DOCUMENT RESUME

ED 464 140 TM 033 823

AUTHOR	Curtis, Deborah A.; Araki, Cheri J.
TITLE	Effect Size Statistics: An Analysis of Statistics Textbooks Used in Psychology and Education.
PUB DATE	2002-04-00
NOTE	42p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 1-5, 2002).
PUB TYPE	Information Analyses (070) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Educational Research; *Effect Size; Literature Reviews; Psychological Studies; *Statistics; Textbook Content; *Textbooks

ABSTRACT

The purpose of this research was to analyze recent statistics textbooks (n=22) to examine the ways in which authors addressed the issue of effect size (ES) and the practical significance of research results. In terms of the overall prevalence of ES statistics, the one-sample, matched-pair, and two-sample t-tests, and one-factor analysis of variance were presented in all 22 texts. ES statistics corresponding to each of these significance tests were presented in 4, 9, 16, and 18 texts, respectively. Problems with the way in which ES statistics were presented were identified, including the failure to distinguish between ES parameters and statistics, the use of conceptually uninformative cookbook formulas, and the lack of agreement on how to calculate specific ES statistics. Problems with the ways in which authors discussed the interpretation of the magnitude of the ES statistics were also identified, including the over-reliance on rules of thumb. (Contains 3 tables and 66 references.) (Author/SLD)

Effect Size Statistics: An Analysis of Statistics Textbooks used in Psychology and Education

Deborah A. Curtis

Cheri J. Araki

San Francisco State University

Paper presented at the April 2002 American Educational Research Association, New Orleans, LA.

Abstract

The purpose of this research was to analyze recent statistics textbooks (N=22) to examine the ways in which authors addressed the issue of effect size (ES) and the practical significance of research results. In terms of the overall prevalence of ES statistics, the one-sample, matched-pair, and two-sample *t*-tests, and one-factor ANOVA were presented in all 22 texts. ES statistics corresponding to each of these significance tests were presented in 4, 9, 16, and 18 texts, respectively. Problems with the way in which ES statistics were presented were identified, including the failure to distinguish between ES parameters and statistics, the use of conceptually uninformative cookbook formulas, and the lack of agreement on how to calculate specific ES statistics. Problems with the ways in which authors discussed the interpretation of the magnitude of the ES statistics were also identified, including the over-reliance on "rules of thumb".

Effect Size Statistics: An Analysis of Statistics Textbooks used in Psychology and Education

There has been a considerable amount of discussion in the statistical and methodological literature about the limitations of statistical significance testing (SST), and the value of reporting and interpreting effect size (ES) statistics (see, for example, Kirk, 1996; Snyder & Thompson, 1998; Thompson, 1994, 1996). But recent reviews of published studies in education and psychology journals indicate that most researchers do not report or interpret ES statistics in their research (Finch, Cumming, & Thomason, 2001; Keselman et al., 1998; Kirk, 1996; Snyder &Thompson, 1998; Thompson & Snyder, 1997, 1998; Vacha-Haase &Nilsson, 1998). A number of methodologists have speculated about why there is an over-reliance on SST and a lack of ES reporting in published research. Nickerson (2000), for example, discussed the over-reliance on SST in terms of some common statistical misconceptions, including the belief that a small *p*-value indicates a large effect. Kirk (2001), also commenting on the over-reliance of SST, reflected on his experiences as a statistics textbook writer, and discussed textbook publishers' pressure to "dumb down" statistics textbooks.

For better or worse, the content of the textbooks used in statistics classes will determine, to some extent, the degree to which future researchers learn about ES theory, computation, and interpretation. The focus of this research, therefore, is to analyze recent statistics textbooks regarding the ways in which the authors address the issue of ES and the practical significance of research results.

BACKGROUND

Assessing the Magnitude of Effect

Recent trends in ES statistic reporting. Statisticians have long-argued that *p*-values

should *not* be used as indicators of magnitude of effect (see Carver, 1993; Cohen, 1990, 1994; Shaver, 1993). Shaver noted:

> It is so commonly stressed that the statistical significance of results is directly a function of sample size that one can only wonder at the number of articles in which results are either interpreted as important because of statistical significance or in which the probability level appears to be taken as an indication of magnitude, as suggested by the use of terms such as *highly significant* when the probability is .01 or less. (p. 303)

Researchers in psychology and education are shifting away from interpreting *p*-values as indicators of effect size, and are beginning to understand the value of reporting and interpreting ES statistics. Two years ago, a report by the American Psychological Association Task Force on Statistical Inference (TFSI) was published (Wilkinson & TFSI, 1999). The TFSI members argued that researchers should "always provide some effect-size estimate when reporting a *p*-value" (p. 599). Moreover, they stated that "it helps to add brief comments that place these effect sizes in a practical and theoretical context" (p. 599). The most recent *Publication Manual of the American Psychological Association (APA)* (APA, 2001) has incorporated a number of recommendations made by the TFSI regarding ES statistics:

> 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)

There are now at least 13 education and psychology journals with editorial policies that require the use of ES statistics (Heldref Foundation, 1997; Thompson, 1994).

Computing ES statistics. There is a debate in the literature about when ES statistics should be calculated. Some methodologists, including the members of the TFSI mentioned above, have argued an ES statistic should be calculated for each SST reported, regardless of whether or not it is significant (see Carver, 1993; Thompson, 1996; Thompson, 1999; Wilkinson et al. 1999). As previously mentioned, this is also the stance advocated in the most recent addition of Publication Manual of the APA (APA, 2001). But others have argued that an ES statistic should *only* be calculated following a statistically significant result (see Huberty, 1987; Levin & Robinson, 1999; Robinson & Levin, 1997). The basic argument put forth here is that a two-step process should be used because ES statistics that correspond to nonsignificant SSTs are not trustworthy.

Interpreting ES statistics. It is insufficient to simply report ES statistics—the researcher needs to interpret these statistics as well (Henson & Smith, 2000; Keselman et. al., 1998; Thompson, 1996). In terms of ES interpretation, one common approach is to use rules of thumb. But as Shaver noted, this can be problematic:

> There already is a tendency to use criteria, such as J. 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)

The issue of how to interpret ES statistics is challenging, because interpretation is inherently subjective. As Henson and Smith (2000) argued, in high stakes research such as in certain medical studies, relatively small effects might be more valued when lives might be saved.

What is of "practical" significance in one study may not be the same as in another study.

Preparing Researchers in Education and Psychology

Little attention has been paid to studying what graduate students actually learn about ES theory, computation, and interpretation. In fact, there is an overall lack of research on what graduate students are learning in their statistics classes. Research has been conducted to evaluate graduate student training in "old standard" and advanced statistical procedures, both in education (Curtis & Harwell, 1998) and psychology (Aiken, West, Sechrest, & Reno, 1990). But neither of these studies collected information on graduate student training in ES theory, computation, or interpretation. What these studies do indicate is that there is great variation in how much time students spend studying statistics. In field of education, Curtis and Harwell surveyed 27 doctoral programs, and found that in only 13% of the universities were students in all programs required to take at least two statistics courses. In 44% of the universities, some students could graduate without taking a single statistics course. In the field of psychology, Aiken et al. found that 89% of the 186 departments they studied offered an introductory graduate statistics sequence, and of the departments offering this sequence, 77% were one year long.

Evaluating Statistics Textbooks in Education and Psychology

Hyde (2001) discussed the importance of including the topic of ES in introductory statistics textbooks:

> Authors of textbooks on statistical methods and research design—even those aimed at the average sophomore—should definitely include material on the theory underlying effect sizes, their computation, and their interpretation. Understanding the logic of effect sizes should be as basic to students' training as understanding the logic of underlying null hypothesis testing. The question then arises as to

> whether the average sophomore will be able to "get it," that is, will be able to understand the concepts. I can say with some confidence, based on 10 years of teaching effect sizes to undergraduates, that they can indeed get it. (p. 226)

There is little research on the quality of statistics textbooks used in psychology and (Harwell, et al. 1996). A number of studies have been conducted which evaluate statistics textbooks based on criteria developed by the studies' authors (see, for example, Brogan, 1980; Cobb, 1987, Harwell, et al., 1996; Huberty & Barton 1990; Huberty, 1993; Schact, 1990). But only one of these studies considered ES reporting: Huberty evaluated 19 statistics textbooks published from 1990 to 1992, and found that the authors of only 8 of these texts "even hinted" of ES indices for group comparison studies (p. 330). He further found that the authors of only 2 of these 8 textbooks presented an ES statistic corresponding to a two-group *t*-test. Although Huberty's findings are important, the textbooks he evaluated are now at least 10 years old, and the prevalence of ES coverage may have increased in recent years.

