptors are arbitrary and vary by context) we consider the following confusing/conflicting medical-study conclusion presented by Rosenthal and DiMatteo (2001). The results of a study designed to examine the effect of taking aspirin on heart-attack prevention (Steering Committee of the Physicians' Health Study Research Group, 1988) yielded what is typically regarded as a tiny Pearson *r* of .034. Yet, when the same outcome is interpreted from the perspective of Rosenthal and Rubin's (1982) binomial effect size display (BESD), the "finding is, in fact, very important and translates into substantial reductions in morbidity and mortality" (Rosenthal & DiMatteo, 2001, p. 78). For related discussion on the potential importance of conventionally small effect sizes, see Prentice and Miller (1992).

5. Sample size and sampling variability. The interpretation of effect sizes also varies as a function of sample size. Studies with smaller sample sizes often result in effect sizes being overestimated, whereas investigations with large sample sizes tend to lead to effect sizes being

underestimated (Bakan, 1966; Fern & Monroe, 1996; Hedges & Olkin, 1985). Empirically, Barnette and McLean (1999) demonstrated that standardized effect-size variation is systematic rather than random. In their Monte Carlo investigation, these authors found that the number of groups and sample sizes were almost perfectly predictive (i.e., R^2 = .999) of standardized effect sizes. Thus, comparing effect sizes across studies with very different sample sizes can be misleading.

One of the most repeated criticisms of statistical hypothesis testing is its over-reliance on sample size (Cohen, 1994; Fan, 2001; Kirk, 1996; Onwuegbuzie & Daniel, 2003, in press; Schmidt & Hunter, 1997; Thompson, 1993). Yet, as was noted recently by Fan (2001): "effect size can also be misleading because sample size influences the sampling variability of an effect-size measure" (p. 275). Using Monte Carlo methods, Fan demonstrated that an observed finding that appears to have practical significance (i.e., a large effect size) actually could be the result of sampling error, thereby making any resultant conclusions unreliable and potentially misleading – which lends empirical support to a major facet of the argument promoted by Levin and Robinson (2000; see also Sawilowsky & Yoon, 2002), summarized later. Fan (2001) recommended that information about both statistical significance and effect sizes be reported for observed findings:

> Statistical significance testing and effect size are two related sides that together make a coin; they complement each other but do not substitute for one another. Good research practice requires that, for making sound quantitative decisions in educational research, both sides should be considered. (p. 275)

It should come as no surprise that effect sizes are affected by sample size in much the same way as are *p*-values. Indeed, effect-size statistics represent random variables. Consequently, effect-size measures are affected by sampling variability, as dictated by its underlying sampling distribution. In turn, the amount of sampling variability of an effect-size estimate is influenced by the underlying sample size, in much the same way that *p*-values are affected by the number of cases utilized in the study. When the sample size is small, the discrepancy between the sample effect size and population effect size is larger (i.e., large bias) than when the sample size is large. Also, effect sizes are affected by nonrandom sampling, a condition that applies to the vast majority of empirical studies in education and psychology. Thus, solutions to compensate for the problems stemming from the role of sample size in statistical hypothesis testing (e.g., use of confidence intervals) should also apply to effect sizes.

A valid criticism of hypothesis testing that is supported by data pertains to the low statistical power that prevails in many studies. Indeed, the average power of null hypothesis significance tests typically ranges from .40 to .60 in empirical studies (Cohen, 1962, 1965, 1988, 1994; Schmidt, 1996; Sedlmeier & Gigerenzer, 1989). With an estimated mean across-study power of .50 (Cohen, 1962, 1997), Schmidt and Hunter (1997) decry that "[t]his level of accuracy is so low that it could be achieved just by flipping a (unbiased) coin!" (p. 40). Yet, the finding that power is unacceptably low in most studies indicates to us that researchers' application of statistical hypothesis testing, rather than its logic, is to blame. Indeed, it can be argued that low statistical power represents more of a research design issue than a statistical issue, since acceptable power can be rectified by incorporating a larger sample.

Unfortunately, as was discussed earlier, effect sizes also can fall victim to poor research designs, in general, and to small sample sizes, in particular. In fact, an obsession with effect sizes without considering the associated sample sizes can have the effect of promoting weak research designs. As such, in making decisions about which articles should be published, journal editors should focus less on *p*-values and effect sizes and more on the quality of the underlying research design (for related discussion and references, see Levin, 1998b, p. 45).

6. Distribution nonnormality. Although this may surprise or disturb some readers, many of the commonly used effect-size measures rely heavily on the parametric hypothesis-testing assumptions of normality and homogeneity of variance (see, for example, Fan, 2001, Barnette & McLean, 1999, and Hogarty & Kromrey, 2001).

The numerator of common effect-size measures involves means and mean differences, which are sensitive to extreme observations, especially when sample sizes are small (Huck, 2000). In the small-sample case, an extreme observation in one of the conditions (e.g., the experimental group) can seriously distort the true mean difference, thereby unduly influencing the effect-size estimate. Just as outlying observations affect the *t*-statistic and associated *p*-values (statistical significance), in the independent-samples test of means they also influence the effect size (practical significance). For this reason, nonparametric effect-size measures have been developed and considered.

