Psychological Methods
2004, Vol. 9, No. 2, 164–182

1082-989X/04/$12.00 DOI: 10.1037/1082-989X.9.2.164

# Beyond the *F* Test: Effect Size Confidence Intervals and Tests of Close Fit in the Analysis of Variance and Contrast Analysis

James H. Steiger
Vanderbilt University

This article presents confidence interval methods for improving on the standard *F* tests in the balanced, completely between-subjects, fixed-effects analysis of variance. Exact confidence intervals for omnibus effect size measures, such as $\omega^2$ and the root-mean-square standardized effect, provide all the information in the traditional hypothesis test and more. They allow one to test simultaneously whether overall effects are (a) zero (the traditional test), (b) trivial (do not exceed some small value), or (c) nontrivial (definitely exceed some minimal level). For situations in which single-degree-of-freedom contrasts are of primary interest, exact confidence interval methods for contrast effect size measures such as the contrast correlation are also provided.

The analysis of variance (ANOVA) remains one of the most commonly used methods of statistical analysis in the behavioral sciences. Most ANOVAs, especially in exploratory studies, report an omnibus *F* test of the hypothesis that a main effect, interaction, or simple main effect is precisely zero. In recent years, a number of authors (Cohen, 1994; Rosnow & Rosenthal, 1996; Schmidt, 1996; Schmidt & Hunter, 1997; Serlin & Lapsley, 1993; Steiger & Fouladi, 1997) have sharply questioned the efficacy of tests of this “nil” hypothesis. Several of these critiques have concentrated on ways that the nil hypothesis test fails to deliver the information that the typical behavioral scientist wants. However, a number of the articles have also suggested, more or less specifically, replacements for or extensions of the null hypothesis test that would deliver much more useful information.

The suggestions have developed along several closely related lines, including the following:

1. Eliminate the emphasis on omnibus tests, with attention instead on focused contrasts that answer specific research questions, along with calculation of point estimates and approximate confidence interval estimates for some correlational measures of effect size (e.g., Rosenthal, Rosnow, & Rubin, 2000; Rosnow & Rosenthal, 1996).

2. Calculate exact confidence interval estimates of measures of standardized effect size, using an iterative procedure (e.g., Smithson, 2001; Steiger & Fouladi, 1997).

3. Perform tests of a statistical null hypothesis other than that of no difference or zero effect (e.g., Serlin & Lapsley, 1993).

As proponents of the first suggestion, Rosnow and Rosenthal (1996) discussed several types of correlation coefficients that are useful in assessing experimental effects. Their work is particularly valuable in situations in which the researcher has questions that are best addressed by testing single contrasts. Rosnow and Rosenthal emphasized the use of the Pearson correlation, rather than the squared multiple correlation, partly because of concern that the latter tends to present an overly pessimistic picture of the value of “small” experimental effects.

The second suggestion, exact interval estimation, has been gathering momentum since around 1980. The movement to replace hypothesis tests with confidence intervals stems from the fundamental realization that, in many if not most situations, confidence intervals provide more of the information that the scientist is truly interested in. For example, in a two-group experiment, the scientist is more interested in knowing how large the difference between the two groups is (and how precisely it has been determined)

I express my gratitude to Michael W. Browne, Stanley A. Mulaik, the late Jacob Cohen, William W. Rozeboom, Rachel T. Fouladi, Gary H. McClelland, and numerous others who have encouraged this project.

Correspondence concerning this article should be addressed to James H. Steiger, Department of Psychology and Human Development, Box 512 Peabody College, Vanderbilt University, Nashville, TN 37203. E-mail: jsteiger@bellsouth.net

Note: This version includes corrections to the calculations at the bottom of page 173, Equation 51, and the discussion and confidence intervals on page 174, Table 3. Thanks to Professor Kris Kirby of Williams College for alerting me to these errors.

than whether the difference between the groups is exactly zero.

The third suggestion, which might be called tests of *close fit,* has much in common with the approach widely known to biostatisticians as *bioequivalence testing* and is based on the idea that the scientist should not be testing perfect adherence to a point hypothesis but should replace the test of close fit with a “relaxed” test of a more appropriate hypothesis. Tests of close fit share many of their computational aspects with the exact interval estimation approach in terms of the software routines required to compute probability levels, power, and sample size. They remain within the familiar hypothesis-testing framework, while providing important practical and conceptual gains, especially when the experimenter’s goal is to demonstrate that an effect is trivial.

In this article, I present methods that implement; support; and, in some cases, unify and extend major suggestions (1) through (3) discussed above. First I briefly review the history, rationale, and theory behind exact confidence intervals on measures of standardized effect size in ANOVA. I then provide detailed instructions, with examples, and software support for computing these confidence intervals. Next I discuss a general procedure for assessing effects that are represented by one or more contrasts, using correlations. Included is a population rationale, with sampling theory and an exact confidence interval estimation procedure, for one of the correlational measures discussed by Rosnow and Rosenthal (1996).

Although the initial emphasis is on confidence interval estimation, I also discuss how the same technology that generates confidence intervals may be used to test hypotheses of *minimal effect,* thus implementing the *good enough principle* discussed by Serlin and Lapsley (1993).

## Exact Confidence Intervals on Standardized Effect Size

The notion that hypothesis tests of zero effect should be replaced with exact confidence intervals on measures of effect size has been around for quite some time but was somewhat impractical because of its computational demands until about 10 years ago. A general method for constructing the confidence intervals, which Steiger and Fouladi (1997) referred to as *noncentrality interval estimation,* is considered elementary by statisticians but seldom is discussed in behavioral statistics texts. In this section, I review some history, then describe the method of noncentrality interval estimation in detail.

### *Rationale and History*

Suppose that, as a researcher, you test a drug that you believe enhances performance. You perform a simple two-group experiment with a double-blind control. In this case, you are engaging in “reject–support” (R-S) hypothesis testing (rejecting the null hypothesis will support your belief). The null and alternative hypotheses might be

$$H_0\colon \mu_1 \le \mu_2;\, H_1\colon \mu_1 > \mu_2. \qquad (1)$$

The null hypothesis states that the drug is no better than a placebo. The alternative, which the investigator believes, is that the drug enhances performance. Rejecting the null hypothesis, even at a very low alpha such as .001, need not indicate that the drug has a strong effect, because if sample size is very large relative to the sampling variability of the drug effect, even a trivial effect might be declared highly significant. On the other hand, if sample size is too low, even a strong effect might have a low probability of creating a statistically significant result.

Statistical power analysis (Cohen, 1988) and sample size estimation have been based on the notion that calculations made before data are gathered can help to create a situation in which neither of the above problems is likely to occur. That is, sample size is chosen so that power will be high, but not too high.

There is an alternative situation, “accept–support” (A-S) testing, that attracts far less attention than R-S testing in statistics texts and has had far less impact on the popular wisdom of hypothesis testing. In A-S testing, the statistical null hypothesis is what the experimenter actually wishes to prove. Accepting the statistical null hypothesis supports the researcher’s theory. Suppose, for example, an experiment provides convincing evidence that the above-mentioned drug actually works. The next step might be to provide convincing evidence that it has few, or acceptably low, side effects.

In this case two groups are studied, and some measure of side effects is taken. The null hypothesis is that the experimental group’s level of side effects is less than or equal to the control group’s level. The researcher (or drug company) supporting the research wants not to reject this null hypothesis, because in this case accepting the null hypothesis supports the researcher’s point of view, that is, that the drug is no more harmful than its predecessors.

In a similar vein, a company might wish to show that a generic drug does not differ appreciably in bioavailability from its brand name equivalent. This problem of bioequivalence testing is well known to biostatisticians and has resulted in a very substantial literature (e.g., Chow & Liu, 2000).

Suppose that Drug A has a well-established bioavailability level $\theta_A$, and an investigator wishes to assess the bioequivalence of Drug B with Drug A. One might engage in A-S testing, that is, test the null hypothesis that

$$\theta_A = \theta_B \qquad (2)$$

and declare the two drugs bioequivalent if this null hypothesis is not rejected. However, the perils of such A-S testing are even greater than in R-S testing. Specifically, simply running a sloppy, low-power experiment will tend to result in nonrejection of the null hypothesis, even if the drugs differ appreciably in bioavailability. Thus, paradoxically, someone trying to establish the bioequivalence of Drug B with Drug A could virtually guarantee success simply by using too small a sample size. Moreover, with extremely large sample sizes, Drug B might be declared nonequivalent to Drug A even if the difference between them is trivial.

Because of such problems, biostatisticians decided long ago that the test for strict equality is inappropriate for bioavailability studies (Metzler, 1974). Rather, a dual hypothesis test should be performed. Suppose that the Food and Drug Administration has determined that any drug with bioavailability within 20% of $\theta_A$ may be considered bioequivalent and prescribed in its stead. Suppose that $\theta_1$ and $\theta_2$ represent these bioequivalence limits. Then establishing bioequivalence of Drug B with Drug A might amount to rejecting the following hypothesis,

$$H_0\colon \theta_B > \theta_2 \text{ or } \theta_B < \theta_1 \tag{3}$$

against the alternative

$$H_a\colon \theta_1 \leq \theta_B \leq \theta_2. \tag{4}$$

In practice, this usually amounts to testing two one-sided hypotheses,

$$H_{01}\colon \theta_B \leq \theta_1 \text{ versus } H_{a1}\colon \theta_B > \theta_1 \tag{5}$$

and

$$H_{02}\colon \theta_B \geq \theta_2 \text{ versus } H_{a2}\colon \theta_B < \theta_2. \tag{6}$$

An alternative approach (Westlake, 1976) is to construct a confidence interval for $\theta_B$. Bioequivalence would be declared if the confidence interval falls entirely within the established bioequivalence limits.

In other contexts, particularly the more exploratory studies performed in psychology, the research goal may be simply to pinpoint the nature of a parameter rather than to decide whether it is within a known fixed range. In that case, reporting the endpoints of a confidence interval (without announcing an associated decision) may be an appropriate conclusion to an analysis. In any case, because the hypothesis test may be performed with the confidence interval, it seems that the confidence interval should always be reported. It contains all the information in a hypothesis test result, and more.