Specific Research Interests

This study had four parts. First, we examined the *prevalence* of ES coverage in current statistics texts. Second, we analyzed textbook authors' explanations of the *logic* underlying various ES indices. As part of this analysis, we examined the formulas that were presented to see if they were conceptually-meaningful *definitional* formulas or conceptually-uninformative *computational* formulas. Third, we compared textbooks to see if the formulas presented for the *same* ES statistic in different texts were *algebraically equivalent*. Finally, we analyzed authors' explanations of how to *interpret* the magnitude of an ES statistic. Here, we considered whether the authors recommended (a) using *p*-values to interpret the magnitude of effect; (b) using rules of thumb to interpret the magnitude of an ES statistic; (c) calculating ES statistics for all SST, or

only for statistically significant SST; and (d) interpreting the magnitude of an ES statistic based on previous research and on the nature of the research question.

A Brief Review of Several ES Statistics

In this section, we will present a brief review of several ES statistics—namely d, $\hat{f}$, r^2_{pb}, $\hat{\eta}^2$, *and* $\hat{\omega}^2$. We assume that the reader has a reasonable background in ES theory, computation, and interpretation. The information presented here is meant to provide the reader with a framework for interpreting the findings from this particular study.

The standardized difference between two means. The standardized difference between two means, d, is an appropriate ES statistic for all three t-tests considered here. (For the sake of clarity, we added our own subscripts for d). For the one-sample case, $d_1 = (\bar{x} - \mu) / s_x$, where $\bar{x}$ is the sample mean, μ is the hypothesized population mean, and s_x is the standard deviation of the sample. For the matched-pair case, $d_2 = \bar{x}_d / s_d$, where $\bar{x}_d$ and s_d are the mean and standard deviation of the difference scores, respectively. For the two-sample independent-groups case, we can calculate $d_3 = (\bar{x}_1 - \bar{x}_2) / s_p$ where $\bar{x}_1$ and $\bar{x}_2$ are the sample means of the first and second groups, respectively, and where $s_p = \sqrt{MSW}$ is the pooled standard deviation, or $d_4 = (\bar{x}_1 - \bar{x}_2) / s_c$, where s_c is the standard deviation of the control group.

Cohen's f. Cohen (1988) proposed an ES parameter, f, as an extension of the standardized difference between two means. Instead of computing the difference between two means, as in d, the spread among the means is represented by "a quantity formally like a standard deviation" (p. 275), and is divided by the common standard deviation of the populations. Thus, $f = \sigma_{means} / \sigma_{within}$, where σ_{means} is the standard deviation of population means, $[\sum(\mu_k - \mu)^2 / K]^{1/2}$, and σ_{within} is the within-population standard deviation. As Cohen indicated:

> f is thus also a pure number, the standard deviation of the standardized means. That is to say that if all the values of the combined populations were to be converted to **z** 'standard' scores (Hays, 1973, p. 250f), using the within-population standard deviation, **f** is the standard deviation of these **k** mean **z** scores. (p. 275).

Cohen (1988) proposed as a population parameter, f, for the purpose of conducting power analyses. Various methods for estimating f have been proposed by others, and these methods will be presented in a later section of this paper.

Measures of explained variance: r_{pb}^2, $\hat{\eta}^2$ *and* $\hat{\omega}^2$. The coefficient r_{pb}^2 is an alternative ES statistic to d for comparing two independent groups. For more complex designs, $\hat{\eta}^2$ and $\hat{\omega}^2$ can be used. (In the two-sample case, r_{pb}^2 and $\hat{\eta}^2$ are equal). Eta-squared is defined as the sums of squares for a given effect divided by the total sums of squares. Thus, in a two-factor fully crossed design, $\hat{\eta}_A^2 = SS_A / SS_{total}$, $\hat{\eta}_B^2 = SS_B / SS_{total}$, and $\hat{\eta}_{AxB}^2 = SS_{AxB} / SS_{total}$. In a repeated measures design, where the same group of participants is measured under three different conditions, $\hat{\eta}_{conditions}^2 = SS_{conditions} / SS_{total}$.

It is also possible to calculate *partial* $\hat{\eta}^2$ (Cohen, 1973). For example, in a two-factor design, $\hat{\eta}_{partial}^2$ for factor A can be calculated by removing the sums of squares for the main effect of B and the A x B interaction from the denominator: $\hat{\eta}_{A(partial)}^2 = SS_A / (SS_{total} - SS_B - SS_{AxB})$. In a repeated measures design, $\hat{\eta}_{partial}^2$ can be calculated by removing the sums of squares between the subjects from the denominator: $\hat{\eta}_{conditions(partial)}^2 = SS_{conditions} / (SS_{total} - SS_{subjects})$.

Although $\hat{\eta}^2$ is a reasonable statistic for describing the proportion of variance accounted for in a sample, it is an optimistic estimate of the true relationship in the population (Hays, 1963). The population parameter is given by $(\sigma_y^2 - \sigma_{y|x}^2) / \sigma_y^2$, where σ_y^2 is the marginal variance

of Y, and $\sigma^2_{y|x}$ is the conditional variance of Y given any X. Hays proposed ω^2 as an alternative measure of explained variance and provided an estimate as:

$$\hat{\omega}^2 = \frac{SS_{between} - (K-1)MS_{within}}{SS_{total} + MS_{within}}. \quad (1)$$

For the two-group case with equal variance, this reduces to:

$$\hat{\omega}^2 = (t^2_{obs} - 1)/(t^2_{obs} + N_1 + N_2 - 1). \quad (2)$$

For further understanding of these measures in more complex ANOVA designs, the reader is referred to Dodd and Schultz, 1973; Dwyer (1974); Glass and Hakstian (1969); Halderson and Glasnapp (1972); and Vaughan and Corballis, 1969.

METHODS

Sample. We had two main criteria for selecting textbooks. First, the textbook needed to be recent, which we defined as a publication date of 1993 or later. Second, the textbook needed to be written specifically for psychology, education, or behavioral science audiences.

The textbooks were obtained from two sources. First, we accessed the *Faculty Online* website (http://www.facultyonline.com). This site provides information from *Monument Information Resource*, which collects sales information from university bookstores and sells this information to textbook publishers. Using this site, we obtained a list of the top selling statistics textbooks in psychology. (There was no separate list of top selling statistics titles in education). There were 27 books on this list. We excluded nine textbooks because they were either general statistics textbooks or were written specifically for the social sciences (i.e. sociology and social welfare). Of the 18 remaining textbooks, four sets had two versions—an introductory version, and a more thorough version. In order to have more advanced-level textbooks, and in order to reduce the redundancy in the sample, the introductory versions were excluded. Thus, 14 textbooks remained: Aron and Aron (1999); Glass and Hopkins (1996); Gravetter and Wallnau

(2002); Heiman (2000); Howell (2002); Hurlburt (1998); Jaccard and Becker (2002); Kiess (2002); McCall (2001); Minium, King, and Bear (1993); Pagano (2001); Runyon, Coleman, and Pittenger (2000); Welkowitz, Ewen, and Cohen (2000); Witte and Witte (2001).

We were concerned that these 14 textbooks might used primarily in undergraduate courses, and not adopted as often in graduate courses. We contacted the publishers of these textbooks to see if there were additional textbooks that they thought were more commonly used in graduate courses. We were given the names of eight more textbooks: Abrami, Cholmsky, and Gordon (2001); Bartz (1999); Diekhoff (1996); Hays (1994); Hinkle, Wiersma, and Jurs (2001); Lehman (1995); Sprinthall (2000); and Thorndike and Dinnel (2001).

