Contemporary Educational Psychology **25,** 241–286 (2000)
doi:10.1006/ceps.2000.1040, available online at http://www.idealibrary.com on IDEAL®

Measures of Effect Size for Comparative Studies: Applications, Interpretations, and Limitations

Stephen Olejnik

University of Georgia

and

James Algina

University of Florida

Although dissatisfaction with the limitations associated with tests for statistical significance has been growing for several decades, applied researchers have continued to rely almost exclusively on these indicators of effect when reporting their findings. To encourage an increased use of alternative measures of effect, the present paper discusses several measures of effect size that might be used in group comparison studies involving univariate and/or multivariate models. For the methods discussed, formulas are presented and data from an experimental study are used to demonstrate the application and interpretation of these indices. The paper concludes with some cautionary notes on the limitations associated with these measures of effect size. © 2000 Academic Press

For more than three decades data analysts have been recommending to researchers in the behavioral sciences that, in addition to a test for statistical significance, an effect size measure should also be reported with their findings (Cohen, 1965; Hays, 1963). The rationale for this recommendation rests on the fact that statistical significance does not imply meaningfulness. ‘‘Significance’’ based on a statistical test provides information on the likelihood of finding the observed relationship by chance alone (sampling error). While statistical ‘‘significance’’ helps to protect the researcher from interpreting an apparently large observed difference as meaning a true difference between populations when sample sizes are small, it does not protect the researcher from interpreting a trivially small observed difference as meaningful when sample sizes are large. Small differences can be statistically ‘‘significant’’

We express our appreciation to James F. Bauman, Nancy Seifert-Kessell, and Leah A. Jones for making the data used in this paper available to us.

Address reprint requests to James Algina, College of Education, University of Florida, Gainesville, FL 32611.

0361-476X/00 $35.00

simply because of a large sample size. This has prompted some to argue that tests of statistical significance are *not* generally useful (Carver, 1978, 1993; Cohen, 1994; Hunter, 1997; Kirk, 1996; Schmidt, 1992). Rather, these critics of significance testing argue that confidence intervals and measures of effect size should be the focal point of research findings. Among the more vocal advocates for the reporting of effect size measures has been Thompson (1996, 1997, 1999a, 1999b), who has been critical of the wording used in *Publication Manual of the American Psychological Association,* fourth edition (American Psychological Association, 1994), which *encourages,* rather than requires, authors to report effect size information (p. 18). While this view of mandating effect size measures is not shared by all methodologists (e.g., Frick, 1999; Levin & Robinson, 1999; Robinson & Levin, 1997), the long-anticipated report from APA Task Force on Statistical Inference recommends that researchers ‘‘always report effect size measures for primary outcomes’’ (Wilkinson et al., 1999, p. 599).

Over the years, several papers have been published tracing the historical developments of effect size measures (Dwyer, 1974; Glass & Hakstian, 1969; Maxwell, Camp, & Arvey, 1981; Richardson, 1996), suggesting alternative indices (Friedman, 1968; McGraw & Wong, 1992; Rosenthal & Rubin, 1982), and offering cautionary notes on the interpretation of these indices (Mitchell & Hartmann, 1981; Muray & Dosser, 1987; O’Grady, 1982; Sechrest & Yeaton, 1982; Strube, 1988). Textbooks on statistical methods have also included procedures for computing a number of indices of effect (e.g., Huberty, 1994; Keppel, 1991; Maxwell & Delaney, 1990; Stevens, 1996). In spite of an abundance of references, applied researchers have apparently not accepted the call for the use of effect size measures (Keselman et al., 1998; McNamara, 1978).

The purpose of the present paper is to present in a single source the computational formulas that can be used for most basic designs involving at least one qualitative factor (grouping factor) used by applied researchers and to apply these formulas using data from a published article. The statistical analyses that are included are contrasts, omnibus single- and multifactor analysis of variance with fixed and random factors, multivariate analysis of variance, univariate and multivariate analysis of covariance, and repeated measures and split-plot analysis of variance. Research designs involving only quantitative factors (regression, correlation) and qualitative response variables (contingency tables, discriminant analysis, and logistic regression) are not considered.

CONTEXT

To demonstrate the application of various measures of effect size, data from an experimental study reported by Bauman, Seifert-Kessell, and Jones (1992) are used. The researchers used a randomized groups pretest–posttest

TABLE 1
Treatment Means and Standard Deviations

		Treatment		
		TA	DRTA	DRA
Pretest	Mean	9.14	9.73	10.50
EDT	*SD*	3.34	2.69	2.97
Pretest	Mean	4.95	5.09	5.27
Strategy	*SD*	1.86	1.99	2.76
Posttest	Mean	7.77	9.77	6.68
EDT	*SD*	3.93	2.72	2.77
Posttest	Mean	8.36	6.22	5.55
Strategy	*SD*	2.90	2.09	2.04
Posttest	Mean	43.45	46.64	42.05
DRP	*SD*	7.86	7.64	6.62

Note. TA = Talk Aloud; DRTA = Directed Reading–Think Aloud; DRA = Directed Reading Activity.

design and were interested in comparing the effectiveness of three interventions designed to affect reading comprehension skills of fourth-grade students. The first intervention taught students to use a Talk-Aloud (TA) approach to implement several comprehension monitoring strategies. The second intervention used a Directed Reading–Thinking Activity (DRTA). The third intervention was a control condition using a Directed Reading Activity (DRA). From the original sample, one student in the TA and one student in the DRTA conditions dropped out of the study (for reasons unrelated to the program), leaving 22 students in each of the three conditions ($N = 66$). Two pretest measures were administered before the interventions began and three posttest measures were administered following the termination of the interventions. The raw data for the study are reported in the Appendix and the group means and standard deviations for the five measures are presented in Table 1. The first pretest measure, an error detection task (EDT), required students to identify 16 sentences that did not make sense within the context of an entire reading passage. The second pretest was a 15-item inventory asking students what strategies they used to understand the reading passage. The first posttest was similar to the first pretest but with a different reading passage. The inappropriate sentences in the first posttest were intentionally made more difficult to detect to avoid a ceiling effect. The second posttest was an 18-item inventory similar to the second pretest. The third posttest was the Degrees of Reading Power (DRP) test (1986), a standardized reading test designed to assess student reading comprehension.

Two orthogonal contrasts were of particular interest to the researchers. The researchers were interested in comparing (a) the two monitoring strategies TA and DRTA with the control condition DRA and (b) TA with DRTA. The researchers analyzed the data using a multivariate analysis of covariance using posttest EDT and DRP as the outcomes and the two pretests as the covariates. Following an omnibus test that was statistically significant at the .05 level, the authors reported that both contrasts were statistically significant at the .05 level for the EDT posttest, and the second contrast was statistically significant at the .05 level for the EDT posttest but not the DRP posttest. For the Strategy posttest both contrasts were statistically significant at the .05 level. Measures of effect size were not reported for any of the statistical tests.

In the sections that follow we will use that data collected by Bauman et al. (1992) to demonstrate a variety of effect size indicators for several designs and data analyses. We will review both standardized mean difference effect sizes and proportion of variance effect sizes. For each type of effect size we will review univariate and multivariate approaches. We will conclude with some cautions regarding the use and interpretation of the measures of effect.

THE STANDARDIZED MEAN DIFFERENCE EFFECT SIZE

In educational and psychological research an effect of interest can often be expressed as a contrast of means, that is, as a weighted sum of means in which the weights sum to zero. If there are *J* cells in a design, a contrast can be written as

$$\psi = c_1\mu_1 + \cdots + c_J\mu_J,$$

where

$$c_1 + \cdots + c_J = 0.$$

Perhaps the most common contrast in educational psychology is the mean difference

$$\mu_j - \mu_{j'},$$

where *j* and *j′* indicate two different means.

A problem with using a contrast to communicate the size of the effect is that a contrast is scale dependent: Its magnitude is expressed in terms of the scale of measurement of the variable for which the means are computed. A standardized mean difference is an attempt to overcome the scale dependence of a contrast. In the following we present both univariate and multivariate standardized mean differences. A univariate standardized mean difference is used when interest focuses on a contrast for just one outcome variable. When interest is focused on a contrast applied to several outcome variables a multivariate standardized mean difference can be used.

Univariate Standardized Mean Difference

Between-Subjects Designs

Single-factor designs. In Bauman et al. (1992) two contrasts were of specific interest. The complex contrast examined the effect of instruction by comparing the average of TA and DRTA with DRA:

$$\psi_1 = \frac{1}{2}(\mu_{\mathrm{DRTA}} + \mu_{\mathrm{TA}}) - \mu_{\mathrm{DRA}}.$$

The pairwise contrast examined the intensity of the instruction by comparing TA with DRTA:

$$\psi_2 = \mu_{\mathrm{DRTA}} - \mu_{\mathrm{TA}}.$$

Using only performance on posttest EDT, a statistical difference at the .05 level was obtained for each contrast ($t(63) = -2.51$, $p \doteq .015$, and $t(63) = 2.08$, $p \doteq .042$, for the first and second contrasts, respectively).

One approach to describing the magnitude of an effect is to estimate the standardized mean difference. Expressing the contrast in standard deviation units standardizes a contrast:

$$\delta = \frac{\psi}{\sigma}.$$

The quantity δ is called a standardized mean difference. When two means are compared the estimate of δ is often called Cohen's *d.* To estimate δ, sample means are used to estimate the population means μ_j. The quantity σ can be estimated in one of three ways presented in Text Box 1. In the following we refer to the estimate of σ as the *standardizer.*

A pooled standard deviation is a square root of a pooled variance. In Option C the pooled variance is

$$S^2_{\mathrm{pooled}} = \frac{(n_1 - 1)S_1^2 + \cdots + (n_J - 1)S_J^2}{(n_1 - 1) + \cdots + (n_J - 1)}$$

Text Box 1

Alternative Methods for Computing Standardizers

A. The standard deviation of one of the groups, perhaps most typically the control group (Glass, 1976).

B. The pooled standard deviation of the group being compared (Cohen, 1969).

C. The pooled standard deviation for all of the cells in the design (Hedges, 1981).

and is equal to the error mean square (*MSE*) from a one-way analysis of variance of the data. If Option B is used, only the sample sizes and variances for the groups included in the contrast are included in the calculation of S^2_{pooled}. An argument against Option B, presented in Hedges (1981) but attributed to Glass (1976), concerns the possibility that two estimated contrasts that are identical in size will have unequal values for $\hat{\delta}$ if Option B is used. If the evidence suggests that all J variances estimate the same population variance (i.e., the variances meet the equality of variance assumption) then Option C is the best option.

If the variance equality assumption is not met then the standard deviation for one of the J groups should be used as the standardizer. In the context of comparing an experimental and a control treatment, Glass, McGraw, and Smith (1981) recommended using the standard deviation for the control group, but pointed out that the experimental group standard deviation could be used. Glass et al. present an example in which $\overline{Y}_E = 52$, $S_E = 2$, $\overline{Y}_C = 50$, and $S_C = 10$. The estimated contrast is

$$\hat{\psi} = (52 - 50) = 2.$$

If the contrast is standardized by using $S_E = 2$, the standardized mean difference is

$$\hat{\delta} = \frac{2}{2} = 1.00.$$

If $S_C = 10$ is used the standardized mean difference is

$$\hat{\delta} = \frac{2}{10} = .20.$$

Both figures are correct. As Glass et al. point out, ‘‘These facts are not contradictory; they are two distinct features of a finding which cannot be expressed by one number’’ (p. 107). Thus, when the equality of variance assumption is violated, the researcher will have to select one standard deviation that expresses the contrast on the scale the researcher thinks is most important, or will have to report the mean difference standardized by several standard deviations and discuss the implications of these figures.

In Table 2 each contrast $[1/2(\mu_{\text{DRTA}} + \mu_{\text{TA}}) - \mu_{\text{DRA}}$ and $(\mu_{\text{DRTA}} - \mu_{\text{TA}})]$ for the EDT posttest was standardized by using each of the three options. When option A was used the standard deviation for the DRA group was used. The DRA treatment was the control treatment in Bauman et al. (1992). The effect sizes only differ slightly depending on the estimate used for σ.

Hedges (1981) pointed out that the standardized mean difference is a biased estimator of δ and recommended that the standardized mean difference be multiplied by a correction factor equaling $1 - 3/(4df - 1)$, where df is

TABLE 2
Univariate Standardized Mean Differences

Standardized mean difference	Contrast $\frac{1}{2}(\mu_{\mathrm{DRTA}} + \mu_{\mathrm{TA}}) - \mu_{\mathrm{DRA}}$	 $(\mu_{\mathrm{DRTA}} - \mu_{\mathrm{TA}})$
$\hat{\delta} = \frac{\hat{\psi}}{S_{\mathrm{c}}}$[a]	$\frac{2.09}{2.769} = .755$	$\frac{2.00}{2.769} = .722$
$\hat{\delta} = \frac{\hat{\psi}}{S_{\mathrm{pooled}}}$[b]	$\frac{2.09}{3.189} = .655$	$\frac{2.00}{3.397} = .589$
$\hat{\delta} = \frac{\hat{\psi}}{S_{\mathrm{pooled}}}$[c]	$\frac{2.09}{3.189} = .655$	$\frac{2.00}{3.189} = .627$

[a] S_{c} is the standard deviation of the control group.
[b] S_{pooled} is the pooled standard deviation of the groups involved in the contrast.
[c] S_{pooled} is the pooled standard deviation for all three groups.

the degrees of freedom for the standard deviation used in the denominator of the estimate of the effect size.