Applyng Monte Carlo methods, Hogarty and Kromrey (2001) demonstrated that the most frequently used effect-size estimates (e.g., Cohen's *d* and Hedges & Olkin's *g*) are sensitive to departures from normality and variance homogeneity (discussed next). Even trimmed effect-size measures (Hedges & Olkin, 1985; Yuen, 1974) exhibit bias when sample sizes are small, as do several nonparametric effect-size indices, including Y_1 (Kraemer & Andrews, 1982) and the Common Language (CL) effect-size statistic (McGraw & Wong, 1992).

7. Score variability (both between and within samples). Other characteristics of the sample also affect interpretation of effect sizes. In particular, the more heterogeneous the sample is with respect to the variable of interest, the greater the effect size typically tends to be. This is the case for both explanatory and predictive studies (O'Grady, 1982). Moreover, homogeneous samples, which more often arise from convenience sampling, can result in range restriction and, subsequently, attenuate effect sizes (Pedhazur & Schmelkin, 1991). Recognition of this complicating situation can be seen in a recent critique of a report challenging the effectiveness of teacher education programs by Darling-Hammond and Youngs (2002):

> The effect size also depends on other context factors, such as the range of variability in the measure used, which can change in different locations and time periods. For example, in some eras and in some locations virtually all teachers held content degrees or were fully certified, so these variables do not strongly predict variations in outcomes. When much more variability is present, these variables are strongly predictive of outcomes. Thus, several studies have found strong measured influences of certification status on student achievement in states like California and Texas during the 1990s when there were wide differences in teachers' qualifications. (p. 15)

It is also possible for variance heterogeneity to reduce the effect size. This can be the case when the sample is *too* diverse and the heterogeneity increases error variance, thereby attenuating the effect size (Lesser, 1959).

Regardless of whether the effect size is increased or decreased by heterogeneous samples, interpreting effect sizes that arise from samples with different degrees of heterogeneity is inadvisable. In particular, researchers should exercise caution in comparing effect sizes across convenience samples. In fact, Daniel and Onwuegbuzie (2000) refer to sampling bias error that results in inconsistency of results across studies as a Type IX error. According to these authors, this type of error relates to "disparities in results generated from numerous convenience samples across a multiplicity of similar studies" (p. 23).

Further, because the denominator of common effect-size measures incorporates the pooled within-conditions variance, heterogeneity of variance affects effect-size estimation similarly to the way that it affects statistical hypothesis testing (and confidence-interval building - as was seen in Levin, 1998b, p. 53). Moreover, the problems caused by departures from normality and heterogeneity of variance when statistical significance testing is involved are very much an issue for effect-size measures associated with more complex family members of the general linear model. For example, the standard effect-size indices (e.g., η^2, ϵ^2, and ω^2) that are often calculated for OVA-type analyses (e.g., ANOVA, ANCOVA, MANOVA) assume equal variances –

an assumption that is not always met (Onwuegbuzie & Daniel, 2003).

However, these weaknesses do not imply that effect sizes should be banned or replaced by some other sort of index, echoing what some researchers (e.g., Carver, 1993) recommend should be the fate of statistical significance testing. Indeed, in cases where such violations come to the fore, nonparametric effect sizes (e.g., Y_1 and CL) may be more appropriate, in much the same way that nonparametric inferential statistics often are more appropriate when the parametric assumptions are violated. The above limitations pertaining to effect sizes identified above suggest that: (*a*) assumptions underlying the selected effect-size method should be subjected to the same stringent scrutiny as are statistical significance tests; (*b*) combining statistical significance testing and effect-size indices, after checking all pertinent assumptions, provides an additional safety net from false or misleading conclusions, compared to using either technique alone; and (*c*) researchers should pay much more attention to maximizing the quality of their research designs (e.g., by selecting an appropriate or optimal sample size) in order to minimize threats to the model assumptions that pertain to both the statistical test and the accompanying effect-size measure of interest.

8. Reliability of the outcome measure (measurement error). Reliability is a concept that receives disproportionately scant attention in the interpretation of an observed finding (Onwuegbuzie & Daniel, 2000, 2001, 2003, in press; Onwuegbuzie, Daniel, & Roberts, in press; Roberts & Onwuegbuzie, 2003; Roberts, Onwuegbuzie, & Eby, 2001; Onwuegbuzie & Weems, in press; Weems & Onwuegbuzie, 2001). Reliability (or more precisely, unreliability) can adversely affect the internal validity of findings via "instrumentation" problems (e.g., Campbell & Stanley, 1963; Onwuegbuzie, 2003), through a reduction in statistical power. Specifically, Onwuegbuzie and Daniel (in press) demonstrated that subgroups with scores that generate markedly different reliability estimates can seriously reduce statistical power, even when the full-sample (i.e., across-groups) reliability coefficient is adequate.