The final sample consisted of 14 "popular" textbooks obtained from *FacultyOnline* and 8 textbooks recommended by the sales representatives. All of the texts were published in 1993 or later, and 14 were published in 2000 or later. Although these texts are a convenience sample, we believe it will be more than adequate for our purposes.

Data analysis. A coding sheet that contained closed- and open-ended items was developed. The closed-ended items consisted of a list of "old standard" statistical tests, namely the Karl Pearson chi-square test of independence; the one-sample , matched-pair, and two-sample *t*-tests; and one-factor, two-factor, and repeated measures ANOVA. For each significance test, we coded (a) whether the test was covered; (b) which ES statistics, if any, were covered; (c) the formula(s) presented for each ES statistic; and (d) whether the authors presented specific "rules of thumb" for interpreting a given ES statistic.

The coding sheet contained open-ended questions that addressed the issue of how to *interpret* ES statistics, as follows: (a) What recommendations, if any, did the authors make in terms of considering previous research in interpreting ES statistics? (b) What recommendations,

if any, did the authors make in terms of considering the nature of the research question in interpreting the magnitude of the ES statistic? (c) What rules of thumb, if any, did the authors present to help interpret the magnitude of the ES statistic? (d) What recommendations, if any, did the authors make in terms of considering the results of corresponding SST in interpreting ES statistics? (f) Did the authors discuss the interpretation of *p*-values as indicators of magnitude of effect? For these open-ended questions, we used a qualitative, inductive approach to analyzing the data.

Both authors independently completed one coding sheet per text. The textbooks were coded in random order.

RESULTS

Analyses Based on the Comparisons of Means

The Prevalence of Various Significance Tests for the Analysis of Means

A summary of the prevalence of the three *t*-tests and three ANOVA models is presented in Table 1, and a more detailed presentation is given in Table 2. As seen in these tables, all

Insert Table 1 about here.

Insert Table 2 about here.

three t-tests were covered in all 22 texts. In terms of the ANOVA models, one-factor ANOVA was covered in all of the texts, and two-factor ANOVA was covered in all but one of the texts. The repeated measures ANOVA model appeared 16 of the 22 texts.

The Prevalence of Various ES Statistics for the Comparison of Means

We found that a number of authors *only* presented the concept of "effect size" as a parameter in the larger context of discussing statistical *power*. For example, an author might have introduced the concept of the standardized difference between two means in the population as $\delta = (\mu_1 - \mu_2)/\sigma$, and might have discussed the role of δ in determining the power of a significance test. But this same author might *not* have described how to estimate δ in the sample or discussed the value of using this estimate in interpreting the practical value of the research findings. For our analyses, we only tallied those cases where an ES statistic was presented in the context of interpreting the practical significance of a research finding, and did not tally those cases where an ES parameter was presented only in the context of discussing statistical power.

A summary of the prevalence of various ES statistics for three *t*-tests and three ANOVA

models is presented in Table 1, and more detailed information is given in Table 2. For the one-group and matched-pair cases, ES statistics were covered infrequently. For the one-group case, *d* was only covered in 4 texts, and in the remaining 18 texts, *no* ES statistics were presented. For the matched-pair case, *d* was only covered in 6 texts, and in 12 texts, *no* ES statistics were presented. For the two-sample *t*-test, *d* (either d_3 or d_4 or both) was presented in 7 of the 22 texts, and r_{pb} was presented in 7 of the 22 texts (See Table 1). For the two-sample t-test, $\hat{\omega}^2$ was presented in 5 texts, which is more often than we had expected, given that an understanding of this statistic would require background knowledge in ANOVA.

Analysis of various ES statistics for I by J Contingency Tables

The Karl Pearson chi-square test of independence for *I* by *J* contingency tables was covered in all 22 textbooks. In eight of these texts, no corresponding ES statistic was presented (Bartz, 1999; Gravetter & Wallnau, 2000; Hurlburt, 1998; Keiss, 2002; McCall, 2001; Minium, King, & Bear, 1993; Pagano, 2001; Thordike & Dinnel, 2001). And in one additional text (Abrami, Cholmsky, & Gordon, 2001), an ES statistic was only presented for the 2 x 2 case.

Of the 15 textbooks where ES statistics were presented for the *I* by *J* case, the two most common statistics were Cramer's *V*, which was covered in 11 texts, and the contingency coefficient (*C*), which was covered in 7 texts. In fact, in five textbooks both *V* and *C* were presented. The only other ES statistic to appear for the *I* by *J* case was Goodman and Kruskall's lambda, which was covered in two textbooks (Hays, 1994; Hinkle, Weirsma, & Jurs, 1998).

The Algebraic Equivalence of Various Formulas for the Same ES Statistic

If we were to compare various textbook formulas for an "old standard" SST such as the matched-pair *t*-test, we might find that the texts vary slightly in terms of the formulas presented. But we would also find these formulas to be algebraically equivalent. To our surprise, this was

not the case for the "standard" ES statistics considered here. Depending on the choice of texts, one could arrive at very different values of the "same" ES statistic for d, $\hat{f}$, r_{pb}^2, and $\hat{\eta}^2$.

Comparing Formulas for d

We found two textbooks with atypical methods for computing d. For the one-sample case, Runyon, Coleman, and Pittenger (2000) presented the formula $d = [(\bar{x}_1 - \mu)/s_x]\sqrt{2}$ (pp. 316, 510), which will yield a larger ES statistic in absolute value than d_1. In the matched-pair case, Thorndike and Dinnel (2001, p. 334) presented the formula $d = \bar{x}_{diff} / s_p$. Because s_p rather than s_d is used in the denominator, Thorndike and Dinnel's statistic will typically be smaller in absolute value than d_2.

Comparing Formulas for $\hat{f}$

We found three different approaches to estimating f, *none* of which are algebraically equivalent. The first approach, taken by Thorndike and Dinnel (2001, p. 435), was attributed to Kirk (1995, pp. 179-181). For the one-factor fixed-effects model, unbiased estimates of the numerator and denominator of $f = \sigma_{mean} / \sigma_{within}$ are given by $[((K-1)/Kn)(MSB - MSW)]^{1/2}$ and $\sqrt{MSW}$, respectively. Kirk's estimate is thus

$$\hat{f}_1 = \sqrt{\frac{\frac{(K-1)}{Kn}(MSB - MSW)}{MSW}}, \tag{3}$$

where n is the number of participants per group, and K is the number of groups. Kirk further noted that this is algebraically equivalent to $\hat{f}_1 = \sqrt{\hat{\omega}^2/(1-\hat{\omega}^2)}$. The second approach was

$$\hat{f}_2 = \sqrt{\hat{\eta}^2/(1-\hat{\eta}^2)}. \tag{4}$$

This approach was presented two texts (Abrami, Cholmsky, & Gordon, 2000, p. 280; Runyon, Coleman, & Pittenger, 2000, pp. 380-381). The third approach was

$$\hat{f}_3 = \sqrt{s_{\bar{x}}^2 / MSW},$$

where

$$s_{\bar{x}}^2 = \frac{1}{(K-1)} \sum (\bar{x}_k - \bar{x})^2 = MSB / n. \quad (5)$$

This approach was also presented in two texts (Aron & Aron, 1999, p. 339; Hurlburt, 1998, p. 299).

Comparing Formulas for r_{pb}^2

In two texts, the formula $r_{pb}^2 = t^2_{obs} / (t_{obs}^2 + df)$ was presented as an ES statistic for the matched-pair *t*-test (Heiman, 2000; Thorndike & Dinnel, 2001, p. 333). But this formula is incorrect, and will not result in r_{pb}^2, but will instead equal $\hat{\eta}_{partial}^2$. To compute r_{pb}^2, a researcher would either need to use a more general formula for the Pearson-Product moment correlation, or analyze the data as a repeated measures ANOVA, and compute $\hat{\eta}_{total}^2 = SS_{conditions} / SS_{total}$.