Although there are few guidelines for interpreting the magnitude of the standardized mean difference, Cohen (1988) recommended standardized mean differences of .2, .5, and .8 for small, medium, and large effects. Using these guidelines, the results indicate medium effects for both contrasts. Cohen (1988, pp. 21–23) also suggested that if normality could be assumed then the percent of overlap between the populations could also be used to provide a sense of meaningfulness for an effect. In the present example the difference between TA and DRTA provides a standardized mean difference of .63; using the standard normal Z distribution this difference would indicate that 50% of the DRTA group performed better than approximately 73% of the TA group. See Fig. 1.

Regardless of whether a contrast is standardized by using a standard deviation for a single group or a pooled standard deviation, when the factor is a manipulated factor and participants are randomly assigned to levels, the standard deviation measures the extent of variation in the full range of talent. This is not true when the factor is an individual difference factor. For example, consider a gender study and assume the standard deviations for males and females are similar enough to justify using S_{pooled}. This standard deviation measures the extent of variation for males alone and for females alone and therefore, if there is a gender effect, measures the extent of variation in a partial range of talent. The standard deviation for the full range of talent is the total standard deviation

$$S_{\mathrm{total}} = \sqrt{\frac{(N-2)S^2_{\mathrm{pooled}} + SS_{\mathrm{gender}}}{N-1}},$$

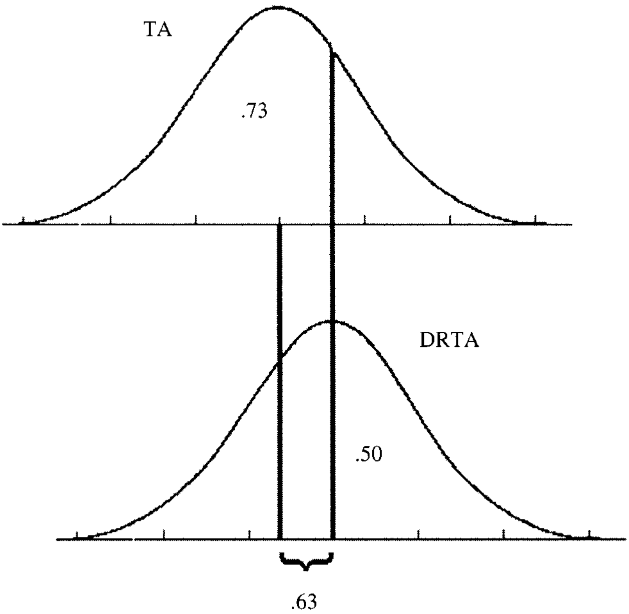

FIG. 1. Effect size as the percent of overlap.

which will be larger than S_{pooled}. It would have been salutary if S_{total} had been used historically to standardize mean differences in studies of individual difference factors. However, the practice of using one of Options A to C is now so firmly entrenched that it is probably a bad idea to change practice. However, the practice has important implications for standardizing mean differences in multifactor designs that include individual difference factors.

Multifactor designs. In a multifactor design, the factors can be manipulated factors, individual difference factors, or a combination of the two types of factors. If all factors are manipulated factors, no new issues arise in selecting the standardizer and the standardizer is calculated by (a) computing cell variances for a cross-classification in which all manipulated factors are included and (b) calculating the standard deviation by using one of the options presented in Text Box 1. If Option C is used the pooled variance is equal to the *MSE* for the design.

If one or more of the factors is an individual difference factor, additional issues arise about the calculation of the standardizer. We will discuss these issues in the context of two-factor designs. The issues concern comparability of standardized mean differences across different designs.

If one factor is a manipulated factor and one factor is an individual difference factor, the issue that arises is whether the standardizer should be calculated by ignoring or controlling the individual difference factor. To illustrate

TABLE 3
Means (and Standard Deviations) for Posttest EDT by Treatment and Ability Group

	Treatment			
Ability	TA	DRTA	DRA	
Low	5.18	8.64	5.45	6.42
	(2.40)	(2.38)	(2.30)	(2.78)
High	10.36	10.91	7.91	9.73
	(3.44)	(2.66)	(2.74)	(3.16)
	7.77	9.77	6.68	8.08
	(3.93)	(2.72)	(2.77)	(3.93)

how the resolution of the issue affects results we separated each of the treatment groups in the Bauman et al. (1992) study into two subgroups. Within each treatment group, students above the group's median on the pretest EDT were assigned to the high-ability group; the other students were assigned to the low-ability group. Thus we have created a 2 (Ability) × 3 (Treatment) between-subjects design. (In practice, when an ability factor is introduced into a design, participants should be classified into ability groups before assigning them to treatments. Our method of creating ability groups is for illustrative purposes only.) In the Appendix, the first 11 scores in each treatment group are for the low-ability students. Means and standard deviations are presented in Table 3.

Suppose we want to compare DRTA and TA. We have two means for TA ($\bar{Y}_{\text{TA, High}}$ and $\bar{Y}_{\text{TA, Low}}$) and two means for DRTA ($\bar{Y}_{\text{DRTA, High}}$ and $\bar{Y}_{\text{DRTA, Low}}$). The contrast of interest[1] is

[1] In this example we assume (a) each cell mean is computed as an arithmetic average of the data in the cell and (b) it is appropriate to compute the DRTA–TA contrast by averaging the DRTA–TA cell mean differences over the levels of the individual difference factor. The use of (a) is not optimal if there is no interaction between the treatment factor and the individual difference factor. When there is no interaction, the cell means can be estimated by using a linear model that includes only the main effects of the treatment and individual difference factor; the sampling variability of the cell mean differences can be improved by using these cell mean estimates. Using the average described in (b) assumes we want to control for the individual difference factor when we define the DRTA–TA contrast. This assumption will not be correct in all situations. For example, suppose the study involves Black and White students rather than high- and low-ability students. If we do not want to control ethnic background, the contrast should be calculated by using

$$p_{\text{B}}(\bar{Y}_{\text{DRTA,B}} - \bar{Y}_{\text{TA,B}}) + p_{\text{W}}(\bar{Y}_{\text{DRTA,W}} - \bar{Y}_{\text{TA,W}}),$$

where p_{B} and p_{W} are the proportion of students who are Black and White, respectively, in the population of interest. In the event of an Ethnic Group × DRTA–TA effect, this contrast can

$$\frac{1}{2}(\overline{Y}_{\text{DRTA, High}} + \overline{Y}_{\text{DRTA, Low}}) - \frac{1}{2}(\overline{Y}_{\text{TA, High}} + \overline{Y}_{\text{TA, Low}})$$

$$= \frac{1}{2}(10.91 + 8.64) - \frac{1}{2}(10.36 + 5.18)$$

$$= 2.00,$$

and is equivalent to a contrast of the marginal means for DRTA and TA. Because the cell frequencies are all equal, the DRTA–TA contrast in the two-factor design is equal to the DRTA–TA contrast in the single-factor design.

To standardize the mean difference we initially calculate the standardizer from the variance pooled over the six cells of the design (i.e., 7.188). The square root of this pooled variance is 2.681 and the standardized mean difference is

$$\hat{\delta} = \frac{2.00}{2.681} = .75.$$

In the single-factor design the pooled standard deviation was 3.189 and the standardized mean difference was .63. The .75 versus .63 difference arises because in the two-factor design, ability is controlled when the 2.681 standard deviation is calculated. In the single-factor design ability is ignored when the 3.189 standard deviation is calculated. Thus the two standardized mean differences are not comparable. The lack of comparability is extremely important. Suppose two investigators compare the same treatments. If the first uses a single-factor design and the second uses a two-factor design with an individual difference factor as the second factor, standardized mean differences for the treatment factor will tend to be larger for the second investigator. Because the difference in the sizes of the standardized mean difference reflects the design and not the treatments, the differences contribute confusion and not clarity to the literature. An indication of the magnitude of the problem introduced by ignoring design factors when estimating effect size was illustrated in Morris and DeShon (1997). Depending on the sample size and effect sizes associated with the individual difference and interaction factors in a two-factor design, the effect size estimated for the manipulated factor can vary from trivial to quite large.

The standard deviation (2.681) in the two-factor design measures the extent of variation in a partial range of talent and is smaller than the standard

be quite different than

$$\frac{1}{2}(\overline{Y}_{\text{DRTA,B}} - \overline{Y}_{\text{TA,B}}) + \frac{1}{2}(\overline{Y}_{\text{DRTA,W}} - \overline{Y}_{\text{TA,W}}).$$

Text Box 2

The Manipulated Factors Method

1. Compute cell variances for a cross-classification that (a) includes all manipulated factors and (b) excludes all individual difference factors, and
2. Calculate the standardizer by using one of the options presented in Text Box 1.

deviation for the full range of talent. The latter standard deviation is estimated in the single-factor design. We think the standardizer should reflect the full range of talent and consequently think that the standard deviation from the two-factor design is an incorrect standardizer. Therefore, when (a) some factors are manipulated factors and some are individual difference factors and (b) the contrast of interest is a contrast of levels of a manipulated factor, we recommend the procedure described in Text Box 2 for calculating the standardizer.[2] Thus instead of using 2.681, we would recommend using one of the standardizers in Table 2. Note that with this method, if Option C in Text Box 1 is used, S^2_{pooled} is not the *MSE* for the full design because the individual difference factors are excluded from the cross-classification.

Glass et al. (1981) agree with this recommendation when the contrast is averaged over all levels of the individual difference factor, as it was in our example. However, suppose a contrast of levels of the manipulated factor involves only one level of the individual difference factor. For example suppose we are interested in the contrast of DRTA and TA for high-ability students only:

$$\bar{Y}_{\text{DRTA, High}} - \bar{Y}_{\text{TA, High}} = 10.91 - 10.36 = .55.$$

Based on the example on page 119, Glass et al. would recommend the pooled standard deviation calculated from the two-factor design (2.681) as the standardizer. We disagree with their recommendation. Because we want the DRTA–TA standardized mean difference for high-ability students to be comparable to the DRTA–TA standardized mean difference for the entire sample, we recommend calculating the standardizer by using the manipulated factors method. In this example the manipulated factors method would

[2] If the researcher plans to standardize by using the standard deviation pooled over all cells obtained by a cross-classification of all manipulated factors (i.e., Option C in Text Box 1), the standard deviation can be calculated from the results in the original analysis by pooling the error variance and the variances of all main and interaction effects that involve the individual difference factors and taking the square root. The pooling is accomplished by adding the sums of squares for error and the sums of squares of all sources of variance involving the individual differences factors (main effect and interaction) and dividing by the sum of the degrees of freedom for all the pooled sources of variance. We present the procedure in Text Box 2 because it allows the use of Options A and B as well as Option C in Text Box 1.

Text Box 3

The Manipulated Factors and Individual Difference Method

1. Compute cell variances for a cross-classification in which (a) all manipulated factors are included, (b) the individual difference factor involved in the contrast is included, and (c) all other individual difference factors are excluded, and
2. Calculate the standardizer by using one of the options presented in Text Box 1.

result in using one of the standardizers in Table 2. Using Option C in Text Box 1, the standardizer would equal 3.189.

Suppose we are interested in comparing the DRTA–TA difference for high-and low-ability students. This is an example of an interaction contrast. The contrast of interest is

$$(\bar{Y}_{\text{DRTA, High}} - \bar{Y}_{\text{TA, High}}) - (\bar{Y}_{\text{DRTA, Low}} - \bar{Y}_{\text{TA, Low}})$$
$$= (10.91 - 10.36) - (8.64 - 5.18)$$
$$= -2.87.$$

Again we recommend calculating the standardizer by using the manipulated factor method, which would result in using one of the standardizers in Table 2.

