Methodological Issues

Editors: David Shannon
Auburn University

Isadore Newman
University of Akron

We are pleased to present this section regarding methodological issues. The purpose of this section is to provide an opportunity to learn about and generate discussion about a variety of methodological issues as they pertain to educational researchers. In this section, we feature two articles.

The first article, written by Frank Baugh and Bruce Thompson of Texas A & M University, discusses various types of effect sizes reported in social science research. The importance of effect size is clear as the number of journals requiring some measure of effect size continues to grow and the 5th edition of the APA Publication Manual emphasizes that it is "almost always necessary." There are many different measures of effect size that are used to demonstrate the magnitude of research effects. Baugh and Thompson summarize the different types of effect size measures and make critical distinctions among them as they pertain to social science research. Finally, they offer guidance for making the most appropriate selection of an effect size measure and reporting such measures.

The second article, written by Anthony Guarino, David Shannon, and Margaret Ross of Auburn University, provides a brief overview of the many different measures of fit used in structual equatioin modeling (SEM). The use of SEM, as well as measures used to assess corresponding models, has grown tremendously in educational research. The review by Guarino, Shannon, and Ross offers a simple framework for organizing measures of fit and guidance for the selection and reporting of such fit indices.

We hope these articles have been valuable and are looking forward to expanding this section in future issues. To do so, we need your input. What types of issues would you like to explore? We look forward to discussing these issues with you and exploring them in future issues. If you would like to submit a manuscript for this section, or simply discuss an idea, please contact: David Shannon, Auburn University, 4036 Haley Center-EFLT, Auburn, AL 36849-5221; telephone (334) 844-3071; e-mail shanndm@auburn.edu or Isadore Newman, University of Akron, Zook Hall-424, Akron OH 44325-4208; telephone (330) 972-6955; e-mail isadore@uakron.edu.

Using Effect Sizes in Social Science Research: New APA and Journal Mandates for Improved Practices

Frank Baugh
Texas A&M University

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

The 2001 edition of the APA Publication Manual emphasizes that effect sizes are "almost always necessary" in reporting and interpreting research results, and 17 journals already require such reports. The present report explains what effect sizes are, some of the dozens of effect indices and choices, and some issues as regards interpreting effect sizes.

Debates about statistical methods are commonplace in the social science literature (cf. Knapp & Sawilowsky, in press; Thompson, in press-a). The great controversy regarding statistical significance testing (null hypothesis testing: NHST) has overshadowed many other discussions since these procedures were first introduced many decades ago (Pedhazur & Schmelkin, 1991). A continuum of views regarding statistical significance testing is evident in the literature.

At one end are methodologists calling for an outright ban of the procedures (Hunter, 1997), who argue that "a massive educational effort is required to extinguish the mindless use of procedures that die hard" (Falk & Greenbaum, 1995, p. 94). Similarly, Rozeboom (1997) described statistical significance tests as "the most bone-headed misguided procedure ever institutionalized in the rote training of science students" (p. 335). Schmidt and Hunter (1997) argued that "Statistical significance testing retards the growth of scientific knowledge; it never makes a positive contribution" (p. 37).

At the other end of the continuum some scholars seemingly argue that all (or much) is well in the world of the

statistical significance test. Noteworthy exemplars include Cortina and Dunlap (1997), Frick (1996), and especially Abelson (1997).

Taking a more moderate stance are others (e.g., Cohen, 1990, 1994; Harris, 1991; Huberty, 1987; Kupersmid, 1988; Rosnow & Rosenthal, 1989; Thompson, 1993, 1996) who have highlighted the myriad limitations of statistical significance testing and called for supplemental analyses. These authors cite the deleterious deficiencies of the tests in evaluating result importance (Thompson, 1993), result replicability (Cohen, 1994), and measuring the size of an effect (Thompson, 1999b).

Effect Size Reporting

Chief among the supplements advocated has been the reporting of effect size. Unfortunately, to date admonitions to report effect sizes have had only limited effects (cf. Keselman et al., 1998; Thompson & Snyder, 1998). Even the American Psychological Association (APA, 1994) *Publication Manual*'s "encouragement" (p. 18) to present effect sizes proved ineffectual (see Keselman et al., 1998; Kirk, 1996; Snyder & Thompson, 1998; Thompson & Snyder, 1997, 1998; Vacha-Haase & Nilsson, 1998).