Comparing Formulas for $\hat{\eta}^2$

We found one major problem with the way in which $\hat{\eta}^2$ was presented in some texts. In some cases, $\hat{\eta}_{partial}^2$ was presented, but the authors did not indicate that it was "partial". Because $\hat{\eta}_{partial}^2$ is typically larger than total $\hat{\eta}^2$, a researcher who unknowingly calculates and presents $\hat{\eta}_{partial}^2$ could easily misinterpret its magnitude. As further evidence of this confusion, note that in the three texts where $\hat{\eta}_{partial}^2$ was presented, Cohen's (1988) rules of thumb for small, medium, and large effects were presented for interpreting $\hat{\eta}_{partial}^2$, and these were the same rules of thumb used for total $\hat{\eta}^2$.

For two-factor ANOVA, $\hat{\eta}_{partial}^2$ was presented two textbooks. In one text (Sprinthall, 2000), the author clearly labeled $\hat{\eta}_{partial}^2$ as "partial", whereas in the other text (Aron & Aron,

1999), the authors explained that they were removing other sources of variance from the base, but never used the "partial" label. For repeated measures ANOVA, $\hat{\eta}^2_{partial}$ was presented in three texts, and *none* of the authors used the "partial" label. In two of these texts—Kiess (2002) and Jaccard and Becker (2002)—the authors explained that they were removing "variation due to individual differences", but in the third text (Sprinthall, 2000), no explanation was offered.

The same problem arose in the two texts where $\hat{\eta}^2$ was presented as an ES statistic for the matched-pair *t*-test (Jaccard & Becker, 2002, p. 314; Kiess, 2002, p. 214). In both texts, the formula for $\hat{\eta}^2$ was erroneously given as $t^2_{obs}/(t^2_{obs} + df)$, which is the same formula erroneously given for r^2_{pb} in the matched-pair case. As noted previously, $t^2_{obs}/(t^2_{obs} + df) = \hat{\eta}^2_{partial}$, but again, the authors of these texts did not indicate that this $\hat{\eta}^2$ was partial.

Conceptualizing Various ES Indices

Definitional versus Computational Formulas for Various ES Statistics

In this section, we would like to make a distinction between a *definitional* formula for a given index (either a statistic or parameter), which typically helps the reader gain a conceptual understanding of that index, and a *computational* formula, which is typically quicker to use but conceptually devoid of meaning. We have no objection to presenting a computational formula for a given ES index *per se*, provided that the reader is first introduced to a given ES index via a definitional formula. But in a number of texts, we found cases where the authors *only* presented short-cut computational formulas. In this section, we will present some of the problems we found in terms of authors' choice of formulas. Our purpose here is not to review every instance of a conceptually-limited computational formula, but rather to give the reader an overview of some pedagogically-questionable approaches we found in explaining ES indices.

Computational formulas for d. Sprinthall (2000) consistently used conceptually

meaningless formulas for d. In the one sample case, he presented the formula: $d = t_{obs} / \sqrt{N}$, where t_{obs} is the value of the test statistic and N is the number of participants (p. 169). For the matched-pair case, he again used this formula, where N is the number of pairs (p. 410). Lastly, for the two-sample case, Sprinthall used the formula $d = t_{obs} / \sqrt{(N_1 + N_2)/(N_1 N_2)}$ (p. 245).

Computational formulas for r_{pb}^2. There are many computational formulas for r_{pb}^2, including $r_{pb}^2 = t_{obs}^2 / (t_{obs}^2 + df)$. We believe that in order for a reader to gain a conceptual understanding of r_{pb}^2, an author would need to introduce the concept of dummy-coding, and then explain that $r_{pb}^2 = t_{obs}^2 / (t_{obs}^2 + df)$ is equivalent to a definitional formula such as $r_{xy}^2 = [(\sum z_x z_y)/(N-1)]^2$, where x is a dummy-coded variable. Several authors did present the formula $r_{pb}^2 = t_{obs}^2 / (t_{obs}^2 + df)$, and did point out the connection of this formula to a definitional formula for r_{xy}^2. But Heiman (2000) did not, and in fact, he indicated that r_{pb}^2 is a *different* correlation coefficient from the Pearson correlation coefficient that is "calculated differently" (p. 178). Thus, a reader of Heiman's text would have no real conceptual understanding of r_{pb}^2.

Computational formulas for $\hat{\eta}^2$. The statistic $\hat{\eta}^2$ is relatively easy to understand as the ratio of between-groups sums of squares to total sums of squares. We did not find any major problems with cookbook approaches to computing $\hat{\eta}^2$ when it was introduced in the context of ANOVA. But in two texts (Kiess, 2002; Jaccard & Becker, 2002), the authors introduced $\hat{\eta}^2$ an appropriate ES statistic for the two-sample t-test, which means that they introduced $\hat{\eta}^2$ before they introduced ANOVA. Because the concept of "sums of squares" was not yet well-developed, the explanation of $\hat{\eta}^2$ was limited.

Computational formulas for $\hat{\omega}^2$. Omega-squared is more complicated to explain and

understand than $\hat{\eta}^2$, because it requires an understanding of the connection between the population parameter, ω^2, and the sample estimate $\hat{\omega}^2$. Equation (1), given earlier, would be conceptually meaningless unless it was explained in terms of variance component estimates of the parameter $\omega^2 = (\sigma_y^2 - \sigma_{y|x}^2)/\sigma_y^2$. Stated differently, a reader would need to understand the correspondence between the variance component estimates in Equation (1) and the population variances given in $(\sigma_y^2 - \sigma_{y|x}^2)/\sigma_y^2$.

Of the nine textbooks where $\hat{\omega}^2$ was presented, the authors of three texts did a thorough job of explaining ω^2 and its estimate, $\hat{\omega}^2$, namely Hays (1994), Howell (2002, pp. 353, 446-449), and Thorndike and Dinnel (2001, p. 434). Howell did a particularly thorough job, and included a table of estimates of variance components for one-factor, two-factor, and three-factor ANOVA designs for fixed and random variables (p. 447).

The remaining six textbooks had more of a cookbook quality to them. The two most problematic textbooks were Diekoff (1996, pp. 217, 253-255) and Runyon, Coleman, and Pittenger (p. 338-339, 380). In Diekoff's text, $\hat{\omega}^2$ was never described as measure of explained variance. In both textbooks, $\hat{\omega}^2$ was never described as an estimate, and in both textbooks, the population equation $(\sigma_y^2 - \sigma_{y|x}^2)/\sigma_y^2$ was never presented or described. For the one-factor ANOVA, Diekoff presented Equation (1), which we would argue is conceptually-meaningless when presented without explanation, whereas Runyon, Coleman, and Pittenger presented the computational formula $\hat{\omega}^2 = [df_{between}(F-1)]/[df_{between}(F-1)+N]$, which we would argue is equally conceptually-meaningless (p. 380). For the two-group situation, the authors of both texts presented Equation (2), which has no conceptual value, and Diekoff also presented Equation (2) for the matched-pair case.

Pagano (2001, pp. 336, 367) and McCall's (2001, pp. 275-276, 385) presentations of

$\hat{\omega}^2$ were almost equally problematic, and both had a cookbook quality to them in terms of explaining $\hat{\omega}^2$. Both authors described $\hat{\omega}^2$ as a measure of explained variance and as an estimate, but neither author presented the equation $\omega^2 = (\sigma_y^2 - \sigma_{y|x}^2) / \sigma_y^2$, and neither discussed $\hat{\omega}^2$ in terms of variance components estimates of the population parameters. Both Pagano and McCall presented Equation (1) and (2) as computational formulas. In terms of attempting to provide some conceptual understanding of $\hat{\omega}^2$, both authors introduced Equation (1) with a brief verbal description of sums of squares, but these explanations were not particularly illuminating.

Computational formulas for $\hat{f}$. We believe that $\hat{f}$ would also be more complicated to explain and understand than $\hat{\eta}^2$, because it also requires an understanding of the connection between the population parameter—f and the sample estimate $\hat{f}$. Because we believe that Kirk's (1995) estimate of $\hat{f}$ -- given earlier in Equation (3)—is the most reasonable of the three estimation procedures we found, we will focus on this estimate.