In structural equation modeling, which includes factor analysis and multiple regression as special cases, statistical testing prior to 1980 was limited to a chi-square test of perfect fit. In this procedure, the statistical null hypothesis is that the model fits perfectly in the population. This hypothesis test was performed, and a model was judged to fit the data “sufficiently well” if the null hypothesis was not rejected. There was widespread dissatisfaction with the test, because no model would be expected to fit perfectly, and so large sample sizes usually led to rejection of a model, even if it fit the data quite well. In this arrangement, enhanced precision actually worked against the researcher’s interests. Steiger and Lind (1980) suggested that the traditional null hypothesis test of perfect fit of a structural model be replaced by a confidence interval on the root-mean-square error of approximation (RMSEA), an index of population badness of fit that compensated for the complexity of the model.

MacCallum, Browne, and Sugawara (1996) suggested augmenting the confidence interval with a pair of hypothesis tests. They considered a population RMSEA value of .05 to be indicative of a close-fitting model, whereas a value of .08 or more was evidence of marginal to poor fit. Consequently, a test of close fit would test the null hypothesis that the RMSEA is greater than or equal to .05 against the alternative that it is less than .05. Rejection of the null hypothesis indicates close fit. A test of not-close fit tests the null hypothesis that the RMSEA is less than or equal to .08 against the alternative that it is greater than .08. Rejection of the null hypothesis indicates that fit is not close. MacCallum et al. demonstrated in detail how, with such hypothesis tests, power calculations could be performed and required sample sizes estimated. These two one-sided tests can be performed easily and simultaneously with a single $1 - 2\alpha$ confidence interval recommended by Steiger (1989). Simply construct the confidence interval and see whether its upper end is below .05 (in which case the test of close fit results in rejection at the alpha level) and whether its lower end exceeds .08 (in which case the test of not-close fit results in rejection at the alpha level). The confidence interval provides all the information in both hypothesis tests, and more.

Fleishman (1980) suggested interval estimation as a supplement for the *F* test in ANOVA. He gave examples of how to compute exact confidence intervals on a number of useful quantities, such as the signal-to-noise ratio, in ANOVA. These confidence intervals offered clear advantages over the traditional hypothesis test. Other authors have noted the existence of exact confidence intervals for the standardized effect size in the simplest special case of ANOVA, the two-sample *t* test (e.g., Hedges & Olkin, 1985).

The rationale for switching from hypothesis testing to confidence interval estimation is straightforward (Steiger & Fouladi, 1997). Unfortunately, the exact interval estimation procedures of Steiger and Lind (1980), Fleishman (1980), and Hedges and Olkin (1985) are virtually impossible to compute accurately by hand. However, by 1990, microcomputer capabilities had advanced substantially. The RMSEA confidence interval was implemented in general purpose structural equation modeling software (Mels, 1989; Steiger, 1989) and, by the late 1990s, had achieved widespread use. Steiger (1990) presented general procedures for constructing confidence intervals on measures of effect size in covariance structure analysis, ANOVA, contrast analysis, and multiple regression. Steiger and Fouladi (1992) produced a general computer program, R2, that performed exact confidence interval estimation of the squared multiple correlation in multiple regression. Taylor and Muller (1995, 1996) have presented general procedures for analyzing power and noncentrality in the general linear model, including an analysis of the impact of restriction of published articles to significant results. Steiger and Fouladi (1997) demonstrated general procedures for confidence interval calculations, and Steiger (1999) implemented these in a commercial software package. Smithson (2001) discussed a number of confidence interval procedures in fixed and random regression models and included SPSS macros for calculating confidence intervals for noncentral distributions. Reiser (2001) discussed confidence intervals on functions of Mahalanobis distance.

### *General Theory of Noncentrality-Based Interval Estimation*

In this section, I review the general theoretical principles for constructing exact confidence intervals for effect size, power, and sample size in the balanced fixed-effects between-subjects ANOVA. For a more detailed discussion of these principles, see Steiger and Fouladi (1997). Throughout what follows, I adopt a simple notational device: When several groups or cells are sampled, I use $N_{\text{tot}}$ to stand for the total sample size and use $n$ to stand for the number of observations in each group.

I begin this section with a brief nontechnical discussion of noncentral distributions. The *t*, chi-square, and *F* distributions are special cases of more general distributions called the noncentral *t*, noncentral chi-square, and noncentral *F*. Each of these noncentral distributions has an additional parameter, called the *noncentrality parameter*. For example, whereas the *F* distribution has two parameters (the numerator and denominator degrees of freedom), the noncentral *F* has these two plus a noncentrality parameter (often indicated with the symbol $\lambda$). When the noncentral *F* distribution has a noncentrality parameter of zero, it is identical to the *F* distribution, so it includes the *F* distribution as a special case. Similar facts hold for the *t* and chi-square distributions. What makes the noncentrality parameter especially important is that it is related very closely to the truth or falsity of the null hypotheses that these distributions are typically used to test. Thus, for example, when the null hypothesis of no difference between two means is correct, the standard *t* statistic has a distribution that has a noncentrality parameter of zero, whereas if the null hypothesis is false, it has a noncentral *t* distribution, that is, the noncentrality parameter is nonzero. The more false the null hypothesis, the larger the absolute value of the noncentrality parameter for a given alpha and sample size.

Most confidence intervals in introductory textbooks are derived by simple manipulation of a statement about interval probability of a sampling distribution. This approach cannot be used to generate exact confidence intervals for many quantities of fundamental importance in statistics. As an example, consider the sample squared multiple correlation, whose distribution changes as a function of the population squared multiple correlation. Confidence intervals for the squared multiple correlation are very informative yet are not discussed in standard texts, because a single simple formula for the direct calculation of such an interval cannot be obtained in a manner analogous to the way one obtains a confidence interval for the population mean $\mu$. Steiger and Fouladi (1997) discussed a general method for confidence interval construction that handles many such interesting examples. The method combines two general principles, which they called the *confidence interval transformation principle* and the *inversion confidence interval principle.* The former is obvious but seldom discussed formally. The latter is referred to by a variety of names in textbooks and review articles (Casella & Berger, 2002; Steiger & Fouladi, 1997), yet it does not seem to have found its way into the standard behavioral statistics textbooks, primarily because its implementation involves some difficult computations. However, the method is easy to discuss in principle and is no longer impractical. When the two principles are combined, a number of very useful confidence intervals result.

*Proposition 1: Confidence interval transformation principle.* Let $f(\theta)$ be a monotone function of $\theta$, that is, a function whose slope never changes sign and is never zero. Let $l_1$ and $l_2$ be lower and upper endpoints of a $1 - \alpha$ confidence interval on quantity $\theta$. Then, if the function is increasing, $f(l_1)$ and $f(l_2)$ are lower and upper endpoints, respectively, of a $100(1 - \alpha)\%$ confidence interval on $f(\theta)$. If the function is decreasing, $f(l_2)$ and $f(l_1)$ are lower and upper endpoints. Here are two elementary examples of this principle.

*Example 1:* Suppose you read in a textbook how to calculate a confidence interval for the population variance $\sigma^2$. However, you desire a confidence interval for $\sigma$. Because $\sigma$ takes on only nonnegative values, it is a monotonic increasing function of $\sigma^2$ over its domain. Hence, the confidence interval for $\sigma$ is obtained by taking the square root of the endpoints for the corresponding confidence interval for $\sigma^2$.

*Example 2:* Suppose one calculates a confidence interval for $z(\rho)$, the Fisher transform of $\rho$, the population correlation coefficient. Taking the inverse Fisher transform of the endpoints of this interval will give a confidence interval for $\rho$.

This is, in fact, the method used to calculate the standard confidence interval for a correlation.

These examples show why Proposition 1 is very useful in practice. A statistical quantity we are very interested in—such as $\rho$—may be a simple function of a quantity—such as $z(\rho)$—we are not so interested in, but for which we can easily obtain a confidence interval. Next, we define the inversion confidence interval principle.

*Proposition 2: Inversion confidence interval principle.* Let $x$ be the observed value of $X$, a random variable with a continuous cdf (cumulative distribution function) $F(x, \lambda) = Pr(X \leq x|\lambda)$ for some numerical parameter $\lambda$. Let $\alpha_1 + \alpha_2 = \alpha$ with $0 < \alpha < 1$ be fixed values. If $F(x, \lambda)$ is strictly decreasing in $\lambda$, for fixed values of $x$, choose $l_1(x)$ and $l_2(x)$ so that $Pr[X \leq x|\lambda = l_1(x)] = 1 - \alpha_2$ and $Pr[X \leq x|\lambda = l_2(x)] = \alpha_1$. If $F(x, \lambda)$ is strictly increasing in $\lambda$, for fixed values of $x$, choose $l_1(x)$ and $l_2(x)$ so that $Pr[X \leq x|\lambda = l_1(x)] = \alpha_1$ and $Pr[X \leq x|\lambda = l_2(x)] = 1 - \alpha_2$. Then the random interval $[l_1(x), l_2(x)]$ is a $100(1 - \alpha)\%$ confidence interval for $\lambda$. Upper or lower $100(1 - \alpha)\%$ confidence bounds (or "one-sided confidence intervals") may be obtained by setting $\alpha_1$ or $\alpha_2$ to zero.

For a simple graphically based explanation of Proposition 2, consult Steiger and Fouladi (1997, pp. 237–239). For a clear, succinct discussion with partial proof, see Casella and Berger (2002, p. 432), who referred to this as "pivoting" the cdf. In this article, I assume $\alpha_1 = \alpha_2 = \alpha/2$, although such an interval may not be the minimum width for a given $\alpha$. Proposition 2 implies a simple approach to interval estimation: Suppose you have observed an $F$ statistic with a value $x$ and known degrees of freedom $\nu_1$ and $\nu_2$. Denote the cumulative distribution of the $F$ statistic by $F(x, \lambda)$, where $\lambda$ is the noncentrality parameter. It can be shown that if $\nu_1$, $\nu_2$, and $x$ are held constant at any positive value, then $F(x, \lambda)$ is strictly decreasing in $\lambda$. Accordingly, Proposition 2 can be used. To calculate a $100(1 - \alpha)\%$ confidence interval on the noncentrality parameter of the $F$ distribution, use the following steps.