Another possible contrast is a contrast of the levels of the individual difference factor. For example the High–Low ability contrast might be of interest:

$$\frac{1}{3}(\bar{Y}_{\text{DRTA, High}} + \bar{Y}_{\text{TA, High}} + \bar{Y}_{\text{DRA, High}}) - \frac{1}{3}(\bar{Y}_{\text{DRTA, Low}} + \bar{Y}_{\text{TA, Low}} + \bar{Y}_{\text{DRA, Low}})$$
$$= \frac{1}{3}(10.91 + 10.36 + 7.91) - \frac{1}{3}(8.64 + 5.18 + 5.45)$$
$$= 3.31.$$

To standardize a contrast of levels of an individual difference factor, we recommend the procedure described in Text Box 3 for calculating the standardizer. The standardizer computed using this method will be comparable to the standardizer that would have been used if the individual difference factor had been the only factor in the design. In the present example the cross-classification described in Point 1 is the cross-classification of the treatment factor and the ability factor. Thus the variances in Point 1 are the squares of the standard deviations in Table 3, and if Option C in text Box 1 were used the standardizer would be 2.681. Note that with this method, if the design had two or more individual difference factors and Option C in Text Box 1 were used, S^2_{pooled} would not be the *MSE* for the full design because all but one of the individual difference factors are excluded from the cross-classification described in Point 1.

To illustrate selecting a standardizer when all factors are individual difference factors consider an example in which one factor is ethnic background (Majority and Minority) and one factor is gender. The gender contrast is

$$\frac{1}{2}(\bar{Y}_{\mathrm{M,Maj}} + \bar{Y}_{\mathrm{M,Min}}) - \frac{1}{2}(\bar{Y}_{\mathrm{F,Maj}} + \bar{Y}_{\mathrm{F,Min}}).$$

If we calculate the standardizer by using the variances obtained from the two-way analysis of the data, the standardizer controls for both gender and ethnic background. In a single-factor gender study, the standardizer would control only for gender. We recommend using a standardizer that is comparable to the one used in the one-factor design: Calculate the standard deviation by treating gender as the only factor in the design. A parallel recommendation applies to the ethnic group contrast

$$\frac{1}{2}(\bar{Y}_{\mathrm{M,Maj}} + \bar{Y}_{\mathrm{F,Maj}}) - \frac{1}{2}(\bar{Y}_{\mathrm{M,Min}} + \bar{Y}_{\mathrm{F,Min}}).$$

Calculate the standard deviation by treating ethnic background as the only factor in the design. The general recommendation for computing a standardizer for a design in which all factors are individual difference factors is to

1. Compute variances for a design in which the individual difference factor involved in the contrast is the only factor, and

2. Calculate the standardizer by using one of the options presented in Text Box 1. If the design that includes all individual differences factors is a balanced design and if Option C is used, the pooled variance can obtained by using

$$S_{\mathrm{pooled}} = \sqrt{\frac{SS_{\mathrm{total}} - SS_{\mathrm{id}}}{N - J}},$$

where SS_{id} is the sums of squares for the individual differences factor of interest, N is the total sample size, and J is the number of levels of the individual differences factor.

Between-Subjects Designs with Covariates

Because of limited resources or access to populations of interest, many research studies are conducted with a small sample of participants. As a consequence of a small sample size statistical power (the likelihood of detecting true differences) can be low. Bauman et al. (1992), for example, had access to only 68 participants, with only 66 individuals completing the study. One strategy that can be used to increase statistical power is to obtain additional information, prior to the beginning the intervention, about the participants that is related to the outcome measures. Using these data, variation in the outcome measure (error variance) can be reduced by using analysis of covariance. Bauman et al. used this strategy in their investigation. Two pretest measures were obtained on each participant before the treatments were begun and their hypotheses about contrasts were tested using univariate and multivariate analyses of covariance. In the univariate analysis, both contrasts were statistically significant at the .05 level ($F_{\psi_1}(1, 61) = 20.14$, $p \doteq .001$; $F_{\psi_2}(1, 61) = 4.90$, $p \doteq .031$).

When analysis of covariance is used the means in the contrast should be estimated by adjusted cell means rather than the posttest cell means. Using the EDT and Strategy pretests as covariates and posttest EDT as the outcome measure the adjusted means were 8.21, 9.81, and 6.19 for the TA, DRTA, and DRA interventions, respectively. The estimate of the contrast of TA and DRTA versus DRA is

$$\frac{1}{2}(8.22 + 9.81) - 6.19$$

$$= 2.83.$$

The contrast for unadjusted posttest means was 2.09. If participants have been randomly assigned to groups the contrast of the adjusted means and the contrast of the unadjusted means will typically be similar because random assignment tends to result in small mean differences on the covariates. When participants are not randomly assigned to groups the difference between the two contrasts can be large. The standardizer should be calculated using the manipulated factors method (Text Box 2) so that it estimates the standard deviation ignoring the covariates. in the present example the only manipulated factor is the treatment factor. Therefore the standardizer is selected from one of the standard deviations used in Table 2. Using the standard deviation pooled over the three groups the standardized mean difference is

$$\hat{\delta} = \frac{2.83}{3.19}$$

$$= .89.$$

The standardizer used here is equivalent to the standard deviation often used by meta-analysts (e.g., Glass et al., 1981, p. 119, Eq. 5.19). This standardizer corrects the adjusted mean square error in ANCOVA as a function of the error degrees of freedom and the pooled within-group correlation between the covariate and dependent variables. We prefer our approach because it is simpler to compute and will help practitioners to understand the meaning of the effect size measure for these designs.

Within-Subjects Designs

In many research settings investigators are interested in comparing related or dependent samples. Two common contexts in which related samples are compared are when (a) participants are observed at several points in time and (b) participants are observed under several different treatment conditions. When the repeated measures on participants are to be compared the study is referred to as a within-subjects design.

In the Bauman et al. (1992) study, participants were observed at two points in time—two pretests and three posttests were given. We use the data on the EDT tasks to illustrate application of the formulas to a single-factor within-subjects design. Pretest EDT scores ($\overline{Y}_{\text{pre}} = 9.79$ and $S_{\text{pre}} = 3.02$) were compared with posttest EDT scores ($\overline{Y}_{\text{post}} = 8.08$ and $S_{\text{post}} = 3.39$). The observed difference was statistically significant at the .05 level ($F(1, 65) = 21.40$, $p \doteq .001$). Overall the test scores declined but it must be recalled that the reading passage used for the posttest EDT was chosen to be more difficult than the passage used for the pretest EDT to avoid a ceiling effect. When the design has a single factor, the formulas presented for the single-factor between-subjects design can be used to calculate the standardizer. For the current example we pooled variances 3.02^2 and 3.39^2 to obtain the pooled standard deviation 3.21. Thus we used Option B in Text Box 1, which is equivalent to Option C because there are only two time points. Option A could also have been used; the pretest standard deviation would most likely be used in Option A. The standardized mean difference is

$$\hat{\delta} = \frac{8.08 - 9.79}{3.21}$$

$$= -.53.$$

It is important to note that although the repeated measures are correlated, the standardizer can be obtained by treating the cell variances as independent estimates of the population variances (Dunlap, Cortina, Vaslow, & Burke, 1996). Option B is consistent with the intent of the approach recommended by Glass et al. (1981, p. 120) for calculating the standardizer.[3] However, to the degree that the cell variances are unequal, the two approaches will yield different numerical results. As with the analysis of covariance procedure discussed above, we believe our approach is simpler to compute and will help researchers understand the true meaning of the effect size measure.

[3] There are typographical errors in Eq. 5.25 in Glass, McGraw, and Smith (1981). Equation 5.25 should be

$$\frac{\sqrt{n}s_{\bar{d}}}{\sqrt{2(1 - r_{EC})}}.$$

TABLE 4
Means (and Standard Deviations) by Noise and Degree

	Offset degree		
Noise	0	4	8
Absent	462 (56.92)	510 (86.02)	528 (78.99)
Present	492 (88.54)	660 (109.54)	761 (116.79)

In our discussion of multifactor between-subjects designs we pointed out the complications that arise due to individual difference factors. These complications do not typically arise in within-subjects designs because typically individual difference factors are not used in these designs. Consequently standardized mean differences in multifactor within-subjects designs can be calculated by using the manipulated factors method described in Text Box 2. Maxwell and Delaney (1990, p. 497) presented synthetic data for a design with two within-subjects factors: The letters *T* and *I* are displayed on a computer screen to $n = 10$ participants. Participants' reaction times were collected when the letter was embedded either in a display of other letters (Noise Present) or displayed alone (Noise Absent). The letters were displayed at the center of the screen, 4° off center, or 8° off center. Means and standard deviations are presented in Table 4.

Suppose we are interested in a contrast of offset versus no offset for the noise absent condition. The contrast is

$$\frac{1}{2}(\bar{Y}_{4,\mathrm{NA}} + \bar{Y}_{8,\mathrm{NA}}) - \bar{Y}_{0,\mathrm{NA}}$$

$$= \frac{1}{2}(510 + 528) - 462$$

$$= 57.$$

Inspecting the standard deviations for the three conditions indicates some difference among the standard deviations. Considering the small sample size, the differences among the standard deviations may be due to sampling error. Nevertheless, the fact that the standard deviations increase with the degree of offset (and with the presence of noise) suggests the differences among the standard deviations are systematic. Consequently, we will use the control standard deviation (i.e., the standard deviation for the noise absent, and no offset cell) to standardize the mean difference:

$$\hat{\delta} = \frac{57}{56.92} = 1.00.$$

Another contrast that might be of interest is the difference, in the linear trend in the means as a function of offset, between noise absent and noise present. For noise present the linear trend for degree is

$$\bar{Y}_{8,\mathrm{NP}} - \bar{Y}_{0,\mathrm{NP}}$$

and for noise absent the linear trend for degree is

$$\bar{Y}_{8,\mathrm{NA}} - \bar{Y}_{0,\mathrm{NA}}.$$

Thus the contrast of interest is

$$\begin{aligned}&(\bar{Y}_{8,\mathrm{NA}} - \bar{Y}_{0,\mathrm{NA}}) - \bar{Y}_{8,\mathrm{NP}} - \bar{Y}_{0,\mathrm{NP}})\\ &= (761 - 492) - (528 - 462)\\ &= 203\end{aligned}$$

and the standardized mean difference is

$$\begin{aligned}\hat{\delta} &= \frac{203}{56.92}\\ &= 3.57.\end{aligned}$$

Split-Plot Designs

In a split-plot design there are both between-subjects and within-subjects factors. The within-subject factors are typically not individual difference factors. The between-subjects factors can be either manipulated factors or individual difference factors. If all the between-subjects factors are manipulated factors the standardizer is calculated by using the manipulated factors method (see Text Box 2). If one or more of the between-subjects factors is an individual difference factor, the procedure for calculating the standardizer depends on whether the levels of a manipulated factor or the levels of an individual difference factor are being compared. If the levels of a manipulated factor are to be compared, we recommend the manipulated factors method. If the levels of an individual difference factor are being compared, we recommend the manipulated and individual difference factor method (see Text Box 3).

Using the EDT pretest and posttest as the within-subject factors and the three interventions (TA, DRTA, and DRA) as the between-subjects factor, the Bauman et al. (1992) study can be used to illustrate a split-plot design. A contrast of interest would be a comparison of the average of mean change for the two treatment groups (TA and DRTA) to mean change for the control group (DRA):

$$\frac{1}{2}(\overline{Y}_{\mathrm{DRTA,\,post}} - \overline{Y}_{\mathrm{DRTA,\,pre}}) + \frac{1}{2}(\overline{Y}_{\mathrm{TA,\,post}} - \overline{Y}_{\mathrm{TA,\,pre}}) - (\overline{Y}_{\mathrm{DRA,\,post}} - \overline{Y}_{\mathrm{DRA,\,pre}})$$

$$\frac{1}{2}(9.77 - 9.73) + \frac{1}{2}(7.77 - 9.14) - (6.68 - 10.50)$$

$$= 3.16.$$

Because neither of the factors (Interventions and Time) is an individual difference factor, we use the manipulated factors method, which is described in Text Box 2, to compute the standardizer. That is, the standardizer is computed from the cell variances for a cross-classification of Interventions and Time. Inspecting the standard deviations in Table 1 for the EDT pretest and posttest, we find that the standard deviations are reasonably similar for all six cells. Therefore, we use Option C and compute the standardizer from the variances pooled over the six cells,

$$S_{\mathrm{pooled}} = \sqrt{\frac{3.34^2 + 2.69^2 + 3.93^2 + 2.72^2 + 2.77^2}{6}}$$

$$= 3.10,$$

and the standardized mean difference is

$$\hat{\delta} = \frac{3.16}{3.10} = 1.02.$$

Multivariate Standardized Mean Difference

The square root of the Mahalanobis D^2 statistic is a multivariate standardized mean difference that is analogous to the univariate standardized mean difference calculated by using the pooled (over all groups) standard deviation as the standardizer. In this section we explain the difference between a univariate and a multivariate standardized mean difference and illustrate calculation of the multivariate standardized mean difference. We limit the presentation to single-factor between-subjects designs.