We would argue that Equation (3) is conceptually meaningless unless it is explained in terms of variance component estimates of the parameter $f = \sigma_{mean} / \sigma_{within}$. (In fact, of the five textbooks where $\hat{f}$ was presented, the parameter f was only given in one text, namely Aron and Aron, 1999). Stated differently, in order to understand this estimate, a reader would need to understand the correspondence between the variance component estimates in Equation (3) and the population standard deviations given in $f = \sigma_{mean} / \sigma_{within}$. Kirk's estimate was used by Thorndike and Dinnel (2001, p. 435), who introduced $\hat{f}$ by saying that it is an average ES for an entire study. But these authors only provided the reader with Equation (3) for calculating $\hat{f}$, and never presented the population equation $f = \sigma_{mean} / \sigma_{within}$. Thus, their presentation was conceptually devoid of meaning.

Interpreting the Magnitude of Effect

Deciding When to Calculate ES Statistics

In 13 of the 18 texts where ES statistics were covered, the authors indicated that an ES statistic should be calculated *following* a statistically significant result. (See Aron & Aron, 1999; Diekoff, 1996; Heiman, 2000; Hinkle, Wiersma, & Jurs, 1998; Howell, 2002; Hurlburt; 1998; Kiess, 2002; McCall, 2001; Pagano, 2001; Runyon, Coleman, & Pittenger, 2000; Sprinthall, 2000; Witte and Witte, 2001; Welkowitz, Ewen, & Cohen, 2000). But these authors varied greatly in terms of the extent to which they provided a rationale for this two-step process. Aron and Aron (1999) gave the most detailed explanation:

> ... in evaluating a study, there are two steps. First, you consider whether the result is statistically significant. If it is, this means you consider there to be a real effect. Then you consider whether the effect size is large enough to make the results useful or interesting. This second step is especially important if the study has any potential practical implications ... If the sample was small, you can assume that a significant result is probably also practically important. But if the sample size is very large, you must consider the effect size directly, as it is quite possible in such a case that it is too small to be useful. (p. 240)

Even though the authors of these 13 textbooks all said that ES statistics should be calculated following statistically significant results, with the exception of Aron and Aron (1999), none of them really explained why they thought that ES statistics should *not* be calculated following nonsignificant results. In fact, most of these authors were "fuzzy" on this issue—they only *explicitly* stated that ES statistics should be calculated following significant SSTs, and never *explicitly* said that ES statistics should *not* be calculated following non-significant SSTs. Rather,

the reader would need to infer this recommendation.

There were only three textbooks where the authors *explicitly* argued that ES statistics should be calculated regardless of whether or not the corresponding SST is significant (see Glass and Hopkins, 1996, p. 449; Jaccard, 2002, p. 278; Minium, King, & Bear, 1993, p. 366). As Minium, King, and Bear stated:

> … a nonsignificant result may still be important as judged by the ES. As an example, consider the following two studies described by Rosnow and Rosenthal (1989a). The studies are real, but the names have been changed. Professor Smith uses 80 subjects to compare two styles of leadership and discovers that style A is significantly better than style B at fostering productivity ($t = 2.21$, $df = 78$, $p < .05$). Professor Jones, who invented style B, is not pleased with this result and replicates the study using only 20 subjects; he reports nonsignificant results ($t = 1.06$, $df = 18$, $p > .30$). Although the p-values differ substantially, the estimated effect size is the same for both studies ($d = .50$). Thus, the second study did not really contradict the first. Professor Jones' power to reject the null hypothesis was much lower than Professor Smith's because of the smaller sample size (p. 366).

Using p-values as Indicators of Magnitude of Effect

None of the authors either used *p*-values as indicators of ES in interpreting SST results, or recommended using *p*-values as indicators of ES. And with one exception, none of the authors used terms like "*very significant*" or "*highly significant*" in interpreting SST results. [In an example of the matched-pair *t*-test, Glass and Hopkins stated: "The practice effect was highly significant—H_0 is rejected at the .001 level of statistical significance. Note that a highly statistically significant difference (e.g., $p < .001$) does not necessarily indicate a large difference

in means" (pp. 298-299)].

We were also interested in seeing if the authors recommended *against* using *p*-values as indicators of ES. There were only five textbooks where the authors discussed in any detail misconceptions in interpreting *p*-values as indicators of magnitude of effect (Abrami, Cholmsky, & Gordon, 2001, p. 212; Bartz, 1999, p. 284; Minium, King, & Bear, 1993; Runyon, Coleman, & Pittenger, 2000, pp. 316, 337, 339; Thordike & Dinnel, 2001, p. 319). Abrami, Cholmsky, and Gordon, for example, stated:

> Tests of significance tell us whether a nonchance relationship among variables is likely but do not tell us the magnitude of the relationship. Consequently, it is inappropriate to imply a large effect or use the phrase "very significant" when the exact probability of a calculated value is very small. (p. 212)

Using "Rules of Thumb"

ES statistics were covered in 18 textbooks, and rules of thumb for these statistics were presented in 10 of these texts. A summary of the prevalence of these rules statistics is presented in Table 3, and more specific information for each text is provided in the last column of Table 2.

Insert Table 3 about here.

Virtually all of the rules of thumb were from Cohen (1988), who gave guidelines for "small", "medium", and "large" effects for a number of ES parameters. (In two cases, Cohen's guidelines were not used. Thorndike and Dinnel (2001) gave guidelines for $\hat{\omega}^2$, where less than .03 was considered "weak", .03 to .15 was "moderate"; and .15 or more was "strong". Keiss (2002) gave guidelines for $\hat{\eta}^2$, where .10 to .15 was considered "strong"). Cohen's guidelines for small, medium, and large effects, respectively, are given by .2, .5, and .8 for *d;* .01, .06, and

.14 for r_{pb}^2 and $\hat{\eta}^2$; and .10, .25, and .40 for $\hat{f}$.

We also examined authors' recommendations regarding the use of rules of thumb in interpreting practical significance, and found that authors' recommendations fell along a continuum. Some authors presented these rules of thumb as rough guidelines to be taken quite lightly, whereas other authors applied these rules quite rigidly in interpreting ES statistics in their texts.

Interpreting the Magnitude of ESS based on Previous Literature and Research Context

In addition to examining rules-of-thumb, we also were interested in whether these authors offered other methods to interpret the magnitude of the ESS in terms of practical significance. Cohen (1988), introduced his rules-of-thumb as a *convention*. He saw them as a useful tool when guidance was needed beyond the researchers evaluation of prior research or theory and as a means for examining power. He commented as follows:

> The terms "small," "medium," and "large" are relative, not only to each other, but to the area of behavioral science or even more particularly to the specific content and research method being employed in any given investigation. In the face of this relativity, there is a certain risk inherent in offering conventional operational definitions for these terms for use in power analysis in as diverse a field of inquiry as behavior science. (p. 25)

As indicated by his comment, one can see that he did not believe that his rules-of-thumb were more than just that--rules-of-thumb. In our investigation, we found that 8 of the 18 textbooks that covered ESS also pointed out that research context and previous research should be considered when interpreting the magnitude of an ESS. The discussion of these alternative ways to interpret ESS ranged from very simple statements (e.g. McCall, 2001) to more elaborate

discussion using real research. For example, in one textbook (Hinkle, Wiersma, and Jurs, 2001) the authors mentioned that a particular ES value, although considered minimum in one discipline, could be considered large and meaningful in another discipline. Researchers must therefore use their judgement and knowledge-base in their fields of study to assess the magnitude of ESS. In three texts (Jaccard & Becker, 2002; Runyon et al., 2000; Witte & Witte, 2001) the authors used real life research examples (e.g., findings in AIDS research, effect of aspirin on heart attacks) to illustrate how a low effect size, say $d = .01$, in one discipline could be considered substantial when impacting even a few individuals' lives.