Some journal editors, recognizing the weakness of such an encouragement, have responsibly delineated journal-specific policies that "are considerably more enlightened" (Thompson, 1999c, p. 138). For example, author guidelines for *Educational and Psychological Measurement* have long stated, "Authors reporting statistical significance will be required to both report and interpret effect sizes" (Thompson, 1994b, p. 845, emphasis in original). Likewise, the editorial policies for APA's Journal of Applied Psychology declare,

> If an author decides not to present an effect size estimate along with the outcome of a significance test, I will ask the author to provide specific justification for why effect sizes are not reported. So far, I have not heard a good argument against presenting effect sizes. Therefore, unless there is a real impediment to doing so, you should routinely include effect size information in the papers you submit. (Murphy, 1997, p. 4)

The following 17 journals now require the reporting of effect sizes:

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 & Clinical Psychology
Journal of Counseling and 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.

In addition, the appointment and subsequent report (Wilkinson & APA Task Force on Statistical Inference, 1999) of the APA Task Force on Statistical Inference (TFSI) provides evidence of APA's recognition of the necessity of effect size reporting. Specifically, the Task Force (Wilkinson & TFSI, 1999) suggested that researchers "*Always* provide some effect-size estimate when reporting a p-value" (p. 599, emphasis added), and emphasized that "reporting and interpreting effect sizes in the context of previously reported effects is essential to good research" (p. 599, emphasis added).

Now the fifth edition of the APA (2001) *Publication Manual* has been released. The new edition goes considerably beyond the previous edition's "encouragement" (p. 18) to report effect sizes. The new manual emphasizes:

> For the reader to fully understand the importance of your findings, it is almost always necessary to include some index of effect size or strength of relationship in your Results section. You can estimate the magnitude of effect or the strength of the relationship with a number of common effect size estimates... 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, emphasis added)

Purpose of the Present Article

Clearly, the importance and value of effect size reporting has slowly permeated the field (Henson & Smith, 2000). As the shift in focus from p values to effect sizes continues, it is imperative that graduate students and researchers fully understand the concepts underlying effect size. Unfortunately, as Cohen (1988) indicated, "it [effect sizes] is the least familiar of the concepts surrounding statistical inference among practicing behavioral scientists" (p. 10). Therefore, the purposes of this paper are to present an accessible explication of effect sizes, provide a framework outlining the possible choices in effect size reporting, and describe appropriate reporting and interpretation procedures.

What is an Effect Size?

Cohen (1969) is credited with popularizing the term effect size in his seminal book on power analysis, though discussions by multiple authors including Pearson (1901) and Fisher (1925) of such estimates appeared in the literature much earlier (Kirk, 1996). Cohen described effect sizes as the "degree to which a certain phenomenon exists in a

population" (p. 9).

Cohen (1969) went on to describe the estimates in terms of the degree of departure of the actual sample results from those specified in the familiar null hypothesis. He noted that "It is convenient to use the phrase 'effect size' to mean the degree to which the null hypothesis [of no difference] is false.... When the null hypothesis is false [which it always is], it is false to some specific degree, i.e., *the effect size (ES) is some specific nonzero value in the population*" (pp. 9-10, emphasis in original). Another accessible definition of effect sizes was put forth by Snyder and Lawson (1993) who stated, "A magnitude-of-effect [i.e., effect size] statistic tells us how much of the dependent variable can be controlled, predicted, or explained by the independent variable(s)" (p. 335).

For example, if a researcher states a null hypothesis that the presence of some phenomenon in a population occurs at the rate of 50%, and the sample results for some identified subset of the population detects the phenomenon at 55%, then the effect size is 5%. Correspondingly, if an experimental group mean and control group mean are compared to determine the effect of a counseling intervention on depression as measured by the Beck Depression Inventory (BDI), the null hypothesis of no difference would occur in the sample when the experimental and control groups means on the posttest BDI were exactly equal. A difference in the means of 2 on the inventory is interpretable as an effect size estimate of 2.

Of course, as we shall see, there are literally dozens of ways to characterize the degree to which sample results diverge from the null hypothesis (i.e., effect sizes). No one choice fits every research situation.

Why are Effect Sizes Important?