Suppose a study has been conducted with two groups (A and B) and two variables (1 and 2). Let d_1 and d_2 be the standardized mean differences for the two variables, respectively:

$$d_j = \frac{\overline{Y}_{\mathrm{A}j} - \overline{Y}_{\mathrm{B}j}}{S_{\mathrm{pooled},j}}, j = 1, 2.$$

Then

$$D^2 = \frac{1}{1 - r_{12}^2}(d_1^2 + d_2^2 - 2r_{12}d_1d_2),$$

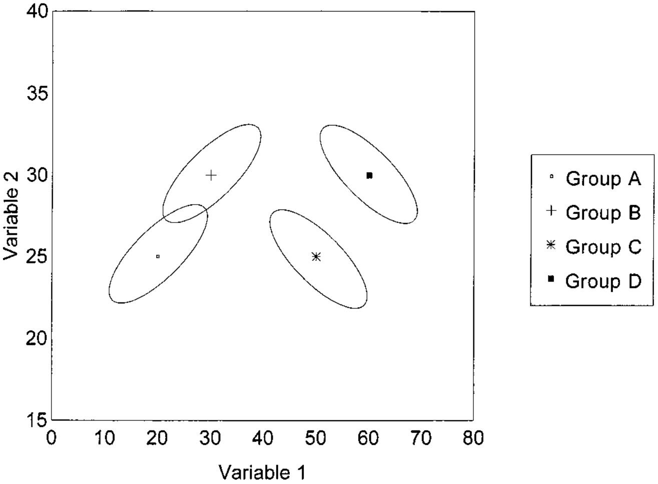

FIG. 2. Bivariate scatter plots for four groups.

where r_{12} is the pooled within-groups correlation between the two variables and measures the association between the two variables within either group. Thus Mahalanobis D^2 reflects not only the univariate standardized mean differences but also the correlation between the two variables. A general formula for the multivariate standardized mean difference is

$$D = \sqrt{\mathbf{d}'\mathbf{R}^{-1}\mathbf{d}},$$

where **d** is a vector of univariate standardized mean differences for the contrast of interest and **R** is the pooled within-groups correlation matrix. However, D can be computed from any of the multivariate test statistics (Roy's largest root, Lawley–Hotelling trace, Pillai–Bartlett trace, or Wilks's lambda) that can be used to test the significance of a miltivariate contrast. Subsequently, we illustrate using Wilks's lambda to calculate D. It should be noted that by the appropriate choice of **R,** D could be made comparable across different designs. When D is calculated from a multivariate test statistic, it may not be comparable across different designs.

To understand the difference between univariate and multivariate standardized mean differences, consider the scatterplots and means depicted for four groups (A, B, C, and D) in Fig. 2. Multivariate standardized mean differences comparing groups A and B and comparing groups C and D are to be computed. The mean difference for variable 1 is 10 for groups A and B and also for groups C and D. The standard deviation on variable 1 is the same for all four groups. Therefore the univariate standardized mean difference comparing groups A and B on variable 1 is equal to that for groups C and D. Similarly, the univariate standardized mean difference comparing groups A and B on variable 2 is equal to that for groups C and D. However, Fig. 1 suggests the separation between groups C and D is larger than that between groups A and B. By incorporating the $r_{12}d_1d_2$ product term, the multivariate standardized mean difference captures this feature of the data: For groups A and B, $r_{12}d_1d_2$ is a positive number and for groups C and D, $r_{12}d_1d_2$ is a negative number. Therefore the multivariate mean difference is larger for groups C and D.

Bauman et al. (1992) considered the EDT posttest and the DRP posttest to measure a single construct, reading comprehension. Consequently, they compared the three interventions with respect to the vector outcomes of the two posttests. Neither the TA&DRTA vs DA nor the TA vs DRTA multivariate contrast was statistically significant at the .05 level ($F(2, 62) = 3.122$, $p \doteq .052$, and $F(2, 62) = 2.130$, $p \doteq .128$, respectively). In Table 5 Wilks's lambda is used to compute D for the two contrasts of interest. Wilks's lambda is reported in SAS when multivariate contrasts are requested. SPSS (version 9.0) will only provide Wilks's lambda for pairwise contrasts in the multivariate GLM program if a separate program is run for each contrast. The square roots of the two measures of distance are .658 and .628 and are slightly larger than the corresponding univariate standardized mean differences reported in Table 2.

PROPORTION OF VARIANCE EFFECT SIZES

An alternative to a standardized mean difference for describing the magnitude of an effect is the proportion of the variation in an outcome measure that is explained by (shared with) the grouping variable. Many researchers are familiar with the concept of shared variance in the context of a correlation coefficient between two quantitative measures. In this context the squared Pearson product moment correlation (r_{xy}^2) provides an index of the strength of a relationship. A proportion of variance effect size can be used to express the size of a contrast or an omnibus effect. We review and illustrate both uses of proportion of variance effect sizes. The standardized mean difference is used more frequently than the proportion of variance to measure the size of a contrast. Consequently, we limit our coverage of proportion of variance effect sizes for contrasts to single-factor between-subjects designs.

Contrasts

Univariate approach. The magnitude of each contrast effect can be reported in three ways: eta squared (Pearson, 1905), epsilon squared (Kelley, 1935), and omega squared (Hays, 1963). Eta squared is computed as the ratio of the sum of squares for the contrast to the total sum of squares,

$$\hat{\eta}^2 = \frac{SS_{\text{contrast}}}{SS_{\text{total}}},$$

TABLE 5
Multivariate Standardized Mean Difference Effect Sizes

	Contrast	
Standardized mean difference	$\frac{1}{2}(\mu_{\text{DRTA}} + \mu_{\text{TA}}) - \mu_{\text{DRA}}$	$(\mu_{\text{DRTA}} - \mu_{\text{TA}})$
$D = \sqrt{\frac{df_{\text{error}}(1 - \Lambda)\sum c_j^2/n_j}{\Lambda}}$	$\sqrt{\frac{63(1 - .9085)\left[\frac{.5^2}{22} + \frac{.5^2}{22} + \frac{-1^2}{22}\right]}{.9085}} = .658$	$\sqrt{\frac{63(1 - .9357)\left[\frac{1^2}{22} + \frac{-1^2}{22}\right]}{.9357}} = .628$

where

$$SS_{\text{contrast}} = \frac{(c_1\overline{Y}_1 + \cdots + c_J\overline{Y}_J)^2}{\dfrac{c_1^2}{n_1} + \cdots + \dfrac{c_J^2}{n_J}}.$$

Eta squared, however, overestimates the population strength of the relationship and is best thought of as a descriptor of the amount of variation in the dependent variable that is shared with the grouping variable for a particular *sample* (Maxwell et al., 1981). Two corrections have been suggested. Epsilon squared corrects the numerator of eta squared by subtracting the error mean square from the contrast sum of squares:

$$\hat{\varepsilon}^2 = \frac{SS_{\text{contrast}} - MS_{\text{error}}}{SS_{\text{total}}}.$$

Omega squared further adjusts epsilon squared by adding the error mean square groups to the total sum of squares in the denominator of epsilon squared:

$$\hat{\omega}^2 = \frac{SS_{\text{contrast}} - MS_{\text{error}}}{SS_{\text{total}} - MS_{\text{error}}}.$$

Generally epsilon squared and omega squared will only differ slightly (Carroll & Nordholm, 1975). Estimated epsilon squared and estimated omega squared can be negative and negative estimates are typically set equal to zero.

Recall that Bauman et al. (1992) were interested in two contrasts: TA& DRTA versus DRA and DRTA versus TA. Table 6 presents the formulas

TABLE 6
Univariate Proportion of Variance Effect Sizes for Contrasts

	Contrast	
Proportion of variance	$\frac{1}{2}(\mu_{\text{DRTA}} + \mu_{\text{TA}}) - \mu_{\text{DRA}}$	$(\mu_{\text{DRTA}} - \mu_{\text{TA}})$
$\hat{\eta}^2 = \frac{\text{SS}_{\text{contrast}}}{\text{SS}_{\text{total}}}$	$\frac{64.121}{748.621} = .086$	$\frac{44.000}{748.621} = .059$
$\hat{\varepsilon}^2 = \frac{SS_{\text{contrast}} - MS_{\text{error}}}{SS_{\text{total}}}$	$\frac{64.121 - 10.167}{748.621} = .072$	$\frac{44.000 - 10.167}{748.621} = .045$
$\hat{\omega}^2 = \frac{SS_{\text{contrast}} - MS_{\text{error}}}{SS_{\text{total}} + MS_{\text{error}}}$	$\frac{64.121 - 10.167}{748.621 + 10.167} = .071$	$\frac{44.000 - 10.167}{748.621 + 10.167} = .045$

Note. $SS_{\text{contrast}} = \dfrac{(c_1\overline{Y}_1 + \cdots + c_j\overline{Y}_j)^2}{c_1^2/n_1 + \cdots + c_J^2/n_J}$.

and estimates for the Bauman et al. study. All three measures of association provide similar estimates of effect for both contrasts. While there are few guidelines for interpreting the numeric values for these effect size measures, Cohen (1988, pp. 280–287) suggested that values of .01, .06, and .14 be used to indicate small, medium, or large associations between the variables, respectively. As with the standardized mean differences presented earlier, the results presented here might be described as a medium effect size for both contrasts.

Whether a researcher reports the effect size in terms of a mean difference in standard deviation units or the proportion of variance shared is to some degree arbitrary at this point in time. Both effect size measures are scale free. That is, the scale of the dependent variable does not affect the magnitudes of these measures. Therefore standardized effect sizes and proportions of variance can be compared for different dependent variables and different contrasts within the same design. If the issues of comparability are taken into account standardized mean differences can be compared across different designs. Proportion of variance effect sizes may not be comparable across different designs. We will return to this issue in more detail when we introduce proportion of variance effect sizes for omnibus tests in multifactor designs. Some might prefer the proportion of variance shared because there are defined limits for this index, 0 to 1, and Rosenthal (1994) believes that this index is easier to interpret than is a standardized mean difference. On the other hand Rosenthal and Rubin (1979) suggested that researchers may not recognize the meaningfulness of an effect size reported in terms of proportion of variance explained, particularly when the proportion of variance is near zero. Saying that 6% of the variance in an outcome is explained by the grouping variable indicates that 94% of the variance is not explained. For many 6% appears trivial rather than meaningful as is implied by Cohen’s description of the effect as a medium effect. On the other hand, a standardized mean difference has no numerical bounds and can only have meaning relative to other standardized mean differences. A standardized mean difference can be transformed into the scale of any measure that is meaningful to a researcher, an advantage not enjoyed by a proportion of variance effect size. For example, if two treatments resulted in a .5 standardized mean difference on the verbal SAT, the difference is a 50-point difference on the SAT scale. A researcher might then judge the meaning of such a difference for practical purposes.

Proportion of variance effect sizes can also be used when covariates are incorporated into the design. Bauman et al. (1992) used an analysis of covariance, with the EDT and Strategies pretests as covariates, to test the contrasts of interest. The adjusted means for the EDT posttest were 8.22, 9.81, and 6.19 for the TA, DRTA, and DRA groups, respectively. Proportion of variance effect sizes are reported in Table 7.

TABLE 7
Univariate Proportion of Variance Effect Sizes for Contrasts: Analysis of Covariance

	Contrast	
Proportion of variance	$\frac{1}{2}(\mu_{\mathrm{DRTA}} + \mu_{\mathrm{TA}}) - \mu_{\mathrm{DRA}}$	$(\mu_{\mathrm{DRTA}} - \mu_{\mathrm{TA}})$
$\hat{\eta}^2 = \frac{SS_{\mathrm{contrast}}}{SS_{\mathrm{total}}}$	$\frac{113.97}{748.62} = .152$	$\frac{27.72}{748.62} = .037$
$\hat{\omega}^2 = \frac{SS_{\mathrm{contrast}} - MS_{\mathrm{error}}}{SS_{\mathrm{total}} + MS_{\mathrm{error}}}$	$\frac{113.97 - 5.66}{748.62 + 5.66} = .144$	$\frac{27.72 - 5.66}{748.621 + 5.66} = .029$

Multivariate approach. Eta squared and omega squared for the two contrasts of interest can be computed as a function of Wilks's lambda and are reported in Table 8. The N in these formulas is the total number of observations on which the means in the contrast are computed. In the present example the complex contrast involved three groups of 22 participants each ($N = 66$). The pairwise contrast involved two groups of 22 participants ($N = 44$). The multivariate eta and omega squared statistics are slightly greater than the univariate measures of association reported in Table 2. Eta squared is also larger than omega squared. Tatsuoka (1988, p. 97) reported that both statistics overestimate the association in the population and offered an adjustment using the number of outcome measures, the degrees of freedom, and the sample size. Formulas and applications for the adjusted eta and omega squared statistics are presented in Table 8. The adjusted multivariate omega squared statistic for the pairwise contrast was negative and would be interpreted as indicating no relationship between group membership and the vector outcome.