Importantly, however, low reliability indices adversely affect not just statistical hypothesis testing; they also negatively impact effect-size measures. After all, low reliability coefficients stem from scores that do not behave in a consistent manner (Onwuegbuzie & Daniel, 2000, 2001) and it is these scores that are used to calculate both inferential test statistics and effect-size measures. Thus, effect-size measures are subject to the same limitations stemming from inadequate reliability as are *p*-values. Indeed, effect sizes should always be interpreted with respect to the reliability of the outcome measure, just as has been recommended for statistical hypothesis testing.

Specifically, there is an inverse relationship between the reliability of any of the variables of interest (whether the independent or dependent variable) and the corresponding effect size. In fact, such reliability provides an upper bound for the effect size (Lord & Novick, 1968; Nunnally & Bernstein, 1994). Because a study's reliability is a function of the study's obtained scores rather than *a priori* test norms (Onwuegbuzie & Daniel, 2000, 2002a, 2002b; Thompson & Vacha-Haase, 2000; Vacha-Haase, Kogan, & Thompson, 2000; Wilkinson & Task Force on Statistical Inference, 1999), effect sizes should not be compared across studies without taking into account the individual studies' outcome-measure reliabilities. For further discussion of reliability and effect size in both correlational and experimental study contexts, see O'Grady (1982, pp. 767-770).

9. Scale of measurement. The type and range of measure used can affect the size of the effect. It is not unusual for researchers studying a phenomenon to use different measures. In particular, in a study of an affective variable, whereas one researcher might use a Likert-type scale, another researcher might employ a rating scale. Still another researcher might employ a semantic differential scale or a Thurstone or Guttman scale. Similarly, in an investigation of a cognitive outcome, whereas one researcher might administer a multiple-choice test, another researcher might administer some other type of closed-ended instrument (e.g., true-false, matching), and still another researcher might administer an open-ended measure such as an essay.

Although all of these measures yield scores that can be analyzed statistically, each type of scale might not be measuring exactly the same construct. For instance, multiple-choice and essay

examinations often target different levels of learning in Bloom's taxonomy of cognitive objectives (Bloom, 1956). As such, the effect size likely would vary as a function of the type of measure used. Although this apples-and-oranges situation is typically offered as the primary rationale for meta-analytic effect-size *combinations* (e.g., Hunt, 1997; Rosenthal & DiMatteo, 2001), it rarely is recognized as a study-*comparison* concern.

Even if scales with the same item format (e.g., a Likert-scale format) are used across studies, both the number/type of items and the number/type of response options employed can affect the size of the effect. With respect to the former, compared to their counterparts with more items, scales with a smaller number of items lead to restriction of range, thereby attenuating effect sizes. Similarly, the proportion of negatively worded and positively worded items can influence the effect size (Onwuegbuzie & Weems, in press; Weems & Onwuegbuzie, 2001). With regard to the latter, the number of response options also can influence the effect size. Specifically, a reduction in the number of response options attenuates the range of scores, which, in turn, may reduce the magnitude of the effect.

Similarly, and as was mentioned earlier, a restriction in the variability of one or more variables typically decreases the effect size. This holds for a study's independent variables, as well as its outcome measures. As noted by Onwuegbuzie and Daniel (2003), lacking the realization that nearly all parametric analyses represent the general linear model, many analysts inappropriately categorize independent variables in nonexperimental research designs in order to perform analyses such as analysis of variance. Disturbingly, findings from such analyses are then used to make causal inferences, when all that has occurred is a discarding of relevant variance – see, for example Cliff (1987); Pedhazur (1982); Prosser (1990); and Thompson (1986, 1988, 1992a).

Yet, categorizing a continuous variable has been found repeatedly to reduce the effect size. For instance, a median split of a continuous variable can reduce the observed correlation by 20% (Cohen, 1983; Hunter & Schmidt, 1990) – see also Vargha, Rudas, Delaney, & Maxwell (1996). If the cutpoint used for splitting the continuous variable differs from the median, then the reduction in the relationship between the variables can be expected to be even larger (Fern & Monroe, 1996).

Moreover, as the number of categorized groups decreases, less variance in the dependent variable is accounted for by the categorical variable, compared to the continuous variable, and thus the effect size is attenuated (Peet, 1999). With regard to type of response options, the use of midpoint categories (e.g., neutral response options) has been found to affect both score reliability and effect size (Weems & Onwuegbuzie, 2001). Therefore, comparing effect sizes across studies using different types and formats of scales is questionable.

In addition, it does not appear to be obvious to some researchers that effect sizes are a function of the scale of measurement used. Evidence of this is provided by McLean et al. (2000), who demonstrated that "gain" effect sizes were different for the raw scores, scaled scores, and Normal Curve Equivalent (NCE) scores for students in Grades 4, 6, and 8 on a national norm-referenced test. Specifically, as McLean et al. expected, the effect sizes for NCE scores were lower than those for raw and scaled scores. The researchers appropriately concluded that when effect sizes are computed, researchers should take into account the scale of measurement on which they are based.