Two separate but related issues characterize the importance of effect sizes. First, the differences between "practical" (cf. Kirk, 1996) and "clinical" (cf. Thompson, in press-b) versus "statistical" significance are noteworthy. Statistical significance is solely concerned with the probability that a sample result is due to sampling error or chance, given the sample size and a presumption that the null (usually "nil" null of zero effect) hypothesis is exactly true in the population (Cohen, 1994; Thompson, 1996). Conversely, practical significance focuses on the usefulness of a result in real life circumstances (Kirk, 1996). Clinical significance (Kendall, 1999) involves the degree to which an intervention may make targeted participants comparable to non-clinical samples (e.g., formerly depressed patients become comparable to non-depressed samples as regards depression).

Many researchers (cf. Mittag & Thompson, 2000) erroneously believe that statistical significance testing speaks to the practicality or importance of results. In this line of reasoning a statistically significant or unlikely result is deemed inherently interesting. Nothing could be further from the truth (Thompson, 1996)! Various authors (Carver, 1978; Shaver, 1985; Snyder & Thompson, 1998; Thompson, 1993, 1996) have repeatedly exposed this fallacy.

Thompson and Kieffer (2000) illustrated that large or noteworthy effects can be statistically nonsignificant while minute or uninteresting results can be statistically significant depending solely on sample size. Because *p* values tell us nothing about result importance or replicability, Snyder and Lawson (1993) noted that "...use of ME [magnitude of effect or effect size] indices can assist the researcher in clarifying whether statistically significant findings are of practical or meaningful significance within the context of an empirical investigation" (p. 335).

Second, effect sizes and their interpretation require us as scientists to think (a good thing). Many would like to believe scientific research is purely an objective endeavor and that subjective interpretations have no place within the enterprise. But "to assume science is entirely objective is to suffer from psychological denial" (Henson & Smith, 2000, p. 292). Thompson (1993) makes this point clear:

> Statistics can be used to evaluate the probability of an event. But importance is a question of human values, and math cannot be employed as an atavistic escape (à la Fromme's *Escape From Freedom*) from the existential human responsibility for making value judgments. If the computer package did not ask you your values prior to its analysis, it could not have considered your value system in calculating ps, and so ps cannot be blithely used to infer the value of research results. Like it or not, empirical science is inescapably a subjective business. (p. 365)

Empirical research is fraught with decisions that must be made by fallible researchers. Denial regarding such decisions does not allay the responsibility to proceed in a careful manner while exercising thoughtful, informed judgment (Henson & Smith, 2000; Kirk, 1996, 2001; Vacha-Haase & Nilsson, 1998). Some have pointed out potential biases that exist in the interpretation of effect sizes (Levin, 1998; Robinson & Levin, 1997). Nevertheless, effect size interpretation based on informed judgment has considerable potential to contribute to the cumulation of knowledge (cf. Cohen, 1990; Kirk, 2001; Thompson, 1999b).

Effect Size Choices

The term "effect size" can be used as a generic descriptor referring to the entire domain of effect size indices. Some confusion about measures of effect size (cf. Snyder & Lawson, 1993) may be the result of authors using differing terms to characterize the large domain of possible estimates. Examples of such terms include: magnitude of effect, magnitude of the experimental effect, explained variance, strength of association, strength of relationship, variance-accounted-for, effect size, or effect magnitude (Murray & Dosser, 1987; O'Grady, 1982; Snyder & Lawson, 1993).

It is important to note that some of these terms also delineate specific classes of effect size. Unfortunately, the practice of using multiple terms to describe a statistic or class of statistics, common in all areas of statistics, could confuse the most dedicated graduate student or scholar.

There are numerous effect size type estimates and no single best choice (Thompson, 1999a). Most measures of effect size fall into one of two major classes: (a) variance-accounted-for (also known as strength of association) and (b) standardized differences (standardized differences of means). In addition, Kirk (1996) describes a third category of "other measures" which are used less frequently and are not discussed here.

Effect sizes may also be used (a) "uncorrected" or, alternatively, the estimated positive bias in the estimates, due to capitalization on sample-specific sampling error variance, may be removed. Estimates in the latter category are (b) "corrected" to more accurately estimate either true population effects or the effects likely to occur in independent future samples.

Variance-accounted-for

Variance-accounted-for estimates characterize the proportion (or percentage) of variance in the dependent variable(s) explained by independent variable(s) (Olejnik & Algina, 2000; Snyder & Lawson, 1993). These effects are in a squared metric.