SUMMARY AND RECOMMENDATIONS

Summary

In light of Huberty's (1993) earlier finding that only 2 of 19 statistics textbooks published from 1990 to 1992 contained an ES statistic for the two-group *t*-test, the overall prevalence of ES statistics for the significance tests considered here was higher than we had expected. Although we used a number of different textbooks than Huberty, our study provides some evidence that the prevalence of ES reporting has increased. But prevalence rates only tell part of the story. Of the textbooks that included ES indices, we found a number of problems in the ways they were presented, some of which were quite unexpected. First, in terms of the ways that effect size theory was introduced, we found that most textbook authors did not distinguish between ES *parameters* and ES *statistics*. Second, when comparing formulas across textbooks for what was supposedly the "same" ES statistic, we found cases where these formulas were not algebraically equivalent. Third, when examining the formulas that textbook authors chose to introduce a given ES statistic, we found a number pedagogically-unsound cases where only "cookbook" computational formulas were provided. Fourth, when evaluating the ways in which authors addressed the issue of interpreting the magnitude of ES statistics, we found a number of cases where the authors relied solely on Cohen's (1988) rules of thumb or suggested the use of previous research or importance of the consequences of the effect, and other cases where the issue of interpretation wasn't discussed at all. In sum, although the prevalence of ES statistics might have been higher than expected, the presentation of ES theory, computation, and interpretation was often problematic.

Recommended "Best Practices" for Statistics Textbooks in Psychology and Education

We would like to end this paper with some recommendations for best practices in

presenting effect size theory, computation, and interpretation in statistics textbooks. In writing these recommendations, we realized that we had to take into consideration the practical reality that there are different kinds of statistics textbooks with different audiences and goals in mind. At one end of the continuum, there are texts that are written for a basic 3-unit first-course for undergraduate or master's degree students. These texts tend to be brief and cover just the "essentials". At the other end of the continuum, there are longer and more detailed textbooks that are written for a more rigorous first-course for doctoral students. Regardless of the intended level of detail of a given textbook, we believe that all statistics textbook authors should integrate effect size theory, computation, and interpretation as a core component of their texts. We make the following recommendations:

1. Distinguish between ES statistics and parameters. Cohen (1994) recommended that effect sizes be routinely reported in the form of confidence limits. We appreciate the fact that authors of introductory statistics textbooks might not have the space to present more advanced issues in effect size theory--such as bias-corrected ES estimates or CI estimation of ES parameters--in their texts. But in a number of statistics textbooks, it was never even made clear that a given ES statistic could be viewed as an estimate of a population parameter. The danger of not clearly identifying an ES statistic as a *statistic* is that the reader might not appreciate the fact that statistical issues such as sampling error—what Cohen called the "crud" factor--still need to be considered when interpreting the magnitude of an ES statistic.

2. Present an appropriate ES statistic for all "old standard" statistical significance tests covered. SSTs and ES indices need to be presented as integrated analytic tools. By only presenting a particular SST without a corresponding ES statistic, the reader may turn to *p*-values as a means of interpreting the magnitude of effect.

3. Introduce each ES index with the most conceptually-meaningful definitional formula possible. If a reader doesn't have a good conceptual understanding of a given ES index, then he or she will not be able to meaningfully interpret a particular research result, and may, in fact, be more likely to apply rules of thumb in a rigid and rote "cookbook" way.

4. For "basic" introductory textbooks, present d as an appropriate ES statistic for the one-sample, matched-pair, and two-sample t-tests. The lowest prevalence of ES statistics coverage was for the one-sample and matched-pair *t*-tests. In a number of cases, this appears to be because the textbook author(s) wanted to be consistent across-the-board for the various analyses of means, and focused only on measures of explained variance. The problem with this consistent approach is that there is no measure of explained variance for the one-sample *t*-test, and an appropriate measure of explained variance for the matched-pair *t*-test really needs to be presented in terms of repeated measures ANOVA. By presenting *d* as an appropriate ES statistic for these three *t*-tests, there would be a higher prevalence of ES coverage. Moreover, from a pedagogical perspective, presenting *d* makes more sense than presenting either $\hat{\eta}^2$ or $\hat{\omega}^2$ for the matched-pair and two-sample *t*-tests, because $\hat{\eta}^2$ and $\hat{\omega}^2$ can't be well-understood until ANOVA is introduced. Finally, we would argue that it is important to present *d* for all three *t*-tests because it is a common ES index and will be encountered regularly in published research.

5. For "basic" introductory textbooks, present $\hat{\eta}^2$ as an appropriate ES statistic for ANOVA designs. Although $\hat{\omega}^2$ might be considered more desirable than $\hat{\eta}^2$ because it is a less biased estimate, from a pedagogical perspective, $\hat{\eta}^2$ is easier to understand, and $\hat{\omega}^2$ requires more depth of understanding of ANOVA and variance estimation. For a "basic" textbook that does not cover ANOVA in detail, the reader may only gain a "cookbook" understanding of $\hat{\omega}^2$, and may not be able to meaningfully interpret its value. We would argue that it is better to have

a good conceptual understanding of a more-biased statistic ($\hat{\eta}^2$) than a cookbook understanding of a less-biased statistic ($\hat{\omega}^2$).

6. For textbooks where an author chooses to present $\hat{\eta}^2_{partial}$ as an appropriate ES statistic, $\hat{\eta}^2_{total}$ should also be presented. It is beyond the scope of this paper to argue the merits of $\hat{\eta}^2_{partial}$ versus $\hat{\eta}^2_{total}$. But from a pedagogical perspective, we believe that if a reader is going to learn $\hat{\eta}^2_{partial}$, then he or she needs to be able to understand the distinction between $\hat{\eta}^2_{partial}$ and $\hat{\eta}^2_{total}$ in order to make an intelligent choice between the two. Moreover, a reader who only learns $\hat{\eta}^2_{partial}$ may be more likely to falsely interpret the magnitude of a given result, because $\hat{\eta}^2_{partial}$ is usually larger than $\hat{\eta}^2_{total}$. Finally, a reader who only learns $\hat{\eta}^2_{partial}$ will be at a disadvantage when reading and interpreting published research, because he or she is more likely to encounter $\hat{\eta}^2_{total}$.

7. For "advanced" introductory textbooks where an author presents $\hat{\omega}^2$ as an appropriate ES, $\hat{\eta}^2$ should also be presented and its limitations discussed. Because $\hat{\eta}^2$ is such a common ES statistic, a reader who only learns $\hat{\omega}^2$ will be at a disadvantage when reading and interpreting published research, because he or will also encounter $\hat{\eta}^2$. Moreover, from a pedagogical perspective, we would argue that it would be better to introduce the reader to $\hat{\omega}^2$ by first explaining the main problem with $\hat{\eta}^2$, namely that it is a more-biased estimate of the population parameter.

8 Take an explicit stance on whether ES statistics should be presented for all statistical significance tests, regardless of whether they are statistically significant, or for only statistically significant results. A discussion of these merits of these two different schools of thought is beyond the scope of this paper. But we did find that many authors were "fuzzy" on this issue. (In fact, we thought that we could draw inferences about various authors' stances on this issue by

finding examples of nonsignificant results of SSTs in the textbooks, and seeing if the author(s) computed ES statistics for these nonsignificant results. This wasn't possible, because practically all examples of SSTs in the textbooks were statistically significant). By taking an explicit stance and by justifying this stance, the textbook author(s) will then have to grapple with issue of the value of SSTs in general as well as to address the issue of the "dependability" of an ES statistic.

9. Present Cohen's (1988) rules of thumb for interpreting the magnitude of an ES statistic, and take an explicit stance on whether these rules should be used, and if so, how. We are not proponents of using Cohen's rules of thumb for interpreting the magnitude of an effect. But because they are commonly used in published research, we believe that a reader needs to know how to use them (if at all) as well as to know how to avoid misusing them.

10. For each statistical significance test and corresponding ES statistic illustrated in the textbook with actual data, the author(s) should interpret the practical significance of the finding. We found a number of cases where textbook authors "side-stepped" the issue of interpreting the magnitude of an ES statistic. Given the complexities and ambiguities involved in such interpretation, in some ways, this side-stepping is easy to understand. But by not interpreting the magnitude of an effect, a void is left for the reader, and he or she has nothing to use as a model for his or her own research. By recommending that authors interpret the practical significance of the findings for each analysis of actual data, the issue of using real (or at least realistic) data in statistics texts arises. We believe that it would be much harder to model the interpretation of the practical significance of a finding that is based on hypothetical data than to model the interpretation of the practical significance of a study based on real (or at least realistic) data. Thus, we believe that that textbook examples of data analyses need to be based on real or realistic data.