Summary

We have highlighted nine general concerns about effect-size indices. When researchers design their studies, they must make numerous decisions. Each of these decisions can affect the magnitude of the effect-size estimate. Unfortunately, the extent to which the effect-size index is influenced by the decisions is almost always unknown. This suggests that researchers are not justified in reflexively applying Cohen's (1988) effect-size magnitude and adjectival guidelines across studies in different domains or across studies that have different research design and analytical factors. Even more importantly, because effect sizes vary as a function of research-related factors, effect sizes should be compared only when all of these factors are comparable. Assessing the substantive significance of an observed finding based solely on the effect size

may be misleading and no more diagnostic than is a test of a statistical hypothesis (Fern & Monroe, 1996).

This does not mean that effect sizes are useless. As noted by Fern and Monroe (1996), if the goal of the researcher is to determine the size of an effect given the unique combination of factors that underlie the data, then a computed measure of effect size is informative. On the other hand, effect sizes cannot be used as a meaningful basis for comparison across studies "unless the researcher understands what, if any, unique factors contributed to the effect-size estimate" (Fern & Monroe, 1996, p. 102). In any case, when reporting effect sizes, researchers should always specify as many design, analysis, and psychometric characteristics as possible to help subsequent researchers decide the extent to which they can compare their effect sizes with previous estimates. In other words, researchers should contextualize their effect sizes (i.e., they should interpret their effect sizes within study's specific parameters).

Many researchers who criticize statistical hypothesis testing, in general, and those who advocate replacing *p*-values with effect size measures, in particular, fail to mention any of the limitations associated with effect-size reporting. Thus, methodologists who criticize hypothesis testing without also discussing the limitations of effect sizes are not providing a balanced analysis but are focusing on the bad practices that have traditionally been linked to the former approach. Unfortunately, the just-mentioned concerns about effect sizes typically are not mentioned by their advocates. In discussing the limitations, we argue that effect sizes are not the hoped-for panacea for empirical research in the social sciences.

Further, we contend that if *only* effect sizes were used to interpret statistical data, social-science research would not be in any better position than it would if only statistical hypothesis testing were used in quantitative studies. In fact, in an effect-size-only world, we submit that social-science research would be in a *worse* position, in that progress would be retarded (Thompson, 1992b) to an even greater extent than that imagined by hypothesis-testing critics, in that statistically "chance" findings would unjustifiably be promoted by researchers as "real." We reconsider that unfortunate situation in the following concluding section.

Toward a Détente

The effect-size flaws that we have reviewed support the assertion that statistical hypothesis testing and effect-size reporting should be used in combination. A logical, internally consistent, way of combining these two procedures is through Robinson and Levin's (1997) two-step suggestion for analyzing empirical data – namely, that effect sizes are reported if and only if the observed finding is statistically significant.

That is, statistical hypothesis testing should serve as a gatekeeper, guarding against spurious effect-size estimation. As noted by Robinson and Levin (1997), the goal of these two complementary approaches is to prevent the over-interpretation of seemingly impressive effect sizes "in the absence of formal assessments of their likelihood" (p. 23). We therefore recommend that statistical hypothesis testing and effect-size estimation be used in tandem to establish a reported outcome's believability and magnitude, respectively. As such, tests of significance serve a valuable purpose in determining whether effect-size measures should be ignored or reported, a position endorsed by Fan (2001), Levin (1993), Robinson and Levin (1997), Knapp and Sawilowsky (2001), and even – we think – Gliner, Leech, and Morgan (2002).

Let us take a moment to consider the last part of the foregoing sentence. We say "even" because Gliner et al.'s recommendation appeared in a journal whose editorial policy specifically calls for effect-size inclusions even in the absence of statistical confirmation: "Furthermore, authors are required to report and interpret magnitude-of-effect measures in conjunction with every *p* value that is reported" (Journal of Experimental Education, 2002, p. 94). We say "we think" because Gliner et al. are internally inconsistent in their position about *always* reporting and interpreting effect sizes in their position.

For example, they agree with Levin and Robinson's (2000) distinction between single-study investigations and multiple-study syntheses: "Our opinion is that effect sizes should accompany all reported *p* values for possible future meta-analytic use, but they should not be presented as

findings in a single study in the absence of statistical significance" (Gliner et al., 2000, p. 86). Yet, in the penultimate sentence of their article they write: "We also recommend reporting effect size for nonsignificant outcomes" (p. 91). Addressing this blanket effect-size reporting recommendation, one of us has pointed out previously:

> This practice is absurdly pseudoscientific and opens the door to encouraging researchers to make something of an outcome that may be nothing more than a "fluke," a chance occurrence. Without an operationally replicable screening device such as statistical hypothesis testing, there is no way of separating the wheat (statistically "real" relationships or effects) from the chaff (statistically "chance" ones), where "real" and "chance" are anchored in reference to either conventional or researcher-established risks or "confidence levels."...In its extreme form, effect-size-only reporting degenerates to strong conclusions about differential treatment efficacy that are based on comparing a single score of one participant in one treatment condition with that of another participant in a different condition. (Levin, 1998b, p. 45)