1. Calculate the cumulative probability $p$ of $x$ in the central $F$ distribution. If $p$ is below $\alpha/2$, then both limits of the confidence interval are zero. If $p$ is below $1 - \alpha/2$, the lower limit of the confidence interval is zero, and the upper limit must be calculated (go to Step 3). Otherwise, calculate both limits of the confidence interval, using Steps 2 and 3.

2. To calculate the lower limit, find the unique value of $\lambda$ that places $x$ at the $1 - \alpha/2$ cumulative probability point of a noncentral $F$ distribution with $\nu_1$ and $\nu_2$ degrees of freedom.

3. To calculate the upper limit, find the unique value of $\lambda$ that places $x$ at the $\alpha/2$ cumulative probability point of a noncentral $F$ distribution with $\nu_1$ and $\nu_2$ degrees of freedom.

Calculating a confidence interval for $\lambda$ thus requires iterative calculation of the unique value of $\lambda$ that places an observed value of $F$ at a particular percentile of the noncentral $F$ distribution.[1] In what follows, I give a variety of examples of confidence interval calculations. Some will be at the 95% level of confidence, others at the less common 90% level. In a later section, I discuss why, when confidence intervals are used to perform a hypothesis test at the .05 level, a 90% interval may be appropriate in some situations and a 95% interval in others. At that point, I describe how to select confidence intervals at the appropriate level to perform a particular hypothesis test.

## Measures of Standardized Effect Size

Now I examine some more ambitious examples. For simplicity of exposition, I assume in this section that either the freeware program NDC (noncentral distribution calculator; see Footnote 1) or other software is available to compute a confidence interval on $\lambda$, the noncentrality parameter of a noncentral $F$ distribution. Consider the one-way, fixed-effects ANOVA, in which $p$ means are compared for equality, and there are $n$ observations per group. The overall $F$ statistic has a distribution that is a noncentral $F$, with degrees of freedom $p - 1$ and $p(n - 1) = N_{\text{tot}} - p$.

The noncentrality parameter $\lambda$ can be expressed in a number of ways. One formula that appears frequently in textbooks is

$$\lambda = n \sum_{j=1}^{p} \left(\frac{\alpha_j}{\sigma}\right)^2. \tag{7}$$

The $\alpha_j$ values in Equation 7 are the *effects* as commonly defined in ANOVA, that is,

$$\alpha_j = \mu_j - \mu. \tag{8}$$

If $\mu_j$ is the mean of the $j$th group, and $\mu$ is the overall mean, then $\mu$ is, in the case of equal $n$, simply the arithmetic average of the $\mu_j$. More generally (although in what follows I assume a balanced design unless stated otherwise),

$$\mu = \sum_{j=1}^{p} \frac{n_j}{N_{\text{tot}}} \mu_j. \tag{9}$$

[1] NDC (noncentral distribution calculator), a freeware Windows program for calculating percentage points and noncentrality confidence intervals for noncentral $F$, $t$, and chi-square distributions, is available for direct download from the author's website (http://www.statpower.net).

The quantity $\alpha_j/\sigma$ is a *standardized effect*, that is, the effect expressed in standard deviation units. The quantity $\lambda/n$ is therefore the sum of squared standardized effects. There are numerous ways one might convert the sum of squared standardized effects into an overall measure of effect size. For example, suppose we average these squared standardized effects in order to obtain an overall measure of strength of effects in the design. The arithmetic average of the *p* squared standardized effects, sometimes called the *signal-to-noise ratio* (Fleishman, 1980), is as follows:

$$f^2 = \frac{1}{p}\sum_{j=1}^{p}\left(\frac{\alpha_j}{\sigma}\right)^2 = \frac{\lambda}{np} = \frac{\lambda}{N_{\text{tot}}}. \tag{10}$$

One problem with this measure is that it is the average squared effect and so is not in the proper unit of measurement. A potential solution is to simply take the square root of the signal-to-noise ratio, obtaining

$$f = \sqrt{\frac{\lambda}{N_{\text{tot}}}} = \sqrt{\frac{1}{p}\sum_{j=1}^{p}\left(\frac{\alpha_j}{\sigma}\right)^2}. \tag{11}$$

In a one-way ANOVA with *p* groups and equal *n*, the effects are constrained to sum to zero, so there are actually only $p - 1$ independent effects. Thus, an alternative measure, $\lambda/[(p - 1)n]$, is the average squared independent standardized effect, and the root-mean-square standardized effect (RMSSE) is as follows:

$$\Psi = \sqrt{\frac{\lambda}{(p-1)n}} = \sqrt{\frac{1}{p-1}\sum_{j=1}^{p}\left(\frac{\alpha_j}{\sigma}\right)^2}. \tag{12}$$

Equations 11 and 12 demonstrate that the relationships between $\Psi$, *f*, and the noncentrality parameter $\lambda$ are straightforward.

In order to obtain a confidence interval for $\Psi$, we proceed as follows. First, we obtain a confidence interval estimate for $\lambda$. Next, we invoke the confidence interval transformation principle to directly transform the endpoints by dividing by $(p - 1)n$. Finally, we take the square root. The result is an exact confidence interval on $\Psi$.

*Example 3:* Suppose a one-way fixed-effects ANOVA is performed on four groups, each with a sample size of 20, and that an overall *F* statistic of 5.00 is obtained, with 3 and 76 degrees of freedom, with a probability level of .0032. The *F* test is thus “highly significant,” and the null hypothesis is rejected at the .01 level. Some investigators might interpret this result as implying that a powerful experimental effect was found and that this was determined with high precision. In this case, the noncentrality interval estimate provides a more informative and somewhat different account of what has been found.

The 95% confidence interval for $\lambda$ ranges from 1.8666 to 32.5631. To convert this to a confidence interval for $\Psi$, we use Equation 12. The corresponding confidence interval for $\Psi$ ranges from .1764 to .7367. Effects are almost certainly “here,” but they are on the order of half a standard deviation, what is commonly considered a medium-size effect. Moreover, the size of the effects has not been determined with high precision.

*Example 4:* Fleishman (1980) described the calculation of confidence intervals on the noncentrality parameter of the noncentral *F* distribution to obtain, in a manner equivalent to that used in the previous two examples, confidence intervals on $f^2$ and $\omega^2$, the latter of which is defined as

$$\omega^2_{\text{A(partialed)}} = \frac{S^2_{\mu_{\text{A}}}}{S^2_{\mu_{\text{A}}} + \sigma^2_{\text{e}}}, \tag{13}$$

where $S^2_{\mu_{\text{A}}}$ is the variance of *p* means for the levels of a particular effect A, that is,

$$S^2_{\mu_{\text{A}}} = (1/p)\sum_{j=1}^{p}(\mu_j - \bar{\mu})^2 \tag{14}$$

and $\sigma^2_{\text{e}}$ is the within-cell variance. $\omega^2_{\text{A(partialed)}}$ may be thought of as the proportion of the variance remaining (after all other main effects and interactions have been partialed out) that is explained by the effect. (In what follows, for simplicity, I refer to the coefficient simply as $\omega^2$.) There are simple relationships between $f^2$, $\omega^2$, and $\lambda$, specifically,

$$f^2 = \frac{\omega^2}{1-\omega^2} = \frac{\lambda}{pn} = \frac{\lambda}{N_{\text{tot}}} \tag{15}$$

and

$$\omega^2 = \frac{f^2}{1+f^2} = \frac{\lambda}{\lambda + N_{\text{tot}}}. \tag{16}$$

Fleishman (1980) cited an example given by Venables (1975) of a five-group ANOVA with $n = 11$ per cell and an observed *F* of 11.221. In this case the 90% confidence interval for the noncentrality parameter $\lambda$ has endpoints 19.380 and 71.549. Once we obtain the confidence interval for $\lambda$, it is a trivial matter to transform the limits of the interval to confidence limits for $\omega^2$, using Equation 16. For example, the lower limit becomes

$$\frac{19.380}{19.380 + (5)(11)} = \frac{19.380}{19.380 + 55} = \frac{19.380}{74.380} = .261. \tag{17}$$

In a similar manner, the upper limit of the confidence interval can be calculated as .565. The confidence interval has determined with 90% confidence that the main effect accounts for between 26.1% and 56.5% of the variance in the dependent variable.

## *General Procedures for Effect Size Intervals in Between-Subjects Factorial ANOVA*

In a previous example, we saw how easy it is to construct a confidence interval on measures of effect size in one-way ANOVA, provided a confidence interval for $\lambda$ has been computed. In this section, a completely general method is demonstrated for computing confidence intervals for various measures of standardized effect size in completely between-subjects factorial ANOVA designs with equal sample size $n$ per cell.

We begin with a general formula relating the noncentrality parameter $\lambda$ with the RMSSE in any completely between-subjects factorial ANOVA. Let $\theta$ stand for a particular effect, and $n$ the sample size per cell. Then

$$\Psi_\theta = \sqrt{\frac{\lambda_\theta}{n_\theta\, df_\theta}}. \tag{18}$$

In Equation 18, $n_\theta$ is equal to $n$ (the number of observations in each cell of the design) multiplied by the product of the numbers of levels in all the factors not represented in the effect currently under consideration; $df_\theta$ is the numerator degrees of freedom parameter for the effect under consideration.

There are simple relationships between the RMSSE and other measures of standardized effect size. Specifically, for a general factorial ANOVA,

$$f_\theta^2 = \frac{\Psi_\theta^2\, df_\theta}{\text{Cells}_\theta} = \frac{\lambda_\theta}{N_{\text{tot}}}, \tag{19}$$

where $\text{Cells}_\theta$ is, for any main effect, the number of levels of the effect. For any interaction, it is the product of the numbers of levels for all factors involved in the interaction. The relationship between $f^2$ and $\omega^2$ is given in Equation 16.

Some examples of these quantities, for a four-way ANOVA, with $p, q, r,$ and $s$ levels of factors A, B, C, and D, respectively, are given in Table 1. The table may be used also for one-, two-, or three-way ANOVAs simply by eliminating terms involving levels not represented in the design. For example, in a three-way ANOVA, the BC interaction effect has $(q - 1)(r - 1)$ numerator degrees of freedom, and $n_{\text{BC}}$ is $np$, because there is no $s$ in this design. The error degrees of freedom in a three-way ANOVA are $pqr(n - 1)$. In the following two examples, I demonstrate how to compute a 90% confidence interval on various measures of effect, using the information in the table.