11. All end-of-chapter problem sets that require computing a statistical significance test should also require computing and interpreting an appropriate ES statistic. Although we did not conduct a formal analysis of end-of-chapter problem sets, our informal analyses suggested that many data analysis problems only required computing SSTs. We believe that pedagogically, this is problematic because it fails to integrate the findings from SSTs and ES statistics in interpreting practical significance, and instead gives the impression that SSTs are of primary importance.

As a final recommendation, we found many cases of "standard" ES statistics that were not standard at all. In the world of ES statistics, there is still a lot of potential for misinterpreting ES statistics that are presented in published research. To avoid possible confusion, we believe that even the most "standard" ES statistics should be accompanied by computational formulas in journal articles.

REFERENCES

*Abrami, P. C., Cholmsky, P. & Gordon, R. (2001). *Statistical analysis for the social sciences: An interactive approach.* Boston, MA: Allyn & Bacon.

Aiken, L. S., West, S. G., Sechrest, L., & Reno, R. R. (1990). Graduate training in statistics, methodology, and measurement in psychology. *American Psychologist, 45,* 721-734.

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

*Aron, A. & Aron, E. N. (1999). *Statistics for psychology (2nd ed.).* Upper Saddle River, NJ: Prentice-Hall, Inc.

*Bartz, A. E. (1999). *Basic statistical concepts (4th ed.).* Upper Saddle River, NJ: Prentice-Hall, Inc.

Brewer, J. K. (1985). Behavioral statistics textbooks: Source of myths and misconceptions? *Journal of Educational Statistics, 10*(3), 252-268.

Brogan, D. R. (1980). A program of teaching and consultation in research methods and statistics for graduate students in nursing. *The American Statistician, 34,* 26-33.

Carver, R. P. (1993). The case against statistical significance testing, revisited. *Journal of Experimental Education, 61*(4), 287-292.

Cobb, G. W. (1987). Introductory textbooks: A framework for evaluation. *Journal of the American Statistical Association, 82,* 321-339.

Cohen, J. (1973). Eta-squared and partial eta-squared in fixed factor ANOVA designs. *Educational and Psychological Measurement, 33,* 107-112.

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

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45*(12), 1304-1312.

Cohen, J. (1994). The Earth is round ($p < .05$). *American Psychologist, 49*(12), 997-1003.

Curtis, D. A., & Harwell, M. R. (1998). Training doctoral students in educational statistics in the United States: A national survey. *Journal of Statistics Education, 6(1).* [Online journal: http://www.amstat.org/publications/jse/v6n1/curtis.html].

*Diekhoff, G. M. (1996). *Basic statistics for the social and behavioral sciences.* Upper Saddle River, NJ: Prentice-Hall, Inc.

Dodd, D. H. & Schultz, R. F., Jr. (1973). Computational procedures for estimating magnitude of effect for some analysis of variance designs. *Psychological Bulletin, 79*, 391-395.

Dwyer, J. H. (1974). Analysis of variance and magnitude of effects: A general approach. *Psychological Bulletin, 81*, 731-737.

Faculty Online. (http://www.facultyonline.com).

Finch, S., Cumming, G. & Thomason, N. (2001). Reporting of statistical inference in the *Journal of Applied Psychology*: Little evidence of reform. *Educational and Psychological Measurement, 61*(2), 181-210.

Fleiss, J. L. (1969). Estimating the magnitude of experimental effects. *Psychological Bulletin, 72,* 273-276.

Glass, G. V., & Hakstian, A. R. (1969). Measures of association in comparative experiments: Their development and interpretation. *American Educational Research Journal,* 6, 403-414.

*Glass, G. V., & Hopkins, K. D. (1996). *Statistical methods in education and psychology (3rd ed.).* Boston, MA: Allyn and Bacon.

*Gravetter, F. J. & Wallnau, L. B. (2000). *Statistics for the behavioral sciences: A first course*

for students in psychology and education (5th ed.). Belmont, CA: Wadsworth/Thomson Learning.

Halderson, J. S. & Glasnapp, D. R. (1972). Generalized rules for calculating the magnitude of an effect in factorial and repeated measures ANOVA designs. *American Educational Research Journal,* 9, 301-310.

Harwell, M. R., Herrick, M. L., Curtis, D, Mundfrom, D, & Gold, K. (1996). Evaluating statistics texts used in education. *Journal of Educational and Behavioral Statistics, 21(1),* 3-34.

Hays, W. L. (1963). *Statistics for the psychologists.* New York, NY: Holt, Rinehart, and Winston, Inc.

*Hays, W. L. (1994). *Statistics (5th ed.).* Fort Worth, TX: Harcourt College Publishers.

*Heiman, G. W. (2000). *Basic statistics for the behavioral sciences.* Boston, MA: Houghton Mifflin Company.

Heldref Foundation. (1997). Guidelines for contributors. *Journal of Experimental Education,* 65, 95-96.

*Hinkle, D. E., Wiersma, W., & Jurs, S. G. (1998). *Applied statistics for the behavioral sciences (4th ed.).* Boston, MA: Houghton Mifflin Company.

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

Huberty, C. J. (1987). On statistical significance testing. *Educational Researcher, 16*(8), 4-9.

Huberty, C. J. (1993). Historical origins of statistical testing practices: The treatment of Fisher versus Neyman-Pearson views in textbooks. *Journal of Experimental Education, 61*(4), 317-333.

Huberty, C. J. & Barton, R. M. (1990). Applied multivariate statistics textbooks [book review]. *Applied Psychological Measurement, 14*, 95-101.

*Hurlburt, R. T. (1998). *Comprehending behavioral statistics (2nd ed.).* Pacific Grove, CA: Brooks/Cole Publishing Company.

Hyde, J. S. (2001). Reporting effect sizes: The roles of editors, textbook authors, and publication manuals. *Educational and Psychological Measurement, 61*(2), 225-228.

*Jaccard, J. & Becker, M. A. (2002). *Statistics for the behavioral sciences (4th ed.).* Belmont, CA: Wadsworth/Thomson Learning.

Keselman, H. J., Huberty, C. J, Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue, B., Kawalchuk, 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.

*Kiess, H. O. (2002). *Statistical concepts for the behavioral sciences (4th ed.).* Boston, MA: Allyn & Bacon.

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(2),* 213-218.

*Lehman, R. S. (1995). *Statistics in the behavioral sciences: A conceptual approach.* Pacific Grove, CA: Brooks/Cole Publishing Company.

Levin, J. R. & Robinson, D. H. (1999). Further reflections on hypothesis testing and editorial policy for primary research journals. *Educational Psychology Review, 11*(2), 143-155.

*McCall, R. B. (2001). *Fundamental statistics for the behavioral sciences (8th ed.).* Belmont,

CA: Wadsworth/Thomson Learning.

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.

*Minium, E. W., King, B. M., & Bear, G. (1993). *Statistical reasoning in psychology and education (3rd ed.).* New York, NY: John Wiley and Sons, Inc.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5,* 241-301.

*Pagano, R. R. (2001). *Understanding statistics in the behavioral sciences (6th ed.).* Belmont, CA: Wadsworth/Thomson Learning.

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

*Runyon, R. P., Coleman, K. A., & Pittenger, D. J. (2000). *Fundamentals of behavioral statistics (9th ed.).* Boston, MA: McGraw Hill Companies.

Schact, S. P. (1990). Statistics textbooks: Pedagogical tools or impediments to learning? *Teaching Sociology, 18,* 390-396.

Shaver, J. P. (1993). What statistical significance is, and what it is not. *Journal of Experimental Education, 61*(4), 293-316.

Snyder, P. A., & 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,* 335-348.

*Sprinthall, R. C. (2000). *Basic statistical analysis (6th ed.).* Boston, MA: Allyn & Bacon.

Thompson, B. (1994). 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. & Snyder, P. A. (1998). Statistical significance and reliability analyses in recent *Journal of Counseling and Development* research articles. *Journal of Counseling and Development, 76,* 436-441.