Moreover, in a recent survey of the editorial board members of four educational-research journals (Capraro & Capraro, 2003), the 97 respondents (estimated from the data provided) greeted the recommendation that their journals *require* effect-size reporting with overwhelming indifference: On a 7-point Likert scale ranging from "very strongly disagree" to "very strongly agree" the mean rating was 4.26, $t(96) = 1.33$, $p = .19$, for testing the hypothesis that respondents' mean ratings do not differ from the scale midpoint of 4. Given the study's relatively large sample size, this nonrejection of the indifference hypothesis should be taken with more than a grain of salt.

One additional internal-inconsistency irony – or at least an example of journal non-policing – is worth mentioning. In an article published by one of the present authors (Hwang & Levin, 2002) in the same issue of the *Journal of Experimental Education* that proclaims the above effect-size policy, effect sizes were *not* reported for every *p*-value included; nor were they reported for statistically nonsignificant outcomes. Yet, somehow, some way, the article was published anyway! And this is not an isolated event.

A colleague, Dan Robinson, has experienced effect-size nonenforcement with two of his articles that were published in the same journal (Katayama & Robinson, 2000; Robinson, Katayama, Dubois, & Devaney, 1998), a journal that has promoted its effect-size policy since 1997 (D. H. Robinson, personal communication, January 13, 2003). As with Thompson's (e.g., 1996) argument in other contexts, perhaps *JEE* should be *encouraged* to take a closer look at its own editorial policy, for in that journal effect-size endorsement clearly does not translate into effect-size enforcement. As an informative aside, the *Journal of Experimental Education* is apparently not alone in its effect-size non-enforcement practices for D. H. Robinson (personal communication, January 22, 2003) indicates a similar phenomenon with another effect-size mandated journal, *Contemporary Educational Psychology.* Out of 11 intervention experiments that he tallied for that journal in 2001, only two were accompanied by effect-size estimates.

Even those who contend that effect sizes should *replace* statistical significance testing (e.g., Carver, 1993; Schmidt, 1996) recommend the use of confidence intervals alongside effect sizes. A two-sided confidence interval, characterized by lower and upper bounds, identifies a probable range of magnitudes for the effect size (Abelson, 1997). As such, confidence intervals can be used to estimate the range of the effect's practical significance – for related discussion, see Onwuegbuzie (2001) and Thompson (2002).

Moreover, insofar as confidence intervals include all the information provided by statistical hypothesis tests, and more (Cohen, 1994; Levin, 1998b; Serlin, 1993), constructing them allows researchers to conduct the corresponding hypothesis tests, if desired (Krantz, 1999). In that sense, then, the provision of an inferential

confidence interval (instead of a hypothesis test) has logical appeal because that approach kills two birds (statistical and practical significance) with one stone. So as not to confuse the issue, it should be made clear that the kind of confidence-interval approach we are endorsing is the single-interval procedure based on a pre-experimentally established Type I error probability, which is inferentially equivalent to applying a Neyman-Pearson statistical test of hypothesis. This approach is fundamentally and logically different from that espoused by certain hypothesis-testing critics, which would have researchers simultaneously provide multiple confidence intervals (for either raw or standardized effects) based on different confidence levels, such as 99%, 95%, 90%, 80%, etc. – see, for example, Schmidt & Hunter (1997) and Thompson (2002).

Alternatively, hypothesis testing *per se* can be substantially improved (strengthened) by applying it in forms that are more intelligent than the one that is currently practiced. Such more intelligent forms call for researchers to formulate/test more theoretically driven and precise hypotheses, to determine (through power calculations) optimal sample sizes to test those hypotheses, and to incorporate equivalence-testing procedures (e.g., Seaman & Serlin, 1998) for better establishing the truth of the null hypothesis (see, for example, Levin, 1998a, pp. 329-330).

At the same time, we contend that hypothesis tests, confidence intervals, and effect sizes do not go far enough in the way of maximizing a domain's knowledge base. This can be accomplished only through independent replications of results (i.e., two or more independent studies yielding similar findings that produce statistically and substantively compatible outcomes). We believe that "a replication is worth a thousandth *p* value" (Levin, 1995), as well as its being worth more than a large effect size based on a single study. In contrast to Carver (1978), however, we do not believe that "replicated results should automatically make statistical significance unnecessary" (p. 393). Such independent replications not only will make "invaluable contributions to the cumulative knowledge in a given domain" (Robinson & Levin, 1997, p. 25) but will also help empirical researchers achieve a common goal.