Between-Subjects Univariate Designs

Single-factor designs. Even when researchers plan to test specific contrasts, many begin by testing the more general omnibus hypothesis of equality of several population means. To illustrate, suppose before the contrasts on the EDT posttest were tested, an analysis of variance F test was used to test the hypothesis that there are no differences in the population EDT posttest means for the three treatments. Using the data in Table 1, there is sufficient evidence to reject this hypothesis at the .05 level of significance ($F(2, 63) = 5.32$, $p \doteq .008$). For an omnibus hypothesis the use of a standardized mean difference is no longer meaningful because the omnibus test simultaneously examines all possible contrasts. Instead a proportion of variance effect size can be computed to provide an index for the magnitude of the omnibus effect.

TABLE 8
Multivariate Proportion of Variance Effect Sizes for Contrasts

	Contrast	
Proportion of variance	$\frac{1}{2}(\mu_{\text{DRTA}} + \mu_{\text{TA}}) - \mu_{\text{DRA}}$	$(\mu_{\text{DRTA}} - \mu_{\text{TA}})$
$\hat{\eta}^2_{\text{Multi}} = 1 - \Lambda$	$1 - .9085 = .092$	$1 - .9357 = .064$
${}_{\text{adj}}\hat{\eta}^2_{\text{Multi}} = \hat{\eta}^2_{\text{Multi}} - \frac{(p^2 + q^2)(1 - \hat{\eta}^2_{\text{Multi}})}{3N}$	$.092 - \frac{(2^2 + 1^2)(1 - .092)}{3(66)} = .069$	$.064 - \frac{(2^2 + 1^2)(1 - .064)}{3(44)} = .029$
$\hat{\omega}^2_{\text{Multi}} = 1 - \frac{N\Lambda}{df_{\text{error}} + \Lambda}$	$1 - \frac{66(.9085)}{63 + .9085} = .062$	$1 - \frac{66(.9357)}{63 + .9357} = .034$
${}_{\text{adj}}\hat{\omega}^2_{\text{Multi}} = \hat{\omega}^2_{\text{Multi}} - \frac{(p^2 + q^2)(1 - \hat{\eta}^2_{\text{Multi}})}{3N}$	$.062 - \frac{(2^2 + 1^2)(1 - .062)}{3(66)} = .038$	$.034 - \frac{(2^2 + 1^2)(1 - .034)}{3(44)} = -.004$

Note. Λ = Wilks's lambda; J is the number of groups in the design; p is the number of dependent variables; q is the number of degrees of freedom for the hypothesis and equals 1 when a single contrast is the focus of interest; $df_{\text{error}} = N - J$.

In selecting a proportion of variance effect size a distinction needs to be made between fixed and random factors. A factor is fixed if the levels of the independent variable were specifically chosen by the researcher for study, or if the levels included in the study exhaust all possible levels of the independent variable (e.g., males and females exhaust the levels of a gender factor). With a fixed factor, eta squared, epsilon squared, or omega squared can be used as an effect size. The following formulas can be used to compute these proportions of variance for any effect in any design that includes only between-subjects factors:

$$\hat{\eta}^2 = \frac{SS_{\text{effect}}}{SS_{\text{total}}},$$

$$\hat{\varepsilon}^2 = \frac{df_{\text{effect}}(MS_{\text{effect}} - MS_{\text{error}})}{SS_{\text{total}}},$$

and

$$\hat{\omega}^2 = \frac{df_{\text{effect}}(MS_{\text{effect}} - MS_{\text{error}})}{SS_{\text{total}} + MS_{\text{error}}}.$$

A factor is considered random if the levels of the factor were selected using a random process from an infinite number of possible levels of the variable and the researcher wants to generalize the findings to levels of the independent variable not included in the specific investigation. With a random factor the purpose of the research is generally to estimate the variation in the means of the levels of the independent variable rather than to compare specific levels of the variable. An example of a random factor is test forms constructed by randomly selecting items from a large pool of test items. Variation in mean test scores associated with the test forms might be investigated. With a random factor, the appropriate measure of association is the intraclass correlation coefficient.

In the Bauman et al. (1992) study, the researchers specifically chose the three types of intervention to investigate, so the type of intervention would be a fixed factor. Table 9 presents eta squared, epsilon squared, and omega squared equaling .144, .117, and .116, respectively. As noted previously, eta squared is positively biased. Using Cohen's guidelines, these results indicate a large effect associated with the interventions. Levin (1967) pointed out that effect sizes associated with the omnibus test can be misleading because a single discrepant intervention can lead to a large effect size measure. Levin's point is demonstrated here where a large effect size is indicated for the omnibus test but the effect size reported for the contrasts that were of particular interest were judged to be of medium magnitude. It might also be pointed

TABLE 9
Proportion of Variance Effect Sizes for a Single-Factor Omnibus Test

Formula	Example
Fixed effect	
$\hat{\eta}^2 = \dfrac{SS_{\text{effect}}}{SS_{\text{total}}}$	$\dfrac{88.122}{748.621} = .144$
$\hat{\varepsilon}^2 = \dfrac{df_{\text{effect}}(MS_{\text{effect}} - MS_{\text{error}})}{SS_{\text{total}}}$	$\dfrac{2(54.061 - 10.167)}{748.621} = .117$
$\hat{\omega}^2 = \dfrac{df_{\text{effect}}(MS_{\text{effect}} - MS_{\text{error}})}{SS_{\text{total}} + MS_{\text{error}}}$	$\dfrac{2(54.061 - 10.167)}{748.621 + 10.167} = .116$
Random effect	
$\hat{\rho}^2 = \dfrac{J(MS_{\text{effect}} - MS_{\text{error}})}{SS_{\text{total}} + MS_{\text{effect}}}$	$\dfrac{3(54.061 - 10.167)}{748.621 + 54.061} = .164$

Note. *J* is the number of groups in the design.

out that the sum of the eta squared effect sizes for the orthogonal contrasts equals the value of eta squared for the omnibus test.

When the independent variable is a random factor the intraclass correlation may be used to estimate the strength of the association. The lower section of Table 9 provides the formula and application. Assuming the three levels of the Bauman et al. (1992) study were randomly selected from a much larger list of interventions, the effect size would be estimated to equal .164. Cohen only considered the fixed effect model, and guidelines for interpreting a random effect have not been suggested.

Multifactor designs. Bauman et al. (1992) investigated the effect of a single independent variable. In many research studies, two or more independent variables are studied simultaneously to investigate their effects as well as the interaction between the variables. Here we will only consider a two-factor design but the formulas generalize to incorporate multiple factors. To demonstrate measures of association for a factorial design we use the data introduced in the section on standardized mean differences for multifactor between-subjects designs: Pretest EDT scores were used to divide participants in each treatment into two subgroups. The means and standard deviations are presented in Table 3.

In a multifactor design the factors may all be fixed, may all be random, or may be a mixture of fixed and random factors. Formulas for calculating proportion of variance effect sizes depend on the nature of the factors in the design. In addition a distinction is made between the proportion of total variance that was discussed for single-factor designs and a proportion of partial

TABLE 10
Proportion of Variance Effect Sizes for Omnibus Test in a Fixed Effects Factorial Design

Formula	Example for treatment factor
Proportion of total variance	
$\hat{\eta}^2 = \dfrac{SS_{\text{effect}}}{SS_{\text{total}}}$	$\dfrac{88.122}{748.621} = .144$
$\hat{\varepsilon}^2 = \dfrac{df_{\text{effect}}(MS_{\text{effect}} - MS_{\text{error}})}{SS_{\text{total}}}$	$\dfrac{2(54.061 - 7.188)}{748.621} = .125$
$\hat{\omega}^2 = \dfrac{df_{\text{effect}}(MS_{\text{effect}} - MS_{\text{error}})}{SS_{\text{total}} + MS_{\text{error}}}$	$\dfrac{2(54.061 - 7.188)}{748.621 + 7.188} = .124$
Proportion of partial variance	
$\hat{\eta}^2 = \dfrac{SS_{\text{effect}}}{SS_{\text{effect}} + SS_{\text{error}}}$	$\dfrac{88.122}{108.121 + 431.273} = .200$
$\hat{\varepsilon}^2 = \dfrac{df_{\text{effect}}(MS_{\text{effect}} - MS_{\text{error}})}{SS_{\text{effect}} + SS_{\text{error}}}$	$\dfrac{2(54.061 - 7.188)}{108.121 + 431.273} = .177$
$\hat{\omega}^2 = \dfrac{df_{\text{effect}}(MS_{\text{effect}} - MS_{\text{error}})}{df_{\text{effect}}MS_{\text{effect}} + (N - df_{\text{effect}})MS_{\text{error}}}$	$\dfrac{2(54.061 - 7.188)}{108.121 + (66 - 2)7.188} = .165$

Note. *J* and *K* are the number of levels in the A and B factors, respectively.

variance. In a proportion of total variance, the variance due to the effect of interest is expressed as a proportion of the sum of the error variance and the effect variances of all factors (i.e., the total variance) in the design. In a proportion of partial variance, the variance due to the effect of interest is expressed as a proportion of the sum of the error variance and the effect (of interest) variance (i.e., the partial variance). The total variation is influenced by all other factors in a design and therefore proportions of total variance are not comparable across studies that incorporate different factors. A difficulty with the proportions of partial variance is that the effects of different factors within the same study cannot be compared because they do not share a common reference (denominator) point. Because different reference points are used the sum of these partial measures of effect can be greater than one even when factors are orthogonal and therefore the concept of a proportion has a diminished meaning.

Using the data in Table 3 and a fixed effects model the interaction effect is not statistically significant at the .05 level ($F(2, 60) = 2.03$, $p \doteq .140$). The main effect for Treatments (T) and Ability (A) are each statistically significant at the .05 level [$F_{\text{T}}(2, 60) = 7.52$, $p \doteq .001$, and $F_{\text{A}}(1, 60) = 25.05$, $p \doteq .000$, respectively]. The formulas for the proportion of total variance are the same as those presented for a single-factor design. Table 10 provides an application of the formulas to the Treatment factor. In the present example the total sum of squares did not change when the second factor was created. Consequently, eta squared for the treatment factor is equal for the single-factor and two-factor designs. Generally, in completely randomized designs as factors are added the total sum of squares increases. When the second factor was created in the present example, the mean square error was reduced because the participants were more homogeneously grouped. With the reduction in the error variance epsilon squared and omega squared increased to .125 and .124, respectively. The increased measure of effect for the blocking design compared to the single-factor design demonstrates a potential hazard of comparing proportion of total variance effect sizes across studies that are based on different research designs.

Table 10 also reports proportions of partial variance effect sizes. The term *partial* refers to the fact that the other factors in the design are controlled by excluding them from the computation of the effect size. Only the treatment and error variations are used to compute the partial effects. Generally speaking, a proportion of partial variance for an effect will be larger than the proportion of total variance for the same effect. Here the proportions of partial variance as measured by eta squared, epsilon squared, and omega squared are .200, .177, and .165; each is larger than the corresponding proportion of total variance: .144, .125, and .124. The difference between the proportion of total variance and the proportion of partial variance depends on the other factors in the design. If additional factors have large effects, the proportion of total variance can be substantially smaller than corresponding proportion of partial variance. The partial effect size measures are also larger than corresponding proportion of total variance for the single-factor design presented earlier. Because the blocking variable (pretest EDT) reduced the within-cell variance the measure of partial effect size is not comparable to the effect size for the single-factor design. This again demonstrates a potential difficulty of comparing effect size measures across different designs. The issues of comparability described in this section are similar to those described in the section on standardized mean differences and are amenable to the same solutions. However, because proportion of variance formulas are well established in the literature, we have not developed these solutions here, but rather point out comparability problems.