Because all General Linear Model (GLM) analyses (ANOVA, regression, etc.) are correlational (Knapp, 1978; Thompson, 1991, 2000), r^2 type effect sizes can be computed for all conventional studies (Thompson, 1999a). Eta-squared (η^2), also called the correlation ratio, is one such effect size in ANOVA. It is computed by dividing the sum of squares (*SS*) explained (between or model) by the *SS*. Likewise, the multiple correlation coefficient, R^2, is an effect size estimate in regression (analogous to η^2) and is calculated by dividing the $SS_{EXPLAINED}$ by SS_{TOTAL}. Both the η^2 and R^2 formulas yield a number between 0 and 1, inclusive, which can then be multiplied by 100 to derive a percent (Kier, 1999). For example, Table 1 reports results of a one-way three-level ANOVA which yielded an effect size of .1315 or 13.15% ($SS_{EXPLAINED}$ / SS_{TOTAL} = 577.4 / 4390.3 = .1315].

Whereas this method holds for univariate analyses, the presence of more than one dependent variable in multivariate analyses requires a different procedure. Thompson (1999a) indicates that 1 – lambda (Λ) in the multivariate situation produces one estimate of multivariate eta^2.

Standardized Differences

Cohen (1988) indicated that the unit of measurement for an effect size must be chosen according to its appropriateness for the data and the statistical model. Discussing effect sizes in terms of the original metric of the dependent variable may be appealing for some well-known instruments (e.g., tests of intelligence). On the other hand, a standardized effect size index that scales measurement across studies into a common metric may be appealing, particularly when different measures yielding different metrics are used in different studies (Cohen, 1988).

Variance-accounted-for effect sizes are one such type of "metric free" (or scale-free) indices. Standardized differences are also scale-free, because score scaling (or spreadoutness) is removed from the effect index by division by some *SD*. However, standardized differences do not discard signs of effects, because they are not in a squared metric. This can be desirable. For example, although a new teaching method may have positive effects in most experiments, the teaching method may have negative treatment effects in some studies, and so attention to the directions of effects across studies can be very important (Kier, 1999).

Standardized difference effect sizes remove the influence of some estimate of the population standard deviation (i.e., the parameter σ) by dividing the mean differences (i.e., $\overline{X}_1 - \overline{X}_2$) by this estimated parameter. Of course, there are dozens of potential estimates of σ, which is indicative of why there are literally dozens of effect indices, no one of which is correct for every research situation.

Table 1
ANOVA Summary Table

Source	*SS*	*df*	*MS*	*F*calc	Effect Size
Explained	577.4	2	288.7	2.04	13.15%
Error	3812.9	27	141.2		
Total	4390.3	29	151.4		

Note. The effect size computed here is eta^2 or the correlation ratio (not the correlation coefficient *r* which, unlike eta^2, is in an unsquared metric).

For example, Cohen's (1969) *d* divides the difference of group means by the "pooled" standard deviation across the groups in a study. In other words, a weighted average of each group's *SD* on the dependent variable is computed. This has the appeal of using a larger n in estimating the parameter, thereby hopefully more accurately estimating the parameter.

The computation for Hedges' (1981) *g* is also based on the "pooled" variance of groups. However, here the "pooled" variance invokes a denominator of *n*-1 rather than *n*. Obviously, if *n* is large, the difference between these two estimates will become quite small.

Glass (1976) believed the control group standard deviation to be the best estimate of the population σ if, for example, an experimental intervention might impact not only the group means, but also the standard deviations of the groups. In such cases Glass' delta (Δ) computes the standardization by dividing mean differences by the *SD* of *only* the control group.

"Corrected" Variance-accounted-for Effect Sizes

Classical GLM analyses all either explicitly or implicitly use least squares weights derived from measured/observed variables (e.g., *Y*) in the sample data to compute scores on latent/synthetic variables (e.g., regression *Y*^). This process capitalizes on *all* variance present in an analysis, including the "sampling error variance" (error associated with the sampling procedure and unique to the specific sample). As a result, "uncorrected" variance-accounted-for effect sizes are positively biased. In other words, the effect size would be overestimated if the same weights were applied to either the population or a future sample (Snyder & Lawson, 1993; Thompson, 1993, 1999a).

Sample size, the number of measured variables, and the size of the effect in the population are all associated with more error variance being present in a sample and thus impact the amount of positive bias. (Smaller samples, more measured variables, and a small population effect size magnify "sampling error variance".) The amount of overestimation is, therefore, a function of the interaction of these three design features (Thompson, 1990).