Conclusion

As was noted by Onwuegbuzie (2003), a primary objective of empirical research – especially research designed to posit causal relationships – is to collect and analyze data that help a researcher make inferences from the sample(s) to the underlying population, leading to meaningful conclusions in which as many rival explanations as possible are eliminated. This is the goal that drives both statistical hypothesis testing and effect-size reporting. The extant literature has documented the limitations of hypothesis testing, whereas in this paper we have illustrated that effect-size interpretation is not without its flaws. No single index by itself is the magic bullet for analyzing and interpreting data. Rather, using both methods in combination, or combining confidence intervals and effect sizes, helps to rule out more rival threats to statistical-conclusion validity (Cook & Campbell, 1979; Shadish, Cook & Campbell, 2002) than would occur if either method were used alone to interpret observed findings. At the same time, however, to minimize both statistical-conclusion validity and external validity threats there is no substitute for independent replications.

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?* Mahwah, NJ: Erlbaum, p. 117-141.

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

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

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin, 66*, 423-437.

Barnette, J. J., & McLean, J. E. (1999, November). *Empirically based criteria for determining meaningful effect size.* Paper presented at the annual meeting of the Mid-South Educational Research Association, Point Clear, AL.

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.

Berkson, J. (1942). Tests of significance considered as evidence. *Journal of the American Statistical Association*, *37*, 325-335.

Bloom, B. S. (Ed.). (1956). *Taxonomy of educational objectives: The classification of educational goals, Handbook I: Cognitive domain*. New York: Longman, Green.

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

Cahan S. (2000). Statistical significance is not a "Kosher Certificate" for observed effects: A critical analysis of the two-step approach to the evaluation of empirical results. *Educational Researcher, 29*(1), 31-34.

Calder, B. J., Phillips, L. W., & Tybout, A. M. (1981). Designing research for application. *Journal of Consumer Research*, *8*, 197-207.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.

Capraro, R. M., & Capraro, M. M. (April, 2003). *Exploring the APA fifth edition* Publication Manual's *impact on the preferences of journal editorial board members.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

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

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

Cliff, N. (1987). *Analyzing multivariate data*. San Diego: Harcourt Brace Jovanovich.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal Psychology*, *65*, 145-153.

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

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement, 7*, 249-253.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* New York: John Wiley.

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

Cohen, J. (1997). The earth is round ($p < .05$). In L.L. Harlow, S.A. Mulaik, & J.H. Steiger (Eds.), *What if there were no significance tests?* Mahwah, NJ: Erlbaum, p. 117-141.

Cook, T. D & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago: Rand-McNally.

Daniel, L. G., & Onwuegbuzie, A. J. (2000, November). *Toward an extended typology of research errors.* Paper presented at the annual meeting of the Mid-South Educational Research Association, Bowling Green, KY.

Darling-Hammond, L., & Youngs, P. (2002). Defining "highly qualified teachers": What does "scientifically-based research" actually tell us? *Educational Researcher*, *31*(9), 13-25.

Falk, R., & Greenbaum, C. W. (1995). Significance tests die hard: The amazing persistence of a probabilistic misconception. *Theory & Psychology*, *5*, 75-98.

Fan, X. (2001). Statistical significance and effect size in education research: Two sides of a coin. *Journal of Educational Research, 94*, 275-282.

Fern, E. F., & Monroe, K. B. (1996). Effect-size estimates: Issues and problems in interpretation. *Journal of Consumer Research*, *23*, 89-105.

Fisher, R. A. (1925/1941). *Statistical methods for research workers* (84th ed.) Edinburgh, Scotland: Oliver & Boyd. (Original work published in 1925).

Frick, R. W. (1995). *Using statistics: Prescription versus practice.* Unpublished manuscript, Department of Psychology, State University of New York at Stony Brook.

Gliner, J. A., Leech, N. L., & Morgan, G. A. (2002). Problems with null hypothesis significance testing (NHST): What do the textbooks say? *Journal of Experimental Education*, *71*, 83-92.

Guttman, L. B. (1985). The illogic of statistical inference for cumulative science. *Applied Stochastic Models and Data Analysis, 1*, 3-10.

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

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

Hogarty, K. Y., & Kromrey, J. D. (2001, April). *We've been reporting some effect sizes: Can you guess what they mean?* Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.

Huck, S. W. (2000). *Reading statistics and research* (3rd ed.). New York: Addison Wesley Longman.

Hunt, M. (1997). *How science takes stock: The story of meta-analysis.* New York: Russell Sage.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings.* Newbury Park, CA: Sage.

Hwang, Y., & Levin, J. R. (2002). Examination of middle-school students' independent use of a complex mnemonic system. *Journal of Experimental Education*, *71*, 25-38.

Keppel, G. (1991). *Design and analysis: A researcher's handbook*. Englewood Cliffs, NJ: Prentice-Hall.

Katayama, A. D., & Robinson, D. H. (2000). Getting students "partially" involved in note-taking using graphic organizers. *Journal of Experimental Education*, *68*, 119-133.

Kerlinger, F. N. (1973). *Foundations of behavioral research*. New York: Holt, Rinehart, & Winston.