*Example 5:* Suppose that, as a researcher, you perform a three-way 2 × 3 × 7 ANOVA, with $n = 6$ observations per cell. In this case, we have $p = 2$, $q = 3$, and $r = 7$.

Suppose that, for the A main effect, you observe an $F$ statistic of 4.2708, which, with 1 and 210 degrees of freedom, has $p = .0400$. We first calculate a confidence interval for $\lambda$. The endpoints of this interval are $\lambda_{\text{lower}} = 0.100597$ and $\lambda_{\text{upper}} = 13.8186$. To convert these to confidence intervals on $\Psi$, $f^2$, $f$, and $\omega^2$, we apply Equations 18, 19, and 16. For the A effect, we have $n_{\text{A}} = (6)(3)(7) = 126$, $df_{\text{A}} = (2 - 1) = 1$, $\text{Cells}_{\text{A}} = 2$, and $N_{\text{tot}} = 252$. Hence, for $\Psi$ we have, from Equation 18,

Table 1
*Key Quantities for Computing Effect Size Intervals in Four-Way Analysis of Variance*

| Source | Levels | $df_\theta$ | $n_\theta$ |
|---|---|---|---|
| A | $p$ | $p - 1$ | $nqrs$ |
| B | $q$ | $q - 1$ | $nprs$ |
| C | $r$ | $r - 1$ | $npqs$ |
| D | $s$ | $s - 1$ | $npqr$ |
| AB | | $(p - 1)(q - 1)$ | $nrs$ |
| AC | | $(p - 1)(r - 1)$ | $nqs$ |
| AD | | $(p - 1)(s - 1)$ | $nqr$ |
| BC | | $(q - 1)(r - 1)$ | $nps$ |
| BD | | $(q - 1)(s - 1)$ | $npr$ |
| CD | | $(r - 1)(s - 1)$ | $npq$ |
| ABC | | $(p - 1)(q - 1)(r - 1)$ | $ns$ |
| ABD | | $(p - 1)(q - 1)(s - 1)$ | $nr$ |
| ACD | | $(p - 1)(r - 1)(s - 1)$ | $nq$ |
| BCD | | $(q - 1)(r - 1)(s - 1)$ | $np$ |
| ABCD | | $(p - 1)(q - 1)(r - 1)(s - 1)$ | $n$ |
| Error | | $pqrs(n - 1)$ | |

*Note.* $\theta$ represents a particular effect; $n$ represents the sample size per cell; and $p$, $q$, $r$, and $s$ represent levels of factors A, B, C, and D, respectively.

$$\Psi_{\text{lower}} = \sqrt{\frac{0.100597}{(126)(1)}} = 0.028256,\ \Psi_{\text{upper}} = \sqrt{\frac{13.8186}{(126)(1)}} = 0.331167. \tag{20}$$

For $f^2$ and $f$ we have, for the lower limits,

$$f^2_{\text{lower}} = \frac{0.100597}{252} = 0.000399194,\ f_{\text{lower}} = 0.01998. \tag{21}$$

For the upper limits, we obtain $f^2_{\text{upper}} = 0.0548357$ and $f_{\text{upper}} = 0.234170$.

We can also convert the confidence limits for $f^2$ into limits for $\omega^2$, using Equation 16. We have

$$\omega^2_{\text{lower}} = \frac{f^2_{\text{lower}}}{1 + f^2_{\text{lower}}} = \frac{0.000399194}{1.000399194} = 0.000399035. \tag{22}$$

In a similar manner, we obtain the upper limit as $\omega^2_{\text{upper}} = 0.0519851$.

*Example 6:* Table 1 can also be used for a two-way ANOVA, simply by letting $r = 1$ and $s = 1$ and ignoring all effects involving factors C and D. Suppose, for example, one were to perform a two-way 2 × 7 ANOVA, with $n = 4$ observations per cell, and the $F$ statistic for the AB interaction is observed to be 2.50. The key quantities are $df_{\mathrm{AB}} = 6$, $df_{\mathrm{error}} = 42$, $n_{\mathrm{AB}} = 4$, and $\mathrm{Cells}_{\mathrm{AB}} = 14$. The confidence limits for $\lambda_{\mathrm{AB}}$ are $\lambda_{\mathrm{lower}} = 0.462800$ and $\lambda_{\mathrm{upper}} = 25.8689$. Consequently, from Equation 18, the confidence limits for the RMSSE are

$$\Psi_{\mathrm{lower}} = \sqrt{\frac{\lambda_{\mathrm{lower}}}{n_{\mathrm{AB}} df_{\mathrm{AB}}}} = \sqrt{\frac{0.462800}{(4)(6)}} = 0.1389, \qquad (23)$$

$$\Psi_{\mathrm{upper}} = \sqrt{\frac{25.8689}{(4)(6)}} = 1.0382. \qquad (24)$$

The confidence intervals for $f^2$ and $f$ are

$$f^2_{\mathrm{lower}} = \frac{0.462800}{56} = 0.008264, f_{\mathrm{lower}} = 0.0909, \qquad (25)$$

$$f^2_{\mathrm{upper}} = \frac{25.8689}{56} = 0.461944, f_{\mathrm{upper}} = 0.6797. \qquad (26)$$

Using Equation 16, we convert the above to the following confidence limits for $\omega^2$:

$$\omega^2_{\mathrm{lower}} = \frac{f^2_{\mathrm{lower}}}{1 + f^2_{\mathrm{lower}}} = \frac{0.008264}{1.008264} = 0.0082, \qquad (27)$$

$$\omega^2_{\mathrm{upper}} = \frac{0.461944}{1.461944} = 0.3160. \qquad (28)$$

## Multiple Regression With Fixed Regressors

One standardized index of the size of effects is to compute the squared multiple correlation coefficient between the independent variable and the scores on the dependent variable. This index, in the population, characterizes the strength of the effect. ANOVA may be conceptualized as a linear regression model with fixed independent variables. In this case, the theory of multiple regression with fixed regressors applies. It is important to realize (e.g., Sampson, 1974) that the theory for fixed regressors, although it shares many similarities with that for random regressors, has important differences, which are especially apparent when considering the nonnull distributions of the variables. The general model is

$$E(\boldsymbol{\eta}) = \mathbf{X}\boldsymbol{\beta}, \qquad (29)$$

where $\boldsymbol{\eta}$ is an $N_{\mathrm{tot}} \times 1$ random vector, $\mathbf{X}$ is an $N_{\mathrm{tot}} \times p$ matrix, and $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown parameters.

This model includes model errors ($\boldsymbol{\epsilon}$) that are assumed to be independently and identically distributed with a normal distribution, zero mean, and variance $\sigma^2$. That is,

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} = \hat{\boldsymbol{\eta}} + \boldsymbol{\epsilon}, \qquad (30)$$

and $\boldsymbol{\epsilon}$ has a multivariate normal distribution with zero mean vector $\mathbf{0}$ and covariance matrix $\sigma^2\mathbf{I}$, with $\mathbf{I}$ an identity matrix. It is common to partition $\boldsymbol{\beta}$ into

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \boldsymbol{\beta}_1 \end{bmatrix}, \qquad (31)$$

where $\beta_0$ is an intercept term. Correspondingly, $\mathbf{X}$ is partitioned as

$$\mathbf{X} = [\mathbf{1} \quad \mathbf{X}_1], \qquad (32)$$

where $\mathbf{1}$ is a column of ones and $\mathbf{X}_1$ contains the original $X$ scores transformed into deviations about their sample means.

Consider now a set of observed scores $\mathbf{y}$, representing realizations of the random variables in $\boldsymbol{\eta}$. If $\mathbf{X}_1$ has $p - 1$ columns, then an $F$ statistic for testing the hypothesis that $\boldsymbol{\beta}_1 = 0$ is

$$F = \frac{R^2/(p-1)}{(1-R^2)/(N_{\mathrm{tot}} - p)}. \qquad (33)$$

This statistic has a noncentral $F$ distribution with $p - 1$ and $N_{\mathrm{tot}} - p$ degrees of freedom, with a noncentrality parameter given by

$$\lambda = \frac{\boldsymbol{\beta}'\mathbf{X}'(\mathbf{I} - \mathbf{P}_1)\mathbf{X}\boldsymbol{\beta}}{\sigma^2} = \frac{\boldsymbol{\beta}'\mathbf{X}'\mathbf{Q}_1\mathbf{X}\boldsymbol{\beta}}{\sigma^2}. \qquad (34)$$

For any matrix $\mathbf{A}$ of full column rank, $\mathbf{P}_{\mathbf{A}}$ is the column space projection operator $\mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'$ and $\mathbf{Q}_{\mathbf{A}}$ the complementary projector $\mathbf{I} - \mathbf{P}_{\mathbf{A}}$.

We now turn to an application of this theory in the context of ANOVA. Consider the simple case of a one-way fixed-effects ANOVA with $n$ observations in each of $p$ independent groups. It is well-known that this model can be written in the form of Equation 29, where $\mathbf{X}$ is a design matrix with $N_{\mathrm{tot}} = np$ rows and $p$ columns, and $\boldsymbol{\beta}$ contains ANOVA parameters.

We are not interested in $R^2$ per se. Rather, we are interested in the corresponding quantity $\rho^2$ in an infinite population of observations in which treatment groups are represented equally. There are several alternative ways of conceptualizing such a quantity. Formally, we can define $\rho^2$ as the probability limit of $R^2$, that is,

$$\rho^2 = \operatorname*{plim}_{n\to\infty}(R^2). \qquad (35)$$

This is the constant that $R^2$ converges to as the sample size increases without bound. It can be proven (see Appendix A) that, with this definition of $\rho^2$, the noncentrality parameter is equivalent to

$$\lambda = N_{\text{tot}} \frac{\rho^2}{1 - \rho^2}, \quad (36)$$

and so

$$\rho^2 = \frac{\lambda}{\lambda + N_{\text{tot}}}. \quad (37)$$

Consequently, a confidence interval for λ may be converted easily into a confidence interval on $\rho^2$ or $\rho$, because $\rho$ is nonnegative. $\rho^2$ represents the coefficient of determination for predicting scores on the dependent variable from only a knowledge of the population means of the groups in an infinite population in which all treatment groups are equally represented.