Dodd and Schultz (1973) provided the formulas for computing the omega squared proportion of total variance for designs in which the two factors are random or one factor is fixed while the other is random (mixed). Additional formulas for eta squared are not needed because as noted earlier eta squared is a descriptive statistic that is relevant only for the current sample and not for inferences to the population. Consequently, the distinction between fixed and random effects is not relevant when eta squared is computed. Tables 11 and 12 present the formulas and applications of these formulas for computing

TABLE 11
Proportion of Variance Effect Sizes for Omnibus Tests in Designs with Two Random Factors

Effect	Formula	Example
	Proportion of total variance	
A	$\hat{\omega}^2 = \dfrac{J(MS_{\mathrm{A}} - MS_{\mathrm{AB}})}{SS_{\mathrm{total}} + MS_{\mathrm{A}} + MS_{\mathrm{B}} - MS_{\mathrm{AB}}}$	$\dfrac{3(54.061 - 14.601)}{748.621 + 54.061 + 180.015 - 14.601} = .122$
B	$\hat{\omega}^2 = \dfrac{K(MS_{\mathrm{B}} - MS_{\mathrm{AB}})}{SS_{\mathrm{total}} + MS_{\mathrm{A}} + MS_{\mathrm{B}} - MS_{\mathrm{AB}}}$	$\dfrac{2(180.02 - 14.601)}{748.621 + 54.061 + 180.015 - 14.601} = .342$
AB	$\hat{\omega}^2 = \dfrac{JK(MS_{\mathrm{AB}} - MS_{\mathrm{error}})}{SS_{\mathrm{total}} + MS_{\mathrm{A}} + MS_{\mathrm{B}} - MS_{\mathrm{AB}}}$	$\dfrac{3(2)(14.601 - 7.188)}{748.621 + 54.061 + 180.015 - 14.601} = .046$
	Proportion of partial variance	
A	$\hat{\omega}^2 = \dfrac{MS_{\mathrm{A}} - MS_{\mathrm{AB}}}{MS_{\mathrm{A}} + nK \times MS_{\mathrm{error}} - MS_{\mathrm{AB}}}$	$\dfrac{(54.061 - 14.606)}{54.061 + 2(11)7.187 - 14.606} = .200$
B	$\hat{\omega}^2 = \dfrac{MS_{\mathrm{B}} - MS_{\mathrm{AB}}}{MS_{\mathrm{B}} + nJ \times MS_{\mathrm{error}} - MS_{\mathrm{AB}}}$	$\dfrac{(180.015 - 14.606)}{180.015 + 3(11)7.187 - 14.606} = .411$
AB	$\hat{\omega}^2 = \dfrac{MS_{\mathrm{AB}} - MS_{\mathrm{error}}}{MS_{\mathrm{AB}} + (n - 1)MS_{\mathrm{error}}}$	$\dfrac{(14.606 - 7.188)}{14.606 + (11 - 1)7.187} = .086$

Note. J and K are the number of levels in the A and B factors, respectively.

the proportions of total and partial variance. In all cases the proportion of partial variance values are greater than the proportion of total variance. Comparing corresponding results in Table 10 (all factors fixed) and Tables 11 and 12 (one or both factors random) demonstrates the importance of recognizing the true nature (random or fixed) of the factors being investigated. The use of an inappropriate formula can substantially change the numerical value of the effect size measure.

For designs involving more than one random independent variable the formulas provided by Vaughan and Corballis (1969) or Dodd and Schultz (1973) would be useful or the appropriate statistics may be derived using the expected mean squares reported in several popular intermediate statistics textbooks (e.g., Keppel, 1991; Kirk, 1982; Maxwell & Delaney, 1990).

Single-factor designs with covariates. Using the EDT and Strategies pretests as covariates and Posttest EDT as the outcome measure the omnibus test was conducted. First, however, the test of the interaction between the pretest measures and the intervention factor (homogeneity of regression slopes) must be examined. Neither pretest measure interacted with the intervention variable ($F_{\mathrm{EDT}}(2, 57) = 2.15$, $p \doteq .126$; $F_{\mathrm{Strategy}}(2, 57) = .10$, $p \doteq .906$). The omnibus test was statistically significant at the .05 level (F(2,

TABLE 12
Proportion of Variance Effect Sizes for Omnibus Tests in Designs with A Fixed and B Random

Effect	Formula	Example
	Proportion of total variance	
A	$\hat{\omega}^2 = \dfrac{(J - 1)(MS_{\text{A}} - MS_{\text{AB}})}{SS_{\text{total}} + MS_{\text{B}} + K \times MS_{\text{AB}} - K \times MS_{\text{error}}}$	$\dfrac{2(54.061 - 14.601)}{748.621 + 180.015 + (2)14.601 - (2)7.188} = .084$
B	$\hat{\omega}^2 = \dfrac{K(MS_{\text{B}} - MS_{\text{error}})}{SS_{\text{total}} + MS_{\text{B}} + K \times MS_{\text{AB}} - K \times MS_{\text{error}}}$	$\dfrac{2(180.015 - 7.188)}{748.621 + 180.015 + (2)14.601 - (2)7.188} = .366$
AB	$\hat{\omega}^2 = \dfrac{JK(MS_{\text{AB}} - MS_{\text{error}})}{SS_{\text{total}} + MS_{\text{B}} + K \times MS_{\text{AB}} - K \times MS_{\text{error}}}$	$\dfrac{(3)(2)(14.601 - 7.188)}{748.621 + 180.015 + (2)14.601 - (2)7.188} = .047$
	Proportion of partial variance	
A	$\hat{\omega}^2 = \dfrac{(J - 1)(MS_{\text{A}} - MS_{\text{AB}})}{(J - 1)\, MS_{\text{A}} + (JK - (J - 1))MS_{\text{AB}} - JK \times MS_{\text{error}}}$	$\dfrac{2(54.061 - 14.606)}{(2)54.061 + ((3)(2) - 2)14.606 - (3)(2)7.188} = .639$
B	$\hat{\omega}^2 = \dfrac{MS_{\text{B}} - MS_{\text{error}}}{MS_{\text{B}} + (Jn - 1)MS_{\text{error}}}$	$\dfrac{180.015 - 7.187}{180.015 + ((3)(11) - 1)7.187} = .422$
AB	$\hat{\omega}^2 = \dfrac{MS_{\text{AB}} - MS_{\text{error}}}{MS_{\text{AB}} + (n - 1)MS_{\text{error}}}$	$\dfrac{14.606 - 7.187}{14.606 + (11 - 1)7.187} = .086$

TABLE 13
Proportion of Variance Effect Sizes for Omnibus Tests in a Single-Factor Design: Analysis of Covariance

Formula	Example
$\hat{\eta}^2 = \dfrac{SS_{\text{effect}}}{SS_{\text{total}}}$	$\dfrac{142.26}{748.621} = .191$
$\hat{\varepsilon}^2 = \dfrac{df_{\text{effect}}(MS_{\text{effect}} - MS_{\text{error}})}{SS_{\text{total}}}$	$\dfrac{2(71.630 - 5.659)}{748.621} = .176$
$\hat{\omega}^2 = \dfrac{df_{\text{effect}}(MS_{\text{effect}} - MS_{\text{error}})}{SS_{\text{total}} + MS_{\text{error}}}$	$\dfrac{2(71.630 - 5.659)}{748.621 + 5.659} = .175$

61) = 12.66, $p \doteq .000$). Proportion of variance effect sizes are reported in Table 13.

Between-Subjects Multivariate Designs

Single-factor designs. Bauman et al. (1992) considered the EDT and Strategies posttests to measure a single construct, reading comprehension. Consequently, they compared the three interventions with respect to the vector outcomes of the two posttests. Proportion of variance effect size indicators for multivariate analyses analogous to the univariate indices presented above have been recommended and are discussed in this section, as is their application. For the Bauman et al. data, the multivariate test comparing the treatment groups on the posttest EDT and Strategies was statistically significant at the .05 level using the Wilks's lambda criterion. ($F(4, 124) = 2.52$, $p \doteq .045$). The univariate test on the EDT posttest was reported earlier. The univariate test for DRP was not statistically significant at the .05 level ($F(2, 63) = 2.23$, $p \doteq .117$). Formulas and applications of measures of association (Huberty, 1972; Sachdeva, 1973; Smith, 1972; Stevens, 1972) for omnibus multivariate tests are presented in Table 14. The formulas for unadjusted and adjusted eta and omega squared statistics are the same as those reported for the contrasts; of course, Wilks's lambda for the omnibus test is used in place of Λ for the contrasts. The values for the multivariate effect size in Table 14 are almost identical to the measures of association for the fixed effects univariate measures of association reported in Table 9. For these data group membership explained little variation in DRP.

In addition to eta and omega squared three additional multivariate proportion of variance effect sizes have been recommended (Cramer & Nicewander, 1979; Serlin, 1982) based on different multivariate test criteria (Hotelling–Lawley trace and Bartlet–Pillai trace). These alternative multivariate effect size measures are briefly discussed by Huberty (1994, pp.

TABLE 14
Multivariate Proportion of Variance Effect Sizes for Omnibus Tests

Formula	Example
$\hat{\eta}^2_{\text{Multi}} = 1 - \Lambda$	$1 - .855 = .145$
${}_{\text{adj}}\hat{\eta}^2_{\text{Multi}} = \hat{\eta}^2_{\text{Multi}} - \dfrac{(p^2 + q^2)(1 - \hat{\eta}^2_{\text{Multi}})}{3N}$	$.145 - \dfrac{(22 + 22)(1 - .145)}{3(66)} = .110$
$\hat{\omega}^2_{\text{Multi}} = 1 - \dfrac{N\Lambda}{df_{\text{error}} + \Lambda}$	$1 - \dfrac{66(.855)}{(66 - 3) + .855} = .116$
${}_{\text{adj}}\hat{\omega}^2_{\text{Multi}} = \hat{\omega}^2_{\text{Multi}} - \dfrac{(p^2 + q^2)(1 - \hat{\eta}^2_{\text{Multi}})}{3N}$	$.116 - \dfrac{(22 + 22)(1 - .116)}{3(66)} = .080$
$\hat{\tau}^2 = 1 - \Lambda^{1/r}$	$1 - .855^{1/2} = .075$
$\hat{\zeta}^2 = U/(r + U)$	$.169/(2 + .169) = .078$
$\hat{\xi}^2 = V/r$	$.145/2 = .072$

Note. Λ = Wilks's lambda, V = Bartlett–Pillai trace, and U = Hotelling–Lawley trace; J is the number of groups in the design; p is the number of dependent variables; q is the number of degrees of freedom for the hypothesis; $r = \min(p, q)$; $df_{\text{error}} = N - J$.

194–195). The bottom of Table 14 reports the formulas and applications of three of these measures of association, $\hat{\tau}^2$, $\hat{\zeta}^2$, and $\hat{\xi}^2$. All three measures of association are lower than the adjusted eta and omega squared. It is worth noting that these statistics are reported by SPSS in the multivariate GLM program under the title of eta squared when effect sizes are requested.

Multifactor designs. Returning to the two-factor design with Ability as the second factor, the multivariate test of the interaction between the Treatments and Ability for the vector of EDT and DRP posttest variables was not statistically significant at the .05 level ($F(4, 118) = 1.87$, $p \doteq .120$). Both main effects were statistically significant at the .05 level, ($F_{\text{T}}(4, 118) = 3.49$, $p \doteq .011$, and $F_{\text{A}}(2, 59) = 12.34$, $p \doteq .000$). For multivariate factorial designs the same formulas reported in Table 14 would be used the effect size. Table 15 presents the formulas and application of the measures of effect for the treatment factor. The factorial effect sizes for the treatment factor are all a little larger then the single-factor multivariate effect sizes for the treatment factor reported in Table 14 because the error matrix was reduced when the ability factor was added to the design. Because Wilks's lambda is computed as the ratio of the determinant of the error matrix and the determinant of the sum of the hypothesis matrix and the error matrix ($\Lambda = |\text{E}|/|\text{H} + \text{E}|$), the values of the measures of association are partial effect sizes rather than proportions of total variance. In our review we did not find any references to multivariate random effects or mixed model effects.

Designs with covariates. Covariates can be used in a multivariate design, as well as in a univariate design, to enhance power. When the omnibus test is conducted, proportion of variance effect sizes can be used to describe the

TABLE 15
Multivariate Proportion of Variance Effect Sizes Applied to the Treatment Effect in Two-Factor Design

Formula	Example
$\hat{\eta}^2_{\text{Multi}} = 1 - \Lambda$	$1 - .799 = .201$
${}_{\text{adj}}\hat{\eta}^2_{\text{Multi}} = \hat{\eta}^2_{\text{Multi}} - \dfrac{(p^2 + q^2)(1 - \hat{\eta}^2_{\text{Multi}})}{3N}$	$.201 - \dfrac{(22 + 42)(1 - .201)}{3(66)} = .120$
$\hat{\omega}^2_{\text{Multi}} = 1 - \dfrac{N\Lambda}{df_{\text{error}} + \Lambda}$	$1 - \dfrac{66(.799)}{(66 - 6) + .799} = .133$
${}_{\text{adj}}\hat{\omega}^2_{\text{Multi}} = \hat{\omega}^2_{\text{Multi}} - \dfrac{(p^2 + q^2)(1 - \hat{\eta}^2_{\text{Multi}})}{3N}$	$.173 - \dfrac{(22 + 42)(1 - .173)}{3(66)} = .089$
$\hat{\tau}^2 = 1 - \Lambda^{1/r}$	$1 - .799^{1/2} = .106$
$\hat{\zeta}^2 = U/(r + U)$	$.251/(2 + .251) = .112$
$\hat{\xi}^2 = V/r$	$.201/2 = .101$

Note. Λ = Wilks's lambda, V = Bartlett–Pillai trace, and U = Hotelling–Lawley trace; J is the number of groups in the design; p is the number of dependent variables; q is the number of degrees of freedom for the hypothesis; $r = \min(p, 9)$; $df_{\text{error}} = N - JK$.

strength of the multivariate omnibus effect. Here we only demonstrate the estimation of effect size for a single-factor univariate and multivariate analysis of covariance but applications to factorial designs are also possible. The multivariate omnibus test was statistically significant at the .05 level ($F(4, 120) = 5.75$, $p \doteq .000$). Table 16 presents the formulas and application for the effect size measures. A comparison of these effect size measures with those reported in Table 14 without the covariates again shows that the effect sizes reported with the covariates are now twice the effect size obtained when the covariates are not used. These results demonstrate that making comparisons between studies that differ in their design and use of control

TABLE 16
Multivariate Proportion of Variance Effect Sizes: Analysis of Covariance

Formula	Example
$\hat{\eta}^2_{\text{Multi}} = 1 - \Lambda$	$1 - .704 = .296$
$\hat{\omega}^2_{\text{Multi}} = 1 - \dfrac{N\Lambda}{df_{\text{error}} + \Lambda}$	$1 - \dfrac{66(.704)}{(66 - 5) + .704} = .247$
$\hat{\tau}^2 = 1 - \Lambda^{1/r}$	$1 - .704^{1/2} = .161$
$\hat{\zeta}^2 = U/(r + U)$	$.420/(2 + .420) = .174$
$\hat{\xi}^2 = V/r$	$.296/2 = .148$

factors can be very misleading. They also raise some question regarding the appropriateness of generalizing Cohen's guidelines for interpreting effect size measures across all different analysis models and research designs. These issues are discussed further under Cautionary Notes below.