Kirk, R. E. (1995). *Experimental design: Procedures for the behavioral sciences* (3rd ed.). Pacific Grove, CA: Brooks/Cole.

Kirk, R. E. (1996). Practical significance. A concept whose time as come. *Education and Psychological Measurement, 56*, 746-759.

Knapp, T. R., & Sawilowsky, S. S. (2001). Constructive criticisms of methodological and editorial practices. *Journal of Experimental Education*, *70*, 65-79.

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

Krantz, D. H. (1999). The null hypothesis testing controversy in psychology. *Journal of the American Statistical Association, 94*, 1372-1381.

Lesser, G. S. (1959). Population difference in construct validity. *Journal of Consulting Psychology, 23*, 60-65.

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

Levin, J. R. (1993). Statistical significance testing from three perspectives. *Journal of Experimental Education*, *61*, 378-382.

Levin, J. R. (1995, April). *The consultant's manual of researchers' common stat-illogical disorders.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Levin, J. R. (1998a). To test or not to test H_0? *Educational and Psychological Measurement, 58*, 313-333.

Levin, J. R. (1998b). What if there were no more bickering about statistical significance tests? *Research in the Schools, 5*, 43-53.

Levin, J. R., & Robinson, D. H. (1999). Further reflections on hypothesis testing and editorial policy for primary research journals. *Educational Psychological Review*, *11*, 143-155.

Levin, J. R., & Robinson, D. H. (2000). Statistical hypothesis testing, effect-size estimation, and the conclusion coherence of primary research studies. *Educational Researcher, 29*(1), 34-36.

Loftus, G. R. (1996). Psychology will be a much better science when we change the way we analyze data. *Current Directions in Psychology*, 5, 161-171.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, Massachusetts: Addison-Wesley.

Maxwell, S. E., & Delaney, H. D. (1990). *Designing experiments and analyzing data: A model comparison perspective*. Belmont, CA: Wadsworth.

McGraw, K. O., & Wong, S. P. (1992). A common language effect size statistic. *Psychological Bulletin*, *111*, 361-365.

McLean, J. E. (1995). *Improving education through action research: A guide for administrators and teachers*. Thousand Oaks, CA: Corwin Press.

McLean, J. E., O'Neal, M. R., & Barnette, J. J. (November, 2000). *Are all effect sizes created equal?* Paper presented at the annual meeting of the Mid-South Educational Research Association, Bowling Green, KY.

Meehl, P. E. (1967). Theory testing in psychology and physics: A methodological paradox. *Philosophy of Science*, *34*, 103-115.

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.

Neyman, J., & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrika, 29A*, Part I: 175-240; part II 263-294.

Nix, T. W., & Barnette, J. J. (1998). The data analysis dilemma: Ban or abandon. A review of null hypothesis significance testing. *Research in the schools, 5*, 3-14.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

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

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

Onwuegbuzie, A. J. (2001). *Towards a framework for comprehensive reporting of empirical findings: The role of statistical significance, theoretical significance, practical significance, and clinical significance*. Unpublished manuscript, Howard University, Washington, DC.

Onwuegbuzie, A. J. (2003). Expanding the framework of internal and external validity in quantitative research. *Research in the Schools, 10*, 71-90.

Onwuegbuzie, A. J., & Daniel, L. G. (2000, November). *Reliability generalization: The importance of considering sample specificity, confidence intervals, and subgroup differences*. Paper presented at the annual meeting of the Mid-South Educational Research Association, Bowling Green, KY.

Onwuegbuzie, A. J., & Daniel, L. G. (2001, April). *Indices of score reliability and their applications.* Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.

Onwuegbuzie, A. J., & Daniel, L. G. (2002a). A framework for reporting and interpreting internal consistency reliability estimates. *Measurement and Evaluation in Counseling and Development, 35*, 89-103.

Onwuegbuzie, A. J., & Daniel, L. G. (2002b). Uses and misuses of the correlation coefficient. *Research in the Schools, 9*, 73-90.

Onwuegbuzie, A.J., & Daniel, L.G. (2003, February 12). Typology of analytical and interpretational errors in quantitative and qualitative educational research. *Current Issues in Education* [On-line], 6(2). Available at http://cie.ed.asu.edu/volume6/number2/

Onwuegbuzie, A. J., & Daniel, L. G. (in press). Reliability generalization: The importance of considering sample specificity, confidence intervals, and subgroup differences. *Research in the Schools*.

Onwuegbuzie, A. J., Daniel, L. G., & Roberts, J. K. (in press). A proposed new "what if" reliability analysis for assessing the statistical significance of bivariate relationships. *Measurement and Evaluation in Counseling and Development*

Onwuegbuzie, A. J., & Weems, G. H. (in press). Characteristics of item respondents who frequently utilize midpoint response categories on rating scales. *Research in the Schools*.

Pedhazur, E. J. (1982). *Multiple regression in behavioral research: Explanation and prediction* (2nd ed.). New York: Holt, Rinehart and Winston.

Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measurement, design, and analysis: An integrated approach*. Hillsdale, NJ: Erlbaum.