Various statisticians have developed "corrected" effect size statistics that adjust downward the positive bias present in "uncorrected" estimates. Kier (1999) indicated that many of the "corrected" effect statistics are based on the formulas developed by Ezekiel (1930) but often attributed to Wherry (1931). The Ezekiel formula, Hays' (1963) omega^2 (ω^2), and Kelley's (1935) epsilon^2 (ε^2) all adjust for the positive bias in estimating the *population* effect.

Other "corrected" variance-accounted-for formulas, such as Herzberg's (1969) and Lord's (1950), are used to estimate the effect size in future samples rather than the population (Kier, 1999; Thompson, 1993). "Uncorrected" estimates are always more than or equal to the corresponding "corrected" estimate. Moreover, population "corrected" effect estimates are larger than "corrected" estimates for future samples. This is true because the latter case requires a downward adjustment for the unique sampling error variance that occurs in both the present and the future samples (Kier, 1999; Snyder & Lawson, 1993).

Decisions about which type of "corrected" variance-accounted-for effect estimate to employ are contingent on the generalizations researchers wish to make. For example, Thompson (1993) remarked,

> From one perspective it might be argued (and has been by some—see Stevens, 1992) that estimates in the last class [future sample estimates] are the most relevant, because in practice scientists extrapolate expectations from previous studies with samples and hope their results will be replicated in future studies with samples. (p. 366)

Both eta^2 and R^2 are popular "uncorrected" effect size estimates. "Adjusted R^2" is often embraced as a "corrected" estimate, perhaps because most statistical packages automatically always report this statistic in regression analyses (Kier, 1999).

The numerical difference between "corrected" and "uncorrected" effect sizes is known as "shrinkage". Even though the variance-accounted-for effect size estimates are in squared metric, "corrected" variance-accounted-for effect sizes can be negative (cf. Thompson, 1999a). Large amounts of "shrinkage", especially to the point of negative "corrected" variance-accounted-for estimates, suggest the (abysmal) absence of reliable results and a lack of statistical power (cf. Cohen, 1988) in the research design. For example, Thompson (1994a) discussed an unfortunate doctoral dissertation in which an R^2 of 44.6% reduced to an adjusted R^2 = 0.45%. Of course, such dramatic "shrinkage" is alarming for any empirical investigation.

"Corrected" Standardized Differences

Both variance-accounted-for and standardized difference effect size statistics have strengths deeming them reasonably useful. For example, variance-accounted-for estimates have the attractive quality of being easily computed in all parametric studies regardless of the analysis employed, even in designs involving a single group (i.e., no subgroups).

But the effect statistics in these two classes can be readily expressed in terms of each other. Cohen (1988, p. 23) provided the following formula for deriving r (e.g., the square root of eta^2 or R^2) from *d* when the groups of interest are fairly equivalent in size:

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

Friedman's (1968, p. 246) formula expresses *d* in terms of *r*:

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

The realizations that (a) *d* can be derived from the square root of a variance-accounted-for effect size (*r*) and (b) "corrected" variance-accounted-for effect sizes can be quite useful led Thompson (in press-b) to suggest that (c) "corrected" standardized differences might be useful. However, his suggestions have not yet been fully explored.

Abridged Summary of Choices

Elmore (2001) recently counted more than 60 different effect sizes. Given the wide array of choices, all formulae and applications cannot be presented here. Kirk (1996) and Snyder and Lawson (1993) and others do provide helpful reviews. Table 2 presents an abridged summary to help communicate the flavor of the diversity of choices.

Interpreting Effect Sizes

Estimates of effect size have the potential to greatly enhance the business of science when researchers understand and interpret them properly (Maxwell, Camp, & Avery, 1981). However, reporting an effect size without interpreting its meaning is almost as common (see Snyder & Thompson, 1998; Thompson & Snyder, 1997) and just as reproachable as not reporting the effect at all. Various dynamics may explain this behavior, including what Thompson (1999c) described as the "atavistic" desire of researchers to avoid responsibility for exercising judgment (see also Kirk, 1996).

As with *p* values no artificial guidelines can reasonably be relied upon to dictate what is an "important" effect size. Cohen's (1990) recommendation is worth remembering: "...don't look for a magic alternative to NHST [null hypothesis significance testing], some other objective mechanical ritual to replace it. It doesn't exist" (p. 1001).