*Example 7:* Suppose that **X** is set up as in Equation 38 to represent a full rank design matrix for a one-way ANOVA, with three groups, and $n = 3$, and that the scores in **y** are 1, 2, 3, 4, 5, 6, 7, 8, 9. In this parameterization, $\beta_0$ corresponds to $\mu_3$, $\beta_1$ corresponds to $\mu_1 - \mu_3$, and $\beta_2$ corresponds to $\mu_2 - \mu_3$. The group means are 2, 5, 8, and the group variances are all 1.

$$\begin{bmatrix} y_{11} \\ y_{21} \\ y_{31} \\ y_{12} \\ y_{22} \\ y_{32} \\ y_{13} \\ y_{23} \\ y_{33} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \epsilon_{11} \\ \epsilon_{21} \\ \epsilon_{31} \\ \epsilon_{12} \\ \epsilon_{22} \\ \epsilon_{32} \\ \epsilon_{13} \\ \epsilon_{23} \\ \epsilon_{33} \end{bmatrix}. \quad (38)$$

In this case, it is easy to show using any standard multiple regression program that the sample squared multiple correlation for predicting **y** from **X** is .90 and that the $F$ statistic for testing the null hypothesis that $\rho^2 = 0$ is

$$F(2, 6) = \frac{R^2/2}{(1 - R^2)/6} = \frac{.9/2}{.1/6} = 27.0. \quad (39)$$

This $F$ statistic is identical to the one obtained by performing a one-way fixed-effects ANOVA on the data. The 90% confidence interval for λ has endpoints of $\lambda_1 = 10.797$ and $\lambda_2 = 119.702$. The lower endpoint for the confidence interval on $\rho^2$, the coefficient of determination, is thus

$$\frac{10.797}{10.797 + 9} = \frac{10.797}{19.797} = .545, \quad (40)$$

and the upper endpoint is

$$\frac{119.702}{119.702 + 9} = \frac{119.702}{128.702} = .930. \quad (41)$$

With one-way ANOVA and equal $n$ per group, this confidence interval is identical to the one for $\omega^2$ discussed earlier. Note also that the sample $R^2$ is positively biased with small sample sizes and will consequently be much closer to the upper end of the confidence interval than the lower.

One of several alternative methods for parameterizing the linear model in Equation 29 is to use what is sometimes called *effect coding.* In this case, the entries in **X** correspond to the contrast weights applied to group means in the ANOVA null hypothesis. For example, the hypothesis of no treatments in a one-way ANOVA with three groups corresponds to two contrasts simultaneously being zero, that is, $\mu_1 - \mu_3 = 0$ and $\mu_2 - \mu_3 = 0$. The contrast weights for the two hypotheses are thus 1, 0, −1 and 0, 1, −1. Thus, omnibus effect size in ANOVA can be expressed as the multiple correlation between a set of contrast weights and the dependent variable.

There has been a fair amount of discussion in the applied literature (Ozer, 1985; Rosenthal, 1991; Steiger & Ward, 1987) about whether the coefficient of determination is overly pessimistic in describing the strength of effects. Those who prefer $\rho$ may convert a confidence interval on $\rho^2$ to a confidence interval on $\rho$ simply by taking the square root of the endpoints of the former.

## Confidence Intervals on Single-Contrast Measures of Effect Size

Rosenthal et al. (2000) argued convincingly for the importance of replacing the omnibus hypothesis in ANOVA with hypotheses that focus on substantive research questions. Often such hypotheses involve single contrasts $\psi$ of the form $\psi = \sum_{j=1}^{p} c_j\mu_j$, with $c_j$, the contrast weights and the null hypothesis being that $\psi = 0$. Rosenthal et al. discussed several different correlational measures for assessing the status of hypotheses on a single contrast. In this section, I discuss methods for exact confidence interval estimation of measures of effect size for a single contrast, including the population equivalent of the correlation measure $r^2_{\text{contrast}}$ discussed by Rosenthal et al.

### *Exact Confidence Intervals for Standardized Contrast Effect Size*

Consider a contrast hypothesis on means, of the form

$$H_0\colon \psi = \sum_{j=1}^{p} c_j\mu_j = 0. \quad (42)$$

With equal sample sizes of $n$ per group, this hypothesis may be tested with a $t$ statistic of the form

$$t = \sqrt{n} \frac{\hat{\psi}}{\sqrt{MS_{\text{within}}(\sum_{j=1}^{p} c_j^2)}}, \quad (43)$$

with

$$\hat{\psi} = \sum_{j=1}^{p} c_j \bar{Y}_{\cdot j}, \tag{44}$$

where $\bar{Y}_{\cdot j}$ represents the sample mean of the *j*th group. The standardized effect size $E_s$ is the size of the contrast in standard deviation units, that is,

$$E_s = \frac{\psi}{\sigma}. \tag{45}$$

The test statistic has a noncentral *t* distribution with $p(n - 1)$ degrees of freedom and a noncentrality parameter of

$$\delta = \sqrt{\frac{n}{\sum_{j=1}^{p} c_j^2}}\; E_s = (L)E_s. \tag{46}$$

To estimate $E_s$, one obtains a confidence interval for $\delta$, using the method discussed by Steiger and Fouladi (1997), and transforms the endpoints of the confidence interval by dividing by $L$ (i.e., the expression under the radical in Equation 46), as shown in the example below.

*Example 8:* The data in Table 2 represent four independent groups of three observations each. Suppose one wished to test the following null hypothesis:

$$\psi = \frac{\mu_1 + \mu_4}{2} - \frac{\mu_2 + \mu_3}{2} = 0. \tag{47}$$

This hypothesis tests whether the average of the means of the first and fourth groups is equal to the average of the means of the other two groups. Suppose we observe $t(8) = 1.7321$. The traditional 95% confidence interval for $\psi$ ranges from $-0.3314$ to 2.3314. Because mean square error is 1 in this example, we would expect a confidence interval for $E_s$ to be similar. Actually, it is somewhat narrower. The 95% confidence interval for $\delta$ ranges from $-0.4429$ to 3.8175. The sum of squared contrast weights is 1, so $L = \sqrt{3}$, and the endpoints of the confidence interval are divided by $\sqrt{3}$ to obtain 95% confidence limits of $-0.2557$ and 2.2041 for $E_s$.

Table 2
*Sample Data for a One-Way Analysis of Variance*

| Group 1 | Group 2 | Group 3 | Group 4 |
|---|---|---|---|
| 1 | 3 | 8 | 12 |
| 2 | 4 | 9 | 13 |
| 3 | 5 | 10 | 14 |

### *Exact Confidence Intervals for* $\rho^2_{\text{contrast}}$

Rosenthal et al. (2000) discussed the sample statistic $r^2_{\text{contrast}}$, which is the squared partial correlation between the contrast weight vector discussed in the previous section and the scores in **y**, with all other sources of systematic between-groups variation partialed out. Consider the data discussed in the preceding example. These weights happen to be the rescaled orthogonal polynomial weights for testing quadratic trend. The remaining sources of between-groups variation may be predicted from any orthogonal complement of the quadratic trend contrast weights. Consequently, if we construct the vectors with columns of repeated linear and cubic contrast weights, the partial correlation between **y** and the contrast weights with the quadratic and cubic weights partialed out is $r^2_{\text{contrast}}$, which may also be computed directly from the standard *F* statistic for the contrast as

$$r^2_{\text{contrast}} = \frac{F_{\text{contrast}}}{F_{\text{contrast}} + df_{\text{within}}}. \tag{48}$$

Rosenthal et al. (2000) did not discuss sampling theory for $r^2_{\text{contrast}}$. However, a population equivalent, $\rho^2_{\text{contrast}}$, may be defined, and it may be shown (see Appendix B) that, with *p* groups in the analysis,

$$F_{\text{contrast}} = \frac{r^2_{\text{contrast}}}{(1 - r^2_{\text{contrast}})/(N_{\text{tot}} - p)} \tag{49}$$

has a noncentral *F* distribution with 1 and $N_{\text{tot}} - p$ degrees of freedom and noncentrality parameter

$$\lambda = N_{\text{tot}} \frac{\rho^2_{\text{contrast}}}{1 - \rho^2_{\text{contrast}}}. \tag{50}$$

Consequently, one may construct a confidence interval for $\rho^2_{\text{contrast}}$ by computing a confidence interval for $\lambda$ and transforming the endpoints, using the result of Equation 37.

*Example 9:* Consider again the data in Table 2. We can compute the *F* statistics corresponding to linear, quadratic, and cubic trend and, for each trend, compute confidence intervals for $\rho^2_{\text{contrast}}$ and/or $\rho_{\text{contrast}}$. For example, consider the test for linear trend. The *F* statistic is 216, with 1 and 8 degrees of freedom, and the 95% confidence interval for the noncentrality parameter $\lambda$ has endpoints of 54.2497 and 483.8839. Consequently, from Equation 37, a 95% confidence interval for $\rho^2_{\text{contrast}}$ has endpoints of

$$\rho^2_{\text{lower}} = \frac{54.2497}{54.2497 + 12} = .819, \rho^2_{\text{upper}} = \frac{483.8839}{483.8839 + 12} = .976 \tag{51}$$

The confidence interval for $\rho_{\text{contrast}}$ (defined as the square root of $\rho^2_{\text{contrast}}$, thus excluding negative values as in Rosenthal et al., 2000) ranges from .905 to .988.

Table 3 shows the results of computing contrast correlations and the associated confidence intervals for linear, quadratic, and cubic trend. Some brief comments are in order. Note, first, that although the $r_{\text{contrast}}$ values for quadratic and cubic trends are appealingly high, the corresponding confidence intervals are quite wide and include zero. On the other hand, the confidence interval for the linear trend is very narrow.

## The Relationship Between Confidence Intervals and Hypothesis Tests—Choosing the Appropriate Interval

Confidence intervals on measures of effect size convey all the information in a hypothesis test, and more. If one selects an appropriate confidence interval, a hypothesis test may be performed simply by inspection. If the confidence interval excludes the null hypothesized value, then the null hypothesis is rejected.