Within-Subjects Designs

Dodd and Schultz (1973) provide the computational formulas for computing the omega squared proportion of total variance for repeated measures designs. Here we also consider the partial effect size measures derived using the expected mean squares for a single within-subjects factor reported by Kirk (1982, p. 245) assuming measures are fixed, subjects are random, and there is no subjects by measures interaction. The repeated measures design can be thought of as a two-factor design (measures and subjects) with one observation per cell. The repeated measure factor is fixed and the subject factor is random. With one observation per cell there is no estimate of within-cell variance and the interaction of subjects and measures is used as the error term. To estimate omega squared, the formulas reported in Table 12 apply, with Factor A representing the Measures factor and Factor B representing the Subjects factor. Because $MS_{\mathrm{AB}} = MS_{\mathrm{error}}$, the MS_{AB} term drops out of the formulas. The resulting formulas are reported in Table 17 and applied to the one-factor within-subjects design with pretest and posttest EDT as the levels of the factor.

Computing a proportion of total variance effect size for the within-subjects factors is a little more difficult than for between-subjects designs. When SPSS is used to analyze data from a within-subjects design, SPSS does not print the SS_{total}. With SAS, MS_{subjects} and SS_{total} are not printed when the Repeated statement is used. With each package the missing quantities will be printed if the data are organized as if the within-subjects factors had been between-subjects factors and the data are analyzed as if all factors had been between-subjects factors.

The proportion of partial variance effect size is provided in SPSS and can be computed by using the output reported in SAS using the Repeated statement. Table 17 presents these results. Eta squared using lambda is identical to the partial eta squared computed using the sum of squares. Omega squared computed using lambda is larger than the partial omega squared computed using the univariate mean squares.

Split-Plot Design

In the Bauman et al. (1992) study participants were randomly assigned to treatment groups and each participant was assessed before and after the treatments. The data can be analyzed by using a split-plot analysis of variance. The interaction between group and measures was statistically significant at the .05 level ($F(2, 63) = 12.64$, $p \doteq .000$). This result is not surprising

TABLE 17
Proportion of Variance Effect Sizes for a Within-Subjects Design

Formula	Example
Proportion of total variance	
$\hat{\eta}^2 = \frac{SS_{\text{effect}}}{SS_{\text{total}}}$	$\frac{96.735}{1438.386} = .067$
$\hat{\varepsilon}^2 = \frac{df_{\text{effect}}(MS_{\text{effect}} - MS_{\text{error}})}{SS_{\text{total}}}$	$\frac{1(96.735 - 4.519)}{1438.386} = .064$
$\hat{\omega}^2 = \frac{df_{\text{effect}}(MS_{\text{effect}} - MS_{\text{error}})}{SS_{\text{total}} + MS_{\text{subjects}}}$	$\frac{1(96.735 - 4.519)}{1438.386 + 16.121} = .063$
Proportion of partial variance	
$\hat{\eta}^2 = \frac{SS_{\text{effect}}}{df_{\text{effect}}MS_{\text{effect}} + SS_{\text{error}}}$	$\frac{(1)96.735}{(1)96.735 + 293.765} = .248$
$\hat{\varepsilon}^2 = \frac{df_{\text{effect}}(MS_{\text{effect}} - MS_{\text{error}})}{df_{\text{effect}}MS_{\text{effect}} + SS_{\text{error}}}$	$\frac{1(96.735 - 4.519)}{(1)96.735 + 293.765} = .236$
$\hat{\omega}^2 = \frac{df_{\text{effect}}(MS_{\text{effect}} - MS_{\text{error}})}{df_{\text{effect}}MS_{\text{effect}} + (N - df_{\text{effect}})MS_{\text{error}}}$	$\frac{1(96.735 - 4.519)}{1(96.735) + (66 - 1)4.519} = .236$
$\hat{\eta}^2_{\text{Multi}} = 1 - \Lambda$	$1 - .752 = .248$
$\hat{\omega}^2_{\text{Multi}} = 1 - \frac{N\Lambda}{df_{\text{error}} + \Lambda}$	$1 - \frac{66(.752)}{65 + .752} = .245$

because individuals were randomly assigned to the groups so only trivial differences were expected on the pretest. Differences due to the treatments on the posttest were expected. Given this result the main effects for measures and treatments would typically be of little interest but are presented here only for completeness. The main effect for measures was statistically significant at the .05 level ($F_M(1, 63) = 29.07$, $p \doteq .000$) but the main effect for treatments did not meet the statistical criteria at the .05 level ($F_G(2, 63) = 1.40$, $p \doteq .255$). Gaebelein and Soderquist (1978) provide the formulas for the variance components for the split-plot design assuming both the repeated measures factor and the between subjects factor are fixed. Table 18 presents the formulas and application of the total and partial proportion of variance effect size measures. In this table factor A is the between-subjects factor and has J levels. Factor B is the within-subjects factor and has K levels. The eta squared effect sizes are computed as described previously. However, the partial eta squared measures of effect size for the between-subjects and the within-subjects factors have different denominators because these effects have different error terms and therefore are not comparable. SPSS provides the partial eta squared statistics for both the between- and the within-subjects factors as an option. Unlike with the within-subjects design described above, when a between subjects factor is included, SAS provides all the information needed to compute the omega squared proportions of total and partial variance. For the present example the partial omega squared value is much smaller than partial eta squared for effects involving the within-subjects factor. Eta squared computed using Wilks's lambda equals the partial eta squared computed using the univariate sum of squares. Computing the partial omega squared statistic using Wilks's lambda results in an estimate of the effect similar to partial eta squared and considerably larger than the effect estimated using the univariate mean squares.

CAUTIONARY NOTES

While many have advocated the reporting one or more measures of effect along with tests of statistical significance, it should be noted that these measures of effect are not without their critics. Effect size measures have been offered as indices of practical significance or meaningfulness. But a basic question raised by some critics is whether these measures actually contribute to a better understanding of the study's results. How meaningful is it to a policy maker to learn that a treatment explains 10% of the variance in an outcome measure or that there is a half standard deviation difference between a treatment group and a control group? What practitioners want to know are answers to such questions as: ''What can the participants of the treatment do because of the intervention that the control group cannot do?'' Such questions are validity issues that depend on the meaning of the measures used (Keren & Lewis, 1979; Porter, Schmidt, Floden, & Freeman, 1978; Yeaton & Sechrest, 1981). Statistical indices of effect are computed, as demonstrated above, completely independent of the meaning of the measures used. At best these statistical indices are relative measures of effect, not absolute measures.

Standards used to evaluate the effect size measures have also been questioned. Cohen (1988) suggested .2σ, .5σ, and .8σ as small, medium, and large effects as a starting point for interpreting the standardized mean differences and 1%, 6%, and 14% as guidelines for interpreting measures of association. But there is little empirical justification for these standards. Feldt (1973), for example, considered a change in average test performance on the Iowa Test of Basic Skills (ITBS) from the 50th to the 75th percentile to be a large improvement. He examined six subtests of the ITBS and found that such an improvement in Grades 3, 5, 7, and 8 would require an increase performance of between .246 and .372 standard deviations. For the Iowa Test of Basic Skills, then, a ''large'' effect would be considerably smaller than that suggested by Cohen. Cohen recognized this limitation and encour-

TABLE 18
Proportion of Variance Effect Sizes for a Split-Plot Design

Effect	Variance	Formula	Example
Between subjects (A)	Total	$\hat{\eta}^2 = \dfrac{SS_{\text{A}}}{SS_{\text{total}}}$	$\dfrac{44.591}{1438.386} = .031$
		$\hat{\omega}^2 = \dfrac{df_{\text{A}}(MS_{\text{A}} - MS_{\text{S/A}})}{SS_{\text{total}} + MS_{\text{S/A}} + N \times MS_{\text{BS/A}}}$	$\dfrac{2(22.295 - 15.925)}{1438.386 + 15.925 + (66)3.328} = .008$
	Partial	$\hat{\eta}^2 = \dfrac{SS_{\text{A}}}{SS_{\text{A}} + SS_{\text{S/A}}}$	$\dfrac{44.591}{44.591 + 1003.295} = .043$
		$\hat{\omega}^2 = \dfrac{df_{\text{A}}(MS_{\text{A}} - MS_{\text{S/A}})}{df_{\text{A}}MS_{\text{A}} + (N - df_{\text{A}})MS_{\text{S/A}} - N \times MS_{\text{BS/A}}}$	$\dfrac{2(22.295 - 15.925)}{2(22.295) + (66 - 2)15.925 - (66)3.328} = .015$
Within subjects (B)	Total	$\hat{\eta}^2 = \dfrac{SS_{\text{B}}}{SS_{\text{total}}}$	$\dfrac{96.735}{1438.386} = .067$
		$\hat{\omega}^2 = \dfrac{df_{\text{B}}(MS_{\text{B}} - MS_{\text{BS/A}})}{SS_{\text{total}} + MS_{\text{S/A}} + N - MS_{\text{BS/A}}}$	$\dfrac{1(96.73 - 3.378)}{1438.386 + 15.925 + (66)3.328} = .056$

	Partial	$\hat{\eta}^2 = \frac{SS_B}{SS_B + SS_{BS/A}}$	$\frac{96.735}{96.735 + 209.659} = .316$
		$\hat{\omega}^2 = \frac{df_B(MS_B - MS_{BS/A})}{df_B MS_B + (KN \times df_B)MS_{BS/A}}$	$\frac{1(96.735 - 3.328)}{1(96.735) + (2(66) - 2)3.328} = .175$
		$\hat{\eta}^2_{\text{Multi}} = 1 - \Lambda$	$1 - .684 = .316$
		$\hat{\omega}^2_{\text{Multi}} = 1 - \frac{N\Lambda}{df_{\text{error}} + \Lambda}$	$1 - \frac{66(.684)}{(66 - 3) + .684} = .291$
Interaction	Total	$\hat{\eta}^2 = \frac{SS_{AB}}{SS_{\text{total}}}$	$\frac{84.106}{1438.386} = .058$
		$\hat{\omega}^2 = \frac{df_{AB}(MS_{AB} - MS_{BS/A})}{SS_{\text{total}} + MS_{S/A} + N \times MS_{BS/A}}$	$\frac{2(42.053 - 3.328)}{1438.386 + 15.925 + (66)3.328} = .046$
	Partial	$\hat{\eta}^2 = \frac{SS_{AB}}{SS_{AB} + SS_{BS/A}}$	$\frac{84.106}{84.106 + 209.659} = .286$
		$\hat{\omega}^2 = \frac{df_{AB}(MS_{AB} - MS_{BS/A})}{df_{AB} MS_{AB} + (KN - df_{AB})MS_{BS/A}}$	$\frac{2(42.053 - 3.328)}{2(42.053) + (2(66) - 2)3.328} = .150$
		$\hat{\eta}^2_{\text{Multi}} = 1 - \Lambda$	$1 - .714 = .286$
		$\hat{\omega}^2_{\text{Multi}} = 1 - \frac{N\Lambda}{df_{\text{error}} + \Lambda}$	$1 - \frac{66(.714)}{(66 - 3) + .714} = .260$

Note. $MS_{S/A}$ is the between-subjects error mean square; $MS_{BS/A}$ is the within-subjects error mean square.

aged researchers to offer alternative standards but, as yet, none have been offered. Further, one might question whether the standards offered for univariate statistics should also be accepted for multivariate effect size measures. It seems unreasonable to think that a common set of criteria can be set for all studies.