After thinking about a broadly defined social science literature, Cohen (1969, 1988) noted typical effect size values that he defined as "small", "medium", and "large". According to his classification of effect typicality in this literature, a d of .2 or the corresponding r^2 of 1.0% can be understood as being relatively small. The d of .5 or $r^2 = 5.9\%$ falls in the "medium" size range. Large effects are equal to or greater than $d = .8$ or $r^2 = 13.8\%$ (Thompson, 1999a).

However, Cohen did not intend these values to be used as rigid cutoffs. As Thompson (1999a) warned, "If in evaluating effect size we apply Cohen's conventions (against his wishes) with the same rigidity with which we have traditionally applied the " = .05 statistical significance testing convention we will merely be being stupid in a new metric" (p. 72). Howell (2002, pp. 228-229) elaborates on these and related issues.

Effect sizes can *only* be interpreted relative to the specific research context and in light of the researcher's personal values (Kirk, 1996). Two scientists could differentially interpret the r^2 of 14%, even when carrying out the same investigation! One may deem the 14% effect size highly noteworthy, while the other considers it trivial. Additionally, even some seemingly "small" effect estimates (e.g., 2%) may be remarkably important when critical outcomes such as human life-or-death issues are involved (see Gage, 1978).

Confidence Intervals for Effects Sizes

Researchers increasingly realize that viewed separately single empirical investigations yield limited useful knowledge (Schmidt, 1996). Yet, researchers have the potential to gain insight as the number of studies focused on a given phenomenon increases. Calculating and reporting confidence intervals is one method for researchers to contribute to the accumulation of knowledge (Thompson, 1999a). For example, Schmidt (1996) suggested that if we interpret the confidence intervals in our study in the context of the intervals in all related previous studies, the true population parameters will eventually be isolated across studies, even if our prior expectations regarding the parameters are wildly wrong.

Indeed, the APA *Publication Manual* published in July, 2001 emphasized that:

> The reporting of confidence intervals... can be an extremely effective way of reporting results. Because confidence intervals combine information on location and precision and can often be directly used to infer significance levels, they are, in general, the *best* reporting strategy. The use of confidence intervals is therefore *strongly recommended*. (p. 22, emphasis added)

It is important to strongly emphasize that the best use of confidence intervals is in relation to intervals from prior studies, and not in relation to whether they subsume zero! The latter practice is merely NHST in sheeps' clothing (Thompson, 1998).

The Task Force (Wilkinson & APA Task Force on Statistical Inference, 1999) recommendations to report (1) effect sizes and (2) confidence intervals quite naturally lead to a suggestion to marry these practices and report (3) confidence intervals about effect sizes. Unfortunately, the estimation of confidence intervals about effect sizes raises thorny technical issues, because such estimates require (a) the use of special statistical distributions that are called "noncentral" (e.g., "noncentral t", "noncentral F"), with which many researchers may be unfamiliar, and (b) the use of computer-intensive estimation procedures, because iterative estimation must be used rather than a computation formula.

Fortunately, new software and/or new programming for old software have overcome these two difficulties. The August, 2001 issue of *Educational and Psychological Measurement* was a special issue focusing on the computation of confidence intervals about effect sizes. User-friendly, accessible explanations of confidence intervals in general, and CIs about effect sizes in particular, were presented. The articles also provide and/or describe software that can be used to derive these estimates (cf. Cumming & Finch, 2001; Fidler & Thompson, 2001; Smithson, 2001).

Conclusion

In 1994, the fourth edition of the APA *Publication Manual* "encouraged" (p. 18) effect size reporting. But 11 studies of reporting practices in one or two post-1994 volumes of 23 different journals found that this "encouragement"

Table 2
Some Effect Size Choices

Effect Size	Common Application	Class	Corrected?	Formula
eta-squared[2] (η^2), also called "correlation ratio"	*ANOVA*	variance-accounted-for	No	$SS_{EXPLAINED}/SS_{TOTAL}$
R^2	regression	variance-accounted-for	No	$SS_{EXPLAINED}/SS_{TOTAL}$
"adjusted" R^2	regression	variance-accounted-for	Yes	$1 - [((n - 1)/(n$ - variables- $1))\ (R^2)]$ OR $R^2 - ((1 - R^2)\ (n$ variables $/\ (n - n$ variables $- 1)))$
omega[2] (ω^2)	*ANOVA*	variance-accounted-for	Yes	$(SS_{BETWEEN} - (k - 1) \times MS_{WITHIN}) / (SS_{TOTAL} + MS_{WITHIN})$, where k is the number of groups
Cohen's *d*	*t* or *ANOVA*	standardized difference	No	$(M_{EXPERIMENTAL} - M_{CONTROL}) / SD_{POOLED}$

was ineffective (Vacha-Haase, Nilsson, Reetz, Lance & Thompson, 2000). Following the wise counsel of various scholars (cf. Kirk, 2001; Hyde, 2001; Vacha-Haase, 2001), the new *Manual* (APA, 2001) goes considerably further in emphasizing the importance of effect size reporting. This is good news.