In such applications, I recommend using the traditional two-sided confidence interval, rather than a one-sided interval (or confidence bound), regardless of whether the hypothesis test is one-sided or two-sided. When a two-sided confidence interval is used to perform the hypothesis test, the confidence level must be matched appropriately both to the type of hypothesis test and to the Type I error rate. Recall that the endpoints of the two-sided confidence interval for a parameter $\theta$ at the $100(1 - \alpha)\%$ confidence level are the values of $\theta$ that place the observed statistic $\hat{\theta}$ at the $\alpha/2$ or $1 - \alpha/2$ cumulative probability point. Suppose the upper and lower limits of the $100(1 - \alpha)\%$ confidence interval are $U$ and $L$, respectively. Then $\hat{\theta}$ is the rejection point at the $\alpha/2$ significance level for one-sided hypothesis tests that $\theta$ is, first, greater than or equal to $U$ and, second, less than or equal to $L$. The observed statistic $\hat{\theta}$ is also equal to (a) the upper rejection point for a two-sided test that $\theta = L$ at the alpha level and (b) the lower rejection point for the two-sided test that $\theta = U$ at the alpha level. Consequently, the endpoints of the confidence interval represent two values of $\theta$ that the observed statistic would barely reject in a two-sided test with significance level alpha. These endpoints are also appropriate for testing one-sided hypotheses at the $\alpha/2$ significance level.

Table 3
*Confidence Intervals (CIs) for Contrast Correlations*

| Statistic | Linear | Quadratic | Cubic |
|---|---|---|---|
| $F$ | 216.00 | 3.00 | 2.40 |
| $r^2_{\text{contrast}}$ | .964 | .273 | .231 |
| CI | .819–.976 | 0–.541 | 0–.520 |
| $r_{\text{contrast}}$ | .982 | .522 | .480 |
| CI | .905–.988 | 0–.741 | 0–.721 |

The preceding paragraph implies a general rule of thumb: to use the confidence intervals to test a statistical hypothesis and to maintain a Type I error rate at alpha:

1. When testing a two-sided hypothesis at the alpha level, use a $100(1 - \alpha)\%$ confidence interval.

2. When testing a one-sided hypothesis at the alpha level, use a $100(1 - 2\alpha)\%$ confidence interval.

*Example 10:* Consider a test of the hypothesis that $\Psi = 0$, that is, that the RMSSE (as defined in Equation 12) in an ANOVA is zero. This hypothesis test is one-sided, because the RMSSE cannot be negative. To use a two-sided confidence interval to test this hypothesis at the $\alpha = .05$ significance level, one should examine the $100(1 - 2\alpha)\% = 90\%$ confidence interval for $\Psi$. If the confidence interval excludes zero, the null hypothesis will be rejected. This hypothesis test is equivalent to the standard ANOVA $F$ test.

*Example 11:* Consider the test that the standardized effect size $E_s$ in Equation 45 is precisely zero. This hypothesis test is two-sided, because $E_s$ can be either positive or negative. Consequently, to use a confidence interval to test this hypothesis at the .05 level, a $100(1 - \alpha)\% = 95\%$ two-sided confidence interval should be used, and the null hypothesis rejected only if both ends of the confidence interval are above zero or if both are below zero.

*Example 12:* Consider a situation in which one wishes to establish that the standardized effect size $E_s$ in Equation 45 is small, and that smallness is defined as an absolute value less than 0.20. To establish smallness, one must reject a hypothesis that $E_s$ is not small. Because $E_s$ can be either positive or negative, $E_s$ can be not small in two directions. The hypothesis that $E_s$ is not small can therefore be tested with two simultaneous one-sided hypothesis tests,

$$H_{01}\colon E_s \leq -0.20 \text{ versus } H_{a1}\colon E_s > -0.20 \qquad (52)$$

and

$$H_{02}\colon E_s \geq 0.20 \text{ versus } H_{a2}\colon E_s < 0.20. \qquad (53)$$

These two hypotheses can both be tested simultaneously at the .05 level by constructing a 90% confidence interval and observing whether the lower end of the interval is above $-0.20$ (to test the first one-sided hypothesis) and the upper end of the interval is below 0.20. What this amounts to is observing whether the entire interval is between $-0.20$ and 0.20. If so, the hypothesis that $E_s$ is not small is rejected, and smallness is indicated.

## Tests of Minimal Effect

### *Rationale and Method*

In many situations, the null hypothesis of zero effect is inappropriate or can be misleading. For example, in R-S testing with extremely large sample sizes, a null hypothesis may be rejected consistently, with a very low probability level, even when the population effect is small. Conversely, in A-S testing, the nil hypothesis of zero effect is often unreasonable, and the hypothesis the experimenter probably wants to test is that the effect is trivial.

Tests of minimal effect are a partial solution to the problems caused by inappropriate testing of a nil hypothesis when the goal is to show that an effect is small. For example, if some "minimal reasonable" effect size can be specified, rejection of the hypothesis that the effect is less than or equal to this value is of practical importance whether or not the sample size is very large. In the traditional A-S situation, in which the experimenter is trying to show that an effect is trivial, the hypothesis that the effect is greater than or equal to a minimal reasonable value can be tested. Serlin and Lapsley (1993) discussed this latter notion in detail and gave numerical examples. In such cases, large sample size will work for, rather than against, the experimenter, because if the effect size is truly below a level that is of practical import, larger samples will yield greater power to demonstrate that fact by rejecting the null hypothesis that the effect is at or above a point of triviality.

The confidence intervals described in the preceding section can be used to test hypotheses of minimal effect: One simply observes whether the appropriately constructed confidence interval contains the target minimal reasonable value. For example, suppose you decide that an RMSSE of 0.25 constitutes a minimal reasonable effect. In other words, effects below that level may be ignored. Effects that are definitely above that level are nontrivial. If you wish to demonstrate that effects are trivial, you might test the hypotheses

$$H_0: \Psi \geq 0.25; H_1: \Psi < 0.25. \qquad (54)$$

On the other hand, if you wish to demonstrate that effects are definitely not trivial, you might test the hypotheses

$$H_0: \Psi \leq 0.25; H_1: \Psi > 0.25. \qquad (55)$$

In each case, rejecting the null hypothesis will support the goal in performing the test, and the problems inherent in A-S testing can be avoided.

A simple approach to simultaneously testing the two hypotheses discussed above is to examine the $1 - 2\alpha$ confidence interval for $\Psi$ and see if it excludes 0.25. If the entire confidence interval is above the point of triviality (i.e., 0.25), then the effect may be judged nontrivial. If the entire confidence interval is below the point of triviality, then the effect has been shown to be trivial. There is a strong similarity between using the effect size confidence interval in this way and the long tradition of bioequivalence testing.

*Example 13:* Suppose you have $p = 6$ groups and $n = 75$ per group. You observe an $F$ statistic of $F(5, 444) = 2.28$, with $p = .046$, so the nil hypothesis of zero effects is rejected at the .05 significance level. However, on substantive grounds, you have decided that a value of $\Psi$ less than 0.25 can be ignored. To demonstrate triviality, you would attempt to reject the null hypothesis that $\Psi$ is greater than or equal to 0.25.

There are two approaches to performing the test. The first approach requires only a single calculation from the noncentral $F$ distribution. Consider the cutoff value of 0.25. Using the result of Equation 12, one may convert this to a value for $\lambda$ via the formula $\lambda = (p - 1)n\Psi^2 = (6 - 1)(75)(.25^2) = 23.4375$. The observed $F$ statistic of 2.28 has a one-sided probability value of .0256 in the noncentral $F$ distribution with $\lambda = 23.4375$, and 5 and 444 degrees of freedom, so the null hypothesis is rejected at the .05 level, and the overall effects are declared trivial.

An alternative approach uses the confidence interval. Note that, because the test is one-sided, we use the 90% confidence interval. The endpoints of the interval for $\lambda$ are 0.1028 and 20.3804. Using the result of Equation 12, we convert this confidence interval into a confidence interval for $\Psi$ by dividing the above endpoints by $(p - 1)n = 375$, then taking the square root. The resulting endpoints for the confidence interval for $\Psi$ are 0.0166 and 0.2331. This confidence interval excludes 0.25, so we can reject the hypothesis that effects are nontrivial, that is, $\Psi \geq 0.25$, at the .05 significance level. The advantage of using the confidence interval is that it provides us with an approximate indication of the precision of the estimation process while still allowing us to perform the hypothesis test.

Significant technical and theoretical issues surround the use of confidence intervals in this manner.

1. The choice of a numerical "point of triviality" for a measure of omnibus effect size should not be treated as a mechanical selection from a small menu of "approved" choices. Rather, it should be considered carefully on the basis of the specific experimental design and the substantive aspects of the variables being measured and manipulated. Whereas .25 might be considered trivial in one experiment, it might be considered very important in another.

2. The power of both hypothesis tests must be analyzed a priori to assess whether sample size is adequate. With low precision (i.e., a wide confidence interval), one might still have high power to demonstrate nontrivial effects if effects are large. However, it is virtually impossible to demonstrate triviality if precision is low, because the triviality point will be close to zero, and a wide confidence interval will not fit between zero and the triviality point.

Full consideration of the technical aspects of estimating the point of triviality, and precision of a parameter estimate and the resulting confidence interval, is beyond the scope (and length restrictions) of this article. However, in the next section, I discuss several theoretical issues that the sophisticated user should keep in mind.

## Conclusions and Discussion

This article demonstrates that the $F$ statistic in ANOVA contains information about standardized effect size, and its precision of estimation, that has not been made available in typical social science reports and is not reported by traditional software packages. Yet this information can readily be calculated, using a few basic techniques.

The fact is, simply reporting an $F$ statistic, and a probability level attached to a hypothesis of nil effect, is so suboptimal that its continuance can no longer be justified, at least in a social science tradition that prides itself on empiricism. A number of the field's most influential commentators on social statistics have emphasized this and urged that, as researchers, we revise our approach to reporting the results of significance tests (e.g., see articles in Harlow, Mulaik, & Steiger, 1997).