Thompson, B. (1999). Journal editorial policies regarding statistical significance tests: Heat is to fire as *p* is to importance. *Educational Psychology Review, 11*(2), 157-169.

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

*Thorndike, R. M. & Dinnel, D. L. (2001). *Basic statistics for the behavioral sciences.* Upper Saddle River, NJ: Prentice-Hall, Inc.

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.

Vaughan, G. M., & Corballis, M. C. (1969). Beyond tests of significance: Estimating strength of effects in selected ANOVA designs. *Psychological Bulletin, 72,* 204-223.

Welkowitz, J., Ewen, R. B., & Cohen, J. (2000). *Introductory statistics for the behavioral sciences (5th ed.).* New York, NY: Harcourt Brace & Company.

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

*Witte, R. S. & Witte, J. S. (2001). *Statistics.* New York, NY: Harcourt College Publishers.

Table 1

Summary of the prevalence of ES statistics that correspond to three t-tests and three ANOVA models (N = 22 textbooks)

Number of Texts	One Sample t-test	Matched Pair t-test	Two Sample t-test	One Factor ANOVA	Two Factor ANOVA	Repeated Measures ANOVA
N that cover significance test	22	22	22	22	21	16
N that cover one or more effect size statistics	4	9	16	18	14	8
d	4	6	7	3	1	1
$\hat{f}$	0	0	0	5	2	2
r^2_{pb}	0	2	7	0	0	0
$\hat{\eta}^2$	0	2	2	11	8	6
$\hat{\omega}^2$	0	0	5	8	6	3
$\hat{\varepsilon}$	0	1	1	1	0	0

Table 2

Summary of Effect Size Procedures for Analyses based upon Means

Author(s)/Year	One Sample t-test	Matched Pair t-test	Two Sample t-test	One Factor ANOVA	Two Factor ANOVA	Repeated Measures ANOVA[1]	Rule of Thumb[2]
Abrami et al./2000	--	d_2	d_3, d_4	$\hat{\eta}^2, \hat{f}, \hat{\omega}^2$	$\hat{\eta}^2, \hat{f}, d$[3]	$\hat{\eta}^2, \hat{f}, \hat{\omega}^2$	$d, \hat{f}, \hat{\omega}^2$
Aron & Aron/1999	--	d_2	d_3	$\hat{\eta}^2, \hat{f}$	$\hat{\eta}^2_{partial}$	--	$d, \hat{\eta}^2, \hat{f}$
Bartz/1999	--	--	--	--	--	N/A	NA
Diekoff/1996	--	--	$\hat{\omega}^2$	$\hat{\omega}^2$	$\hat{\omega}^2$	$\hat{\omega}^2$	--
Glass & Hopkins/1996	--	d_2	d_3, d_4	$d_{kk'}$[4]	--	--	--
Gravetter et al./2000	--	--	--	--	--	--	NA
Hays/1994	--	--	r_{pb}, $\hat{\omega}^2$	$\hat{\eta}^2$, $\hat{\omega}^2$	$\hat{\eta}^2$, $\hat{\omega}^2$	--	--
Heiman/2000	--	r^2_{pb}	r^2_{pb}	$\hat{\eta}^2$[7]	$\hat{\eta}^2$	$\hat{\eta}^2$	--
Hinkle et al./2001	--	--	--	$\hat{\omega}^2$	$\hat{\omega}^2$	--	--
Howell/2002	--	--	d_3, r^2_{pb}	$\hat{\eta}^2, \hat{\omega}^2$	$\hat{\omega}^2$	--	d
Hurlburt/1998	d_1, U	d_2, U[8]	d_3, U	$\hat{\eta}^2, \hat{f}, d_{max}$	--	d_{max}[10]	$\hat{\eta}^2, \hat{f}$
Jaccard & Becker/2002	--	$\hat{\eta}^2, \varepsilon^2$	$r^2_{pb}, \hat{\eta}^2, \varepsilon^2$	$\hat{\eta}^2$	$\hat{\eta}^2$	$\hat{\eta}^2_{partial}$	$\hat{\eta}^2$
Kiess/2002	--	$\hat{\eta}^2$	$\hat{\eta}^2$	$\hat{\eta}^2$	$\hat{\eta}^2$	$\hat{\eta}^2_{partial}$	$\hat{\eta}^2$
Lehman/1995	--	--	--	--	--	--	NA
McCall/2001	--	--	$\hat{\omega}^2$	$\hat{\omega}^2$	$\hat{\omega}^2$	N/A	--
Minium et al./1993	--	--	--	--	--	--	d
Pagano/2001	--	--	$\hat{\omega}^2$	$\hat{\omega}^2$	--	N/A	--

Table 2 (continued)

Author(s)/Year	One Sample t-test	Matched Pair t-test	Two Sample t-test	One Factor ANOVA	Two Factor ANOVA	Repeated Measures ANOVA[1]	Rule of Thumb[2]
Runyon et al./2000	d_1	--	$\hat{\omega}^2$	$\hat{f}, \hat{\omega}^2$	$\hat{f}, \hat{\omega}^2$	$\hat{\eta}^2\ \hat{f}, \hat{\omega}^2$	$d, \hat{f}$
Sprinthall/2000	d_1	d_2	d_3	$\hat{\eta}^2$	$\hat{\eta}^2_{partial}$	$\hat{\eta}^2_{partial}$	$d, \hat{\eta}^2$
Thorndike et al./2001	d_1	d_3, r^2_{pb}	d_3, r^2_{pb}	$\hat{\eta}^2, \hat{f}, d, \hat{\omega}^2$	N/A	N/A	$\hat{\omega}^2$
Welkowitz et al./2000	--	--	r_{pb}	$\hat{\varepsilon}$	--	N/A	--
Witte & Witte/2001	--	--	r_{pb}[13]	$\hat{\eta}^2$	$\hat{\eta}^2$	N/A	$r^2_{pb}, \hat{\eta}^2$

Notes: [1]Repeated measures ANOVA with one within-subjects and one between-subjects variable. [2]This column identifies each ES statistic where the author provided a rule of thumb for interpretation. [3]For comparing marginal and cell means. [4]Calculated all pairwise standardized mean differences. [7]Mentioned $\hat{\omega}^2$. Estimated CI for individual means for one- and two-factor ANOVA. [8]Cohen's nonoverlap U_3. [10]Calculated the difference between largest and smallest mean, and dividing by sqrt(2*$MS_{residual}$) [12]Discussed the importance of calculating *d*, but only provided formula for δ, and never demonstrated how to estimate δ in a sample. [13]Mentioned $\hat{\omega}^2$ but did not cover.

Table 3

Summary of the prevalence of "rules of thumb" for interpreting the magnitude of various ES statistics (N = 22 textbooks)

Effect Size Statistic	N of Texts that Reported "Rules of Thumb" for this Statistic	Total N of Texts that Cover this Statistic
d	6	8
r_{pb}^2	1	6
$\hat{\eta}^2$	6	11
$\hat{\omega}^2$	2	9
f	4	5
Cramer's V	2	11
Contingency Coefficient (C)	0	7

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC®

TM033823

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Effect Size Statistics: An Analysis of Statistics Textbooks used in Psychology and Education	
Author(s): ~~Cheri J~~ Deborah A. Curtis • Cheri J. Araki	
Corporate Source:	Publication Date:

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1 ☒

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

Level 2A ☐

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B ☐

Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, → please

Signature: Deborah Curtis	Printed Name/Position/Title: Deborah Curtis / Associate Professor
Organization/Address: DAIS/College of Education	Telephone: 415 338 1076 / FAX: 415 338 0568
	E-Mail Address: curtis@sfsu.edu / Date: 4/8/02

San Francisco State University
San Francisco CA 94132

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, or, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

ERIC CLEARINGHOUSE ON ASSESSMENT AND EVALUATION
UNIVERSITY OF MARYLAND
1129 SHRIVER LAB
COLLEGE PARK, MD 20742-5701
ATTN: ACQUISITIONS

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
4483-A Forbes Boulevard
Lanham, Maryland 20706

Telephone: 301-552-4200
Toll Free: 800-799-3742
FAX: 301-552-4700
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com

EFF-088 (Rev. 2/2000)