The business of science is concerned with the amassing of important information through empirical investigation. Social science researchers intending to make contributions toward this end must recognize the limits of statistical significance testing and the detrimental impact that blind reliance on these procedures has had in the field. Because resistance to change in reporting practices continues to permeate the field, some methodologists contend that journal editorial policies requiring effect size reporting and interpretation have the potential to bring about the necessary transition (cf. Vacha-Haase, 2001). Happily, more and more journal editors are responding favorably to this encouragement by requiring effect size estimates to be reported and interpreted (e.g., McLean & Kaufman, 2000; Murphy, 1997; Snyder, 2000; Thompson, 1994).

Tests of statistical significance are not capable of yielding information about the size or relative noteworthiness of effects. Statistical test results are of course influenced by effect sizes, but statistical significance is also influenced by various other study features, including most notably sample size (Thompson & Kieffer, 2000). Because *p* values are therefore confounded measures of study effects, *p* values are not themselves directly useful as measures of effect (Thompson, 1999b).

Some have argued that statistical significance tests should be banned (e.g., Hunter, 1997; Schmidt & Hunter, 1997), while others have demurred from this view (cf. Levin, 1998). But regardless of these differences, what the field now seems to agree is that effect sizes are "essential" to interpreting and communicating results (Wilkinson & APA Task Force on Statistical Inference, 1999). Consequently, researchers have a responsibility as ethical scientists to gain a basic understanding of the concepts underlying effect size and in every investigation to provide some estimate of effect.

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.

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

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

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

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45*, 1304-1312.

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 calculation and interpretation of both central and noncentral confidence intervals. *Educational and Psychological Measurement, 61*, 532-574.

Elmore, P. (2001, April). *A primer on basic effect size concepts*. Paper presented at the annual meeting of the American Educational Research Association, Seattle.

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

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

Fidler, F., & Thompson, B. (2001). Computing correct confidence intervals for ANOVA fixed- and random-effects effect sizes. *Educational and Psychological Measurement, 61*, 575-605.

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

Gage, N. L. (1978). *The scientific basis of the art of teaching*. New York: Teachers College Press.

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

Harris, M. J. (1991). Significance tests are not enough: The role of effect-size estimation in theory corroboration. *Theory & Psychology, 1*, 375-382.

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

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

Henson, R. K. & Smith, A. D. (2000). State of the art in statistical significance and effect size reporting: A review of the APA Task Force report and current trends. *Journal of Research and Development in Education, 33*, 285-296.

Herzberg, P. A. (1969). The parameters of cross-validation. *Psychometrika Monograph Supplement, 16*, 1-67.

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

Huberty, C. J (1987). On statistical testing. Educational Researcher, 16, 4-9.

Hunter, J. E. (1997). Needed: A ban on the significance test. *Psychological Science, 8*(1), 3-7.

Hyde, J.S. (2001). Reporting effect sizes: The roles of editors, textbook authors, and publication manuals. *Educational and Psychological Measurement, 61*, 225-228.

Kelley, T. L. (1935). An unbiased correlation ration 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.

Keselman, H. J., Huberty, C. J, Lix, L. M., Olejnik, S., Cribbie, R., Donahue, B., Kowalchuk, R. K., Lowman, L. L., Petoskey, M. D., Keselman, J. C., & Levin, J. R. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research, 68*, 350-386.

Kier, F. J. (1999). Effect size measures: What they are and how to compute them. In B. Thompson (Ed.), *Advances in social science methodology* (Vol. 5, pp. 87-100). Stamford, CT: JAI Press.

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

Kirk, R. E. (2001). Promoting good statistical practices: Some suggestions. *Educational and Psychological Measurement, 61*, 213-218.

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

Knapp, T., & Sawilowsky, S. (in press). Constructive criticisms of methodological and editorial practices. *Journal of Experimental Education, 69*.