Null hypothesis testing is the source of much controversy. I have tried to promote an eclectic, integrated point of view that resists the temptation to downgrade either the hypothesis testing or the interval estimation approaches and emphasizes how they complement each other. Reviewers and other readers of the article have provided much food for thought and have raised several substantive criticisms that enriched my point of view considerably. In the following sections, I discuss some of the limitations of the procedures in this article, deal explicitly with several of the more common objections to my major suggestions, and then summarize my point of view and present some conclusions.

### *Statistical Limitations and Extensions of the Present Procedures*

The procedures discussed in this article provide exact distributional results under standard ANOVA assumptions (independence, normality, and equal variances) and are easily calculated with modern software. However, they are restricted to (a) completely between-subjects fixed-effect ANOVA with (b) equal $n$ per cell. The present article does not present procedures for dealing with the complications that result from unbalanced designs and/or repeated measures, nor does it discuss extensions to random effects or mixed ANOVA models or to multivariate analyses.

In some cases, procedures for these other situations are already available. Consider, for example, the case of one-way random effects ANOVA. The treatment effects are random variables with a variance of $\sigma_{\mathrm{A}}^2$, and $\Psi$ may be redefined as $\sigma_{\mathrm{A}}/\sigma$. A $100(1 - \alpha)\%$ confidence interval for $\Psi$ may therefore be obtained in the equal $n$ case by taking the square root of the well-known (Glass & Hopkins, 1996, p. 542) confidence interval for $\sigma_{\mathrm{A}}^2/\sigma^2$. One obtains, with $p$ groups,

$$\Psi_{\text{lower}} = \sqrt{\max\left[n^{-1}\left(\frac{F_{\text{obs}}}{F^*_{\alpha/2}} - 1\right), 0\right]}, \Psi_{\text{upper}} = \sqrt{\max\left[n^{-1}\left(\frac{F_{\text{obs}}}{F^*_{1-\alpha/2}} - 1\right), 0\right]}. \quad (56)$$

$F_{\text{obs}}$ is the observed value of the $F$ statistic, and $F^*$ is the percentage point from the $F$ distribution with $p - 1$ and $p(n - 1)$ degrees of freedom.

This approach can be generalized to more complicated designs. Burdick and Graybill (1992) discussed general methods for obtaining exact confidence intervals for $\Psi$ and related quantities in random effects models, both in the equal $n$ and unbalanced cases. Computational procedures for the unbalanced case are much more complicated than for the case of equal $n$.

However, on close inspection, some extensions yield challenging complications that require careful analysis. Some examples are as follows.

1. In the unbalanced, fixed-effects case, the noncentrality parameter $\lambda$ is defined as follows:

$$\lambda = \sum_{j=1}^{p} n_j\left(\frac{\alpha_j}{\sigma}\right)^2. \quad (57)$$

Note that with $\mu$ defined as in Equation 9, the quantity $f^2$ as defined in Equation 10 represents the ratio of between-groups to within-group variance in a population with probability of membership in the treatment groups proportional to the sample sizes in the ANOVA. There are situations in which this quantity is of interest (such as when the sampling plan reflects the relative size of natural subpopulations) and others in which it might not be. Cohen (1988, pp. 359–361) discussed this point in detail.

2. In repeated measures ANOVA designs, the noncen-

trality parameter λ unfortunately confounds effects of treatments with the correlation among observations. For example, in a one-way within-subjects design, if the data possess compound symmetry, the noncentrality parameter is

$$\lambda = \frac{n}{1-\rho}\sum_{j=1}^{p}\left(\frac{\mu_j-\mu}{\sigma}\right)^2. \tag{58}$$

The RMSSE, Ψ, as defined in Equation 12, though still an appropriate measure of effect size, cannot be estimated directly using the exact techniques discussed in this article, unless $\rho$ is known. For a detailed discussion of this issue in the context of point estimation in meta-analysis, see Dunlap, Cortina, Vaslow, and Burke (1996).

3. In multivariate analysis, the noncentrality parameter includes information about the variances and correlations of the dependent variables. For example, when two populations are compared on *k* dependent variables, using Hotelling's $T^2$ with two independent samples of size $n_1$ and $n_2$, the standard *F* statistic has *k* and $n_1 + n_2 - k - 1$ degrees of freedom and has a noncentral *F* distribution with a noncentrality parameter λ that is a simple function of the squared population Mahalanobis distance $\Delta^2$:

$$\lambda = \frac{n_1 n_2}{n_1 + n_2}\Delta^2. \tag{59}$$

The latter, computed as

$$\Delta^2 = (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) \tag{60}$$

with $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ the population mean vectors, and $\boldsymbol{\Sigma}$ the common covariance matrix, may be described as a sum of squared orthogonalized and standardized mean differences. Consequently, a natural analogue of Equation 12 that takes into account the number of dependent variables is

$$\Psi = \sqrt{\frac{\Delta^2}{k}}. \tag{61}$$

A confidence interval on $\Delta^2$ may be calculated easily (Reiser, 2001) from a confidence interval on λ, using the results of Equation 59. This interval may, in turn, be transformed into a confidence interval on Ψ using Equation 61.

We see that in one of the contexts discussed above (repeated measures), dependencies between measures are an annoying confound that must be removed from consideration. In another (the case of two populations), they are an essential ingredient for proper evaluation of effect size. In some of the problematic cases discussed above, and in situations where the standard ANOVA statistical assumptions are inappropriate, resampling methods such as bootstrapping can be used to obtain appropriate confidence intervals.

The width of a confidence interval often is described as indicating precision of measurement. However, as Steiger and Fouladi (1997, pp. 254–255) pointed out, this relationship is less than perfect and is seriously compromised in some situations for several reasons. The width of a confidence interval is itself a random variable and is subject to sampling variations. Moreover, the confidence intervals are truncated at zero to avoid improper estimates. In extreme cases, a confidence interval might actually have 0 as both endpoints. This zero-width confidence interval obviously does not imply that effect size was determined with perfect precision.

### *Focused Contrasts or Omnibus Hypotheses?*

In an early version of this article, I concentrated almost exclusively on omnibus measures of effect size. Several reviewers have objected that confidence intervals on measures of standardized effect size such as $\omega^2$ and the RMSSE were, to paraphrase, an elegant solution to the wrong problem. These writers have echoed the view of Rosenthal et al. (2000), who stated that "omnibus questions seldom address questions of real interest to researchers, and are typically less powerful than focused procedures" (p. 1).

I share an enthusiasm for focused contrasts and recommend them in lieu of an omnibus test whenever researchers have clear ideas about linear contrasts. Moreover, I believe that not enough researchers have been trained to look carefully for ways to phrase their ideas as contrasts. However, I think that dismissing the improvements to the use of the *F* statistic suggested in this article ignores several important realities.

First, much research in the social sciences is exploratory, and an omnibus *F* test in such circumstances may be the prelude to subsequent examination of unplanned contrasts. In such cases, an overall measure of the strength of effect sizes, and the precision with which they have been determined, may alert the researcher in advance to a lack of overall precision in the experimental design. Second, when one is comparing several studies that have reported overall *F* tests, comparing confidence intervals on standardized effect size measures can be very useful in resolving apparent disparities in experimental outcomes.

As it turns out, the confidence interval on $\rho^2_{\text{contrast}}$, one standardized measure of omnibus effect size, is closely related, conceptually and computationally, to the procedure for computing a confidence interval on $\omega^2$. The latter index examines the squared multiple correlation between observed data and a set of contrast weights, whereas the former examines the squared correlation between the data and one set of contrast weights with the variation predicted by the complementary contrasts partialed out. Thus, the same technology that I find useful for omnibus tests may be applied directly to contrasts.

I believe that reporting an exact confidence interval on $\rho_{\text{contrast}}$ is substantially more informative than simply reporting the raw coefficient. And, to be clear, I fully support concentration on focused contrasts in lieu of omnibus tests whenever the experimenter has firm questions that suit the contrast analysis framework.

### *Some Recent Objections to Standardized Measures of Effect Size*

Revised hypothesis-testing strategies for ANOVA require specification of target values of a standardized measure of effect size. The confidence interval approach is more relaxed but strongly tends to lead the experimenter to consider which overall effect sizes qualify as trivial and which are nontrivial in a particular application.

Although many writers have emphasized the value of standardized measures of effect size in power analysis and sample size estimation, standardized effect size measures do have some shortcomings. As a nonlinear combination of several sources of variation in an experiment, they reduce several values into one and are of necessity less precise than similar indices computed on a focused contrast. Moreover, ANOVA effects as used in the calculation of the noncentrality parameter $\lambda$ in the omnibus test may or may not correspond to experimental effects as commonly conceptualized (see, e.g., Steiger & Fouladi, 1997, pp. 244–245), and focused contrasts can get at such experimental effects much more effectively than an omnibus procedure. Recently, Lenth (2001) suggested dispensing with standardized measures of effect size altogether in the context of power analysis and sample size estimation. His main justification was that combining information about raw effects (i.e., mean differences) and variation ignored a possible confounding impact of reliability of measurement.

### *Reconciling the Interval Estimation and Minimal-Effect-Testing Approaches*

As stated at the outset, this article discusses two major approaches that might be used to replace the traditional $F$ test in ANOVA. The noncentrality interval estimation approach emphasizes estimation of some function of overall effect size, along with an indication of the precision of the measurement. The dual hypothesis testing approach replaces the hypothesis of nil effect with two hypotheses, one that the effect is trivial, the other that it is nontrivial.

The approach I personally favor is confidence interval estimation on some standardized measure of overall effect size. This approach may be viewed as replacing hypothesis testing entirely, yet it can be used to perform both kinds of hypothesis tests required by the dual hypothesis-testing framework. Specifically, one simply examines, simultaneously, whether the confidence interval excludes a trivial effect value on the left or right. If, for example, the confidence interval lies entirely above the cutoff point for a trivial effect, one rejects the hypothesis of triviality. If the confidence interval lies entirely below the cutoff point, one rejects the hypothesis of nontriviality.

Moreover, the confidence interval approach, being an exact procedure, also provides all the information available in the standard $F$ test. For example, the $F$ test results in rejection at the .05 level if and only if the 90% confidence interval for $\Psi$ excludes zero.