Peet, M. W. (1999, November). *The importance of variance in statistical analysis: Don't throw the baby out of the bathwater*. Paper presented at the annual meeting of the Mid-South Educational Research Association, Point Clear, Alabama.

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

Prosser, B. (1990, January). *Beware the dangers of discarding variance.* Paper presented at the annual meeting of the Southwest Educational Research Association, Austin, TX. (ERIC Reproduction Service No. ED 314 496)

Roberts, J. K., & Onwuegbuzie, A. J. (2003). Alternative approaches for interpreting alpha with homogeneous subsamples. *Research in the School, 10*, 63-69.

Roberts, J. K., Onwuegbuzie, A. J., & Eby, R. (2001, April). *Alternative approaches for interpreting alpha with homogeneous subsamples: The introduction of a new measure of homogeneous alpha.* Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.

Robinson, D. H., Katayama, A. D., Dubois, N. F., & Devaney, T. (1998). Interactive effects of graphic organizers and delayed review on concept acquisition. *Journal of Experimental Education*, *67*, 17-31.

Robinson, D. H., & Levin, J. R. (1997). Reflections on statistical and substantive significance, with a slice of replication. *Educational Researcher*, *26*(5), 21-26.

Rosenthal, R., & DiMatteo, M. R. (2001). Meta-analysis: Recent developments in quantitative methods for literature reviews. *Annual Review of Psychology*, *52*, 59-82.

Rosenthal, R., & Rubin, D. B. (1982). A simple, general purpose display of magnitude of experimental effect. *Journal of Educational Psychology*, *74*, 166-169.

Rozeboom, W. W. (1960). The fallacy of the null hypothesis significance test. *Psychological Bulletin*, *57*, 416-428.

Rozeboom, W. W. (1997). Good science is abductive, not hypothetic-deductive. In L.L. Harlow, S.A. Mulaik, & J.H. Steiger (Eds.), *What if there were no significance tests?* Mahwah, NJ: Erlbaum, p. 335-392.

Sawilowsky, S. S., & Yoon, J. (2002). The trouble with trivials ($p > .05$). *Journal of Modern Applied Statistical Methods*, *1*, 143-144.

Schmidt, F. L. (1992). What do data really mean? *American Psychologist, 47*, 1173-1181.

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?* Mahwah, NJ: Erlbaum, 37-64.

Seaman, M. A., & Serlin, R. C. (1998). Equivalence confidence intervals for two-group comparisons of means. *Psychological Methods, 3*, 403-411.

Sechrest, L., & Yeaton, W. H. (1982). Magnitudes of experimental effects in social science research. *Evaluation Review*, *6*, 579-600.

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

Serlin, R. C. (1993). Confidence intervals and the scientific method: A case for Holm on the range. *Journal of Experimental Education*, *61*(4), 350-360.

Shadish, W. R., Cook, T. D., & Campbell, D.T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Shaver, J. P. (1993). What statistical significance testing is, and what it is not. *Journal of Experimental Education*, *61*, *293-316.*

Steering Committee of the Physicians' Health Study Research Group (1988). Findings from the aspirin component of the ongoing Physicians' Health Study. *New England Journal of Medicine, 318*, 162-264.

Thompson, B. (1986). ANOVA versus regression analysis of ATI designs: An empirical investigation. *Educational and Psychological Measurement, 46,* 917-928.

Thompson, B. (1988). Discard variance: A cardinal sin in research. *Measurement and Evaluation in Counseling and Development, 21*, 3-4.

Thompson, B. (1992a, April). *Interpreting regression results: Beta weights and structure coefficients are both important.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Thompson, B. (1992b). Two and one-half decades of leadership in measurement and evaluation. *Journal of Counseling and Development, 70*, 434-438.

Thompson, B. (1993). The use of statistical significance research: Bootstrap and other alternatives. *The 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. (2001). Significance, effect sizes, stepwise methods, and other issues: Strong arguments move the field. *Journal of Experimental Education, 70*, 80-93.

Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher, 31*(3), 25-32.

Thompson, B., & Vacha-Haase, T. (2000). Psychometrics is datametrics: The test is not reliable. *Educational and Psychological Measurement, 60*, 174-195.

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

Vacha-Haase, T., Kogan, L. R., & Thompson, B. (2000). Sample compositions and variabilities in published studies versus those in test manuals: Validity of score reliability inductions. *Educational and Psychological Measurement, 60*, 509-522.

Vargha, A., Rudas, T., Delaney, H. D., & Maxwell, S. E. (1996). Dichotomization, partial correlation, and conditional independence. *Journal of Educational and Behavioral Statistics, 21*, 264-282.

Weems, G. H., & Onwuegbuzie, A. J. (2001). The impact of midpoint responses and reverse coding on survey data. *Measurement and Evaluation in Counseling and Development, 34*, 166-176.

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

Yuen, K. K. (1974). The two-sample trimmed *t* for unequal population variances. *Biometrika*, 61, 165-170.