Throughout this paper it was noted that measures of effect size are affected by the research design used. If researchers are not careful, serious errors in computing an effect size (e.g., using equations appropriate for a fixed factor when the factor is random) can result. Effect sizes may not be comparable across different designs and could lead to misinterpretations of the magnitude of the effects observed. That is, the effect size computed from a study that includes one or more individual difference factors (e.g., covariates, blocking variables) will not be comparable to the effect size computed from a study that includes only manipulated variables. As demonstrated here, adjustments can be made if sufficient information is available but this may not always be the case.

Another potential problem with omega squared measures of effect is they were derived from the variance components obtained from the expected mean squares for the sources of variation in the model. The expected mean squares assume a balanced design. In many applied research contexts sample sizes are not equal and are generally disproportionate. Vaughan and Corballis (1969) cautioned against the use of omega squared in these situations. Carroll and Nordholm (1975), however, found that in a single-factor design, unequal n had little effect on the estimation of Kelley's ε^2 or Hay's ω^2 if variances are equal, and with equal n heterogeneous variances had little impact on the estimation of these effect sizes. But unequal n and heterogeneous variances lead to an overestimation of the effect size, and Carroll and Nordholm cautioned against their use in these situations.

Carroll and Nordholm (1975) also showed empirically that the standard errors for both ε^2 and ω^2 can be large when sample sizes are small. Even when the total sample size from three populations equaled 90 the standard errors for these effect size measures were unacceptably large. The authors cautioned researchers against interpreting these effect size measures when they are estimated using small samples.

Levin's (1967) point regarding effect size measures for omnibus tests also merits restating. In most multigroup studies the omnibus hypothesis test is rarely useful for answering research questions of real interest to researchers (Olejnik & Huberty, 1993; Rosnow & Rosenthal, 1988). An omnibus effect size measure may be very misleading if the large effect is produced by a single discrepant treatment mean that may or may not be of primary interest to the researcher. Specific contrasts are often much more meaningful to researchers and effect size measures for these contrasts are likely to be more meaningful and should be encouraged.

Finally, an effect size measure is often interpreted relative to effect size measures reported in previous studies. But several critics have pointed out that measures of association are affected by the reliability of the measures, the heterogeneity of the populations being compared, the specific levels of the variables studied, the strength of the treatments, and the range of treatments (Maxwell et al., 1981; O'Grady, 1982; Sechrest & Yeaton, 1982), thus making such comparisons hazardous. Fern and Monroe (1996) presented an especially comprehensive discussion of these factors. Differences on these factors can lead to misleading comparisons of measures of effect:

1. Low reliability increases the error variance and puts a limit on the amount of variance that can be explained by an explanatory variable. Two measures of effect from two studies of the same explanatory variable can be substantially different if the outcome measures used have substantially different reliabilities.

2. Population heterogeneity can reduce the magnitude of the effect size measure. The effect sizes computed in two studies that differ with respect to the variability of the outcome measure may not be comparable. For example, if one investigation studies high school freshmen while a second study involves high school students from all grade levels, the measures of effect size may not be comparable.

3. In fixed effects models the magnitude to the omnibus measures of effect size depends on the specific levels of the variables studied. If different levels of the explanatory variable are investigated, the measure of effect will not be comparable. This of course is also true with tests of statistical significance as well.

4. The strength of the treatment refers to the likelihood that the treatment will have the intended effect (Sechrest & Yeaton, 1982). A strong treatment would explain more variance than a weak treatment but the quantification of the strength of a treatment is generally unknown. For example Nist and Olejnik (1995) studied the effect of context and dictionary use on varying levels of word knowledge. In this study the quality of the dictionary definitions and the quality of the context in which the target words were provided were varied. The quality of dictionary definitions and the quality of context were simply described as weak or strong but not quantified. The amount of variation in the word knowledge scores explained by both of these factors depended a great deal on the type of cues provided in the context and the clarity of the definitions provided. Sechrest and Yeaton (1982) argue that without knowledge of the strength of a treatment the proportion of variance accounted for is meaningless.

5. The range of treatments included in a study can increase or reduce the proportion of variation explained. For example, if the levels of the treatment variable were narrowly defined (e.g., 10, 20, 30, or 40 min of free reading time), the between-treatment variation in comprehension skills would be

lower than in a study with a greater spread in the levels in free reading times (e.g., 30, 60, 90, or 120 min). The increased variability of the latter design is likely to lead to a greater measure of effect.

Because of the limitations of effect size measures describe above, it is often difficult to make meaningful comparisons of effect sizes across different studies or to interpret them against a common standard. Perhaps one reason that measures of effect have been slowly adopted has been because of these limitations.

Conclusion

Although methodologists have long advocated the use of some type of effect size measure, applied researchers have been slow to incorporate them along with their statistical tests in reporting their research findings. It is hoped that by presenting the formulas for several indices of effect size for a variety of univariate and multivariate tests and their applications, researchers will be encouraged to include these measures when reporting their findings. But increased use alone without recognition of the inherent limitations associated with these indices of effect may not add to a better understanding of research findings. With the limitations in mind, however, it is believed that these measure of effect will prove to be useful and meaningful to researchers and practitioners alike.

APPENDIX: Raw Data from the Bauman, Seifert-Kessell, and Jones Study

TA					DRTA					DRA				
Pre 1	Pre 2	Post 1	Post 2	Post 3	Pre 1	Pre 2	Post 1	Post 2	Post 3	Pre 1	Pre 2	Post 1	Post 2	Post 3
7	6	4	8	43	7	2	7	6	27	4	3	5	4	34
4	6	4	10	34	7	6	5	6	36	6	5	9	5	36
7	2	4	4	45	8	7	13	5	51	9	4	5	3	42
7	6	3	9	39	7	6	8	5	51	8	8	7	7	37
6	5	8	5	40	6	2	7	0	50	9	2	4	4	44
4	6	1	10	27	8	4	10	6	55	10	10	9	10	49
6	4	7	9	46	9	6	8	6	52	8	5	3	3	38
6	6	7	9	39	8	4	10	11	48	11	3	4	5	38
8	2	9	5	31	9	5	12	6	53	8	4	2	3	38
5	3	6	8	39	8	6	7	8	45	7	3	5	4	50
8	3	4	6	40	8	5	8	8	47	9	6	7	8	31
11	5	12	8	52	12	4	13	3	51	12	6	8	5	49
14	6	14	12	53	10	1	5	7	30	16	5	10	9	54
13	6	12	11	53	16	8	14	7	50	15	13	9	8	52
9	5	7	11	41	15	7	14	6	55	14	8	12	5	50
12	3	5	10	41	9	6	10	9	48	12	7	5	5	35
13	9	9	9	46	13	7	12	7	52	12	3	8	7	36
13	8	13	1	52	12	8	11	6	46	13	7	12	4	46
11	7	11	12	55	9	4	8	7	36	12	5	4	6	42
12	3	5	13	36	13	6	10	6	45	12	2	8	8	47
11	4	11	7	50	10	2	11	6	49	12	2	6	4	39
14	4	15	7	54	10	6	12	6	49	12	5	5	5	38

Note. Pre 1 and Post 2 are the EDT pretest and posttest; Pre 2 and Post 2 are the Strategies pretest and posttest; Post 3 is the DRP.

REFERENCES

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

Bauman, J. F., Seifert-Kessell, N., & Jones, L. A. (1992). Effect of think-aloud instruction on elementary students' comprehension monitoring abilities. *Journal of Reading,* **24,** 143–172.

Carroll, R. M., & Nordholm, L. A. (1975). Sampling characteristics of Kelley's ε^2 and Hays' ω^2. *Educational and Psychological Measurement,* **35,** 541–554.

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

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

Cohen, J. (1965). Some statistical issues in psychological research. In B. B. Wolman (Ed.), *Handbook of clinical psychology* (pp. 95–121). New York: Academic Press.

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

Cohen, J. (1988). *Statistical power analysis of the behavioral sciences.* (2nd ed.). New York: Academic Press.

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

Cramer, E. M., & Nicewander, W. A. (1979). Some symmetric invariant measures of multivariate association. *Psychometrika,* **44,** 43–54.

Degrees of reading power. (1986). New York: The College Board.

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

Dunlap, W. P., Cortina, J. M., Vaslow, J. B., & Burke, M. J. (1996). Meta-analysis of experiments with matched groups or repeated measures designs. *Psychological Methods,* **1,** 170–177.

Dwyer, J. H. (1974). Analysis of variance and the magnitude of effects. *Psychological Bulletin,* **81,** 731–737.

Feldt, L. (1973). What size samples for methods/materials experiments? *Journal of Educational Measurement,* **10,** 221–231.

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

Frick, R. W. (1999). Defending the status quo. *Theory & Psychology,* **9,** 183–189.

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

Gaebelein, J. W., & Soderquist, D. R. (1978). The utility of within-subjects variables: Estimates of strength. *Educational and Psychological Measurement,* **38,** 351–360.

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

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., McGraw, B., & Smith, M. L. (1981). *Meta-analysis in social research.* Beverly Hills, CA: Sage.

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

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

Huberty, C. J. (1972). Multivariate indices of strength of association. *Multivariate Behavioral Research,* **7,** 523–528.

Huberty, C. J. (1994). *Applied discriminant analysis.* New York: Wiley.

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

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

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

Keren, G., & Lewis, C. (1979). Partial omega squared for ANOVA designs. *Educational and Psychological Measurement,* **39,** 119–128.

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

Kirk, R. E. (1982). *Experimental design: Procedures for the behavioral sciences* (2nd ed.). Belmont, CA: Wadsworth.

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

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

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

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

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

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

McNamara, J. F. (1978). Practical significance and statistical models. *Educational Administration Quarterly,* **14,** 48–63.

Mitchell, C., & Hartmann, D. P. (1981). A cautionary note on the use of omega squared to evaluate the effectiveness of behavioral treatments. *Behavioral Assessment,* **3,** 93–100.

Morris, S. B., & DeShon, R. P. (1997). Correcting effect sizes computed from factorial analysis of variance for use in meta-analysis. *Psychological Methods,* **2,** 192–199.

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

Nist, S. L., & Olejnik, S. (1995). The role of context and dictionary definitions on varying levels of word knowledge. *Reading Research Quarterly,* **30,** 172–193.

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

Olejnik, S., & Huberty, C. J. (1993, April). *Preliminary statistical tests.* Paper presented at the annual meeting of the American Educational Research Association, Atlanta, GA.

Pearson, K. (1905). *Mathematical contributions to the theory of evolution: XIV. On the general theory of skew correlations and nonlinear regression* (Draper's Company Research Memoirs, Biometric Series II). London: Dulau.

Porter, A. C., Schmidt, W. H., Floden, R. E., & Freeman, D. J. (1978). Practical significance in program evaluation. *American Educational Research Association,* **15,** 529–539.

Richardson, J. T. E. (1996). Measures of effect size. *Behavioral Research Methods, Instruments, & Computers,* **28,** 12–22.

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

Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 231–244). New York: Sage.

Rosenthal, R., & Rubin, D. B. (1979). A note on percent variance explained as a measure of the importance of effects. *Journal of Applied Social Psychology,* **9,** 395–396.

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

Rosnow, R. L., & Rosenthal, R. (1988). Focused tests of significance and effect size estimation in counseling psychology. *Journal of Counseling Psychology,* **35,** 203–208.

Sachdeva, D. (1973). Estimating strength of relationship in multivariate analysis of variance. *Educational and Psychological Measurement,* **33,** 627–631.

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

Serlin, R. C. (1982). A multivariate measure of association based on the Pillai–Bartlett procedure. *Psychological Bulletin,* **91,** 413–417.

Schmidt, F. (1992). What do data really mean? Research findings, meta-analysis, and cumulative knowledge in psychology. *American Psychologist,* **47,** 1173–1181.

Smith, I. L. (1972). The ETA coefficient in MANOVA. *Multivariate Behavioral Research,* **7,** 361–372.

Stevens, J. P. (1972). Global measures of association in multivariate analysis of variance. *Multivariate Behavioral Research,* **7,** 373–378.

Stevens, J. P. (1996). *Applied multivariate statistics for the social sciences.* (3rd ed.). Mahwah, NJ: Erlbaum.

Strube, M. J. (1988). Some comments on the use of magnitude-of-effect estimates. *Journal of Counseling Psychology,* **35,** 342–345.

Tatsuoka, M. M. (1988). *Multivariate analysis: Techniques for educational and psychological research.* New York: Collier Macmillan.

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

Thompson, B. (1997). Editorial policies regarding statistical significance tests: Further comments. *Educational Researcher,* **26,** 29–32.

Thompson, B. (1999a). If statistical significance tests are broken/misused, what practices should supplement or replace them? *Theory & Psychology,* **9,** 165–181.

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

Vaughan, G. M., & Corballis, C. (1969). Beyond tests of significance: Estimating strength of effects in selected ANOVA designs. *Psychological Bulletin,* **72,** 204–213.

Wilkinson, L., & the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist,* **54,** 549–604.

Yeaton, W. H., & Sechrest, L. (1981). Meaningful measures of effect. *Journal of Consulting and Clinical Psychology,* **49,** 766–767.