Kupersmid, J. (1988). Improving what is published: A model in search of an editor. *American Psychologist, 43*, 635-642.

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

Lord, F. M. (1950). *Efficiency of prediction when a regression equation from one sample is used in a new sample* (Research Bulletin 50-110). Princeton, NJ: Educational Testing Service.

McLean, J.E., & Kaufman, A.S. (2000). Editorial: Statistical significance testing and other changes to Research in the Schools. *Research in the Schools, 7*(2), 1-2.

Maxwell, S. E., Camp, C. J., & Avery, R. D. (1981). Measures of strength of association: A comparative examination. *Journal of Applied Psychology, 66*, 525-534.

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), 14-20.

Murphy, K. R. (1997). Editorial. *Journal of Applied Psychology, 82*, 3-5.

Murray, L. W., & Dosser, D. A. (1987). How significant is a significant difference? Problems with the measurement of magnitude of effect. *Journal of Counseling Psychology, 34*, 68-72.

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

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

Pearson, K. (1901). On the correlation of characters not quantitatively measurable. *Philosophical Transactions of the Royal Society of London, 195*, 1-47.

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

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

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

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

Schmidt, F. (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.

Shaver, J. (1985). Chance and nonsense. *Phi Delta Kappan, 67*, 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.

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

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

Snyder, P., & Thompson, B. (1998). Use of tests of statistical significance and other analytic choices in a school psychology journal: Review of practices and suggested alternatives. *School Psychology Quarterly, 13*, 335-349.

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

Thompson, B. (1991). A primer on the logic and use of

canonical correlation analysis. *Measurement and Evaluation in Counseling and Development, 24*, 80-95.

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

Thompson, B. (1994a, April). *Common methodology mistakes in dissertations, revisited.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans. (ERIC Document Reproduction Service No. ED 368 771)

Thompson, B. (1994b). Guidelines for authors. *Educational and Psychological Measurement, 54*, 837-847.

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

Thompson, B. (1998). In praise of brilliance: Where that praise really belongs. *American Psychologist, 53*, 799-800.

Thompson, B. (1999a, April). *Common methodology mistakes in educational research, revisited, along with a primer on both effect sizes and the bootstrap.* Invited address presented at the annual meeting of the American Educational Research Association, Montreal. (ERIC Document Reproduction Service No. ED 429 110)

Thompson, B. (1999b). Statistical significance tests, effect size reporting, and the vain pursuit of pseudo-objectivity. *Theory & Psychology, 9*, 191-196.

Thompson, B. (1999c). Why "encouraging" effect size reporting is not working: Etiology of researcher resistance to changing practices. *Journal of Psychology, 133*, 133-140.

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

Thompson, B. (in press-a). Significance, effect sizes, stepwise methods, and other issues: Strong arguments move the field. *Journal of Experimental Education, 69*.

Thompson, B. (in press-b). "Statistical," "practical," and "clinical": How many kinds of significance do counselors need to consider? *Journal of Counseling and Development*.

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.

Thompson, B., & Snyder, P. A. (1997). Statistical significance testing practices in the Journal of Experimental Education. *Journal of Experimental Education, 66*, 75-83.

Thompson, B., & Snyder, P. A. (1998). Statistical significance and reliability analyses in recent JCD research articles. *Journal of Counseling and Development, 76*, 436-431.

Vacha-Haase, T. (2001). Statistical significance should not be considered one of life's guarantees: Effect sizes are needed. *Educational and Psychological Measurement, 61*, 219-224.

Vacha-Haase, T., & Nilsson, J. E. (1998). Statistical significance reporting: Current trends and uses in MECD. *Measurement and Evaluation in Counseling and Development, 31*, 46-57.

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.

Wherry, R. J. (1931). A new formula for predicting the shrinkage of the coefficient of multiple correlation. *Annals of Mathematical Statistics, 2*, 440-451.

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]

Author Note

Frank Baugh may be contacted through his e-mail address at "dbsorento@neo.tamu.edu". Bruce Thompson receives e-mail (and provides reprints) through his Web page URL: "http://www.coe.tamu.edu/~bthompson".

Frank Baugh is a doctoral student in Counseling Psychology. His training has focused on counseling and statistics.

Bruce Thompson is Professor and Distinguished Research Scholar at Texas A & M University and adjunct Professor of Community Medicine at Baylor College of Medicine (Houston). His primary areas of expertise are statistics, measurement, and program evaluation.