The hypothesis-testing approach offers advantages as well. For one, it keeps the analysis within the comfortably familiar bounds of hypothesis testing. For another, it is computationally easier—one may perform the hypothesis test without extensive iteration, and so it may be performed with a wider range of available free software. Another advantage is that, by simultaneously analyzing power for both a test of triviality and a test of nontriviality, the user can be relatively certain that the confidence interval, if calculated, will have enough precision to determine whether effects are trivial or not.

### *Standardized Effects and Coefficients of Determination—A Caution*

Any statistical technique offers opportunity for abuse and misuse, especially if the technique is used mechanically and without taking into account the special circumstances surrounding a particular set of data. Abelson (1995) discussed in detail how important it is to remain open-minded when judging the importance of effect sizes. In some cases, effects that seem small may be quite important. This should be kept in mind before effects that are nonzero, but seemingly trivial, are dismissed. Abelson's comments are similar to Cohen's (1988, pp. 534–535) in his chapter on special issues in power analysis.

### *Casting a Vote for Change*

A fundamental contribution to behavioral statistics by Cohen (1962) was to demonstrate that many studies lack sufficient statistical power. The initial emphasis on power analysis spearheaded by Cohen (1962) has now given way to a more sophisticated emphasis on precision of estimation.

Confidence intervals on standardized measures of effect size allow one to assess how precisely effects have been measured and simultaneously assess whether the experiment has ruled out (a) the notion that effects are trivial and (b) the notion that they are nontrivial. The procedures are straightforward and offer obvious benefits. It is time for a change. Yet there are numerous obstacles to change in behavioral statistics practice. A significant obstacle is the dominant influence a few commercial statistical packages such as SPSS and SAS have on practice in the field. The way psychology has operated in the past, procedures are unlikely to be used until they have been implemented in a widely used statistics package, and commercial statistics packages tend to be conservative toward new approaches.

In the final analysis, the impetus for change may have to come from journal editors and practitioners, some of whom have resisted change for a variety of reasons discussed by Thompson (1999). Fortunately, the Internet makes it possible to distribute innovative software to practitioners very easily at virtually zero cost. There is no longer any reason to report a squared multiple correlation, an ANOVA *F* statistic, or a focused contrast *t* test without providing information about confidence intervals on standardized effects. Each reader of this article can cast votes for change by obtaining the freeware I (and other authors) have made available, and then, when reviewing articles that report omnibus tests and focused contrasts without associated intervals, taking two simple steps: (a) performing their own calculation of confidence intervals on standardized effect size and (b) requesting that the author include this information in the published article.

## References

Abelson, R. P. (1995). *Statistics as principled argument.* Hillsdale, NJ: Erlbaum.

Burdick, R. K., & Graybill, F. A. (1992). *Confidence intervals on variance components.* New York: Dekker.

Casella, G., & Berger, R. L. (2002). *Statistical inference* (2nd ed.). Pacific Grove, CA: Duxbury.

Chow, S.-C., & Liu, J.-P. (2000). *Design and analysis of bioavailability and bioequivalence studies.* New York: Dekker.

Cohen, J. (1962). The statistical power of abnormal–social psychological research. *Journal of Abnormal and Social Psychology, 65,* 145–153.

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

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

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

Fleishman, A. E. (1980). Confidence intervals for correlation ratios. *Educational and Psychological Measurement, 40,* 659–670.

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

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

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

Lenth, R. V. (2001). Some practical guidelines for effective sample size determination. *American Statistician, 55,* 187–193.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods, 1,* 130–149.

Mels, G. (1989). *A general system for path analysis with latent variables.* Unpublished master's thesis, University of South Africa, Pretoria, South Africa.

Metzler, C. M. (1974). Bioavailability: A problem in equivalence. *Biometrics, 30,* 309–332.

Ozer, D. J. (1985). Correlation and the coefficient of determination. *Psychological Bulletin, 97,* 307–315.

Reiser, B. (2001). Confidence intervals for the Mahalanobis distance. *Communications in Statistics, Simulation and Computation, 30,* 37–45.

Rosenthal, R. (1991). Effect sizes: Pearson's correlation, its display via the BESD, and alternative indices. *American Psychologist, 46,* 1086–1087.

Rosenthal, R., Rosnow, R. L., & Rubin, D. B. (2000). *Contrasts and effect sizes in behavioral research: A correlational approach.* New York: Cambridge University Press.

Rosnow, R. L., & Rosenthal, R. (1996). Computing contrasts, effect sizes, and counternulls on other people's published data: General procedures for research consumers. *Psychological Methods, 1,* 331–340.

Sampson, A. R. (1974). A tale of two regressions. *Journal of the American Statistical Association, 69,* 682–689.

Schmidt, F. L. (1996). Statistical significance testing and cumulative research in psychology: Implications for the training of researchers. *Psychological Methods, 1,* 115–129.

Schmidt, F. L., & Hunter, J. E. (1997). Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 38–64). Mahwah, NJ: Erlbaum.

Searle, S. R. (1987). *Linear models for unbalanced data.* New York: Wiley.

Serlin, R. A., & Lapsley, D. K. (1993). Rational appraisal of psychological research and the good-enough principle. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 199–228). Hillsdale, NJ: Erlbaum.

Smithson, M. (2001). Correct confidence intervals for various regression effect sizes and parameters: The importance of noncentral distributions in computing intervals. *Educational and Psychological Measurement, 61,* 603–630.

Steiger, J. H. (1989). *EzPATH: A supplementary module for SYSTAT and SYGRAPH.* Evanston, IL: Systat.

Steiger, J. H. (1990, October). *Noncentrality interval estimation and the evaluation of statistical models.* Paper presented at the meeting of the Society of Multivariate Experimental Psychology, Kingston, RI.

Steiger, J. H. (1999). *STATISTICA power analysis.* Tulsa, OK: StatSoft.

Steiger, J. H., & Fouladi, R. T. (1992). R2: A computer program for interval estimation, power calculation, and hypothesis testing for the squared multiple correlation. *Behavior Research Methods, Instruments, and Computers, 4,* 581–582.

Steiger, J. H., & Fouladi, R. T. (1997). Noncentrality interval estimation and the evaluation of statistical models. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 221–257). Mahwah, NJ: Erlbaum.

Steiger, J. H., & Lind, J. C. (1980, May). *Statistically based tests for the number of factors.* Paper presented at the meeting of the Psychometric Society, Iowa City, IA.

Steiger, J. H., & Ward, L. M. (1987). Factor analysis and the coefficient of determination. *Psychological Bulletin, 99,* 471–474.

Taylor, D. J., & Muller, K. E. (1995). Computing confidence bounds for power and sample size of the general linear univariate model. *The American Statistician, 49,* 43–47.

Taylor, D. J., & Muller, K. E. (1996). Bias in linear model power and sample size calculation due to estimating noncentrality. *Communications in Statistics: Theory and Methods, 25,* 1595–1610.

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

Venables, W. (1975). Calculation of confidence intervals for noncentrality parameters. *Journal of the Royal Statistical Society, Series B, 37,* 406–412.

Westlake, W. J. (1976). Symmetrical confidence intervals for bioequivalence trials. *Biometrics, 32,* 741–744.

# Appendix A

## The Relationship Between $\rho^2$ and $\lambda$ in One-Way ANOVA

Define, for the $p$ sample means,

$$s_{\bar{x}\cdot}^2 = \frac{1}{p-1}\sum_{j=1}^{p}(\bar{x}_{\cdot j} - \bar{x}_{\cdot\cdot})^2. \tag{A1}$$

The corresponding population quantity is

$$s_{\mu}^2 = \frac{1}{p-1}\sum_{j=1}^{p}(\mu_j - \bar{\mu}_{\cdot})^2 = \frac{\boldsymbol{\beta}'\mathbf{X}'\mathbf{Q}_1\mathbf{X}\boldsymbol{\beta}}{n(p-1)}, \tag{A2}$$

where $\boldsymbol{\beta}$, $\mathbf{X}$, and $\mathbf{Q_1}$ are as described in Equations 29 through 38. In a balanced, one-way ANOVA, with $p$ groups and $n$ observations per group, $SS_{\text{treatments}} = n(p-1)s_{\bar{x}\cdot}^2$.

Consider any estimator $\hat{\theta}$ of a parameter $\theta$. The *probability limit* of $\hat{\theta}$, denoted $\text{plim}(\hat{\theta})$, is equal to a value $c$ if and only if for any error tolerance $\delta > 0$, we have

$$\lim_{n\to\infty} Pr(|\hat{\theta} - c| < \delta) = 1.$$

The notion of a probability limit is closely related to that of consistency, in that $\hat{\theta}$ is a consistent estimator for $\theta$ if and only if $\text{plim}_{n\to\infty}(\hat{\theta}) = \theta$. In what follows, for brevity of notation, I simply write $\text{plim}(X)$ rather than $\text{plim}_{n\to\infty}(X)$. I use a number of well-known results. In particular, if $\text{plim}(X)$ and $\text{plim}(Y)$ exist, then

$$\text{plim}(X+Y) = \text{plim}(X) + \text{plim}(Y),$$

$$\text{plim}(X/Y) = \text{plim}(X)/\text{plim}(Y),$$

$$\text{plim}(XY) = \text{plim}(X)\text{plim}(Y).$$

Moreover, the plim of a sample moment is equal to the corresponding population moment. We define $\rho^2$ as $\text{plim}(R^2)$, that is, the value that $R^2$ converges to in an infinite population. Then

$$\begin{aligned}
\rho^2 = \text{plim}(R^2) &= \text{plim}\left(\frac{SS_{\text{treatments}}}{SS_{\text{treatments}} + SS_{\text{error}}}\right) \\
&= \frac{\text{plim}[n(p-1)s_{\bar{x}\cdot}^2]}{\text{plim}[n(p-1)s_{\bar{x}\cdot}^2] + \text{plim}[p(n-1)MS_{\text{error}}]} \\
&= \frac{\text{plim}(s_{\bar{x}\cdot}^2)}{\text{plim}(s_{\bar{x}\cdot}^2) + \lim_{n\to\infty}\left[\dfrac{p(n-1)}{(p-1)n}\right]\text{plim}(MS_{\text{error}})} \\
&= \frac{s_\mu^2}{s_\mu^2 + \dfrac{p}{p-1}\sigma^2} \\
&= \frac{(p-1)s_\mu^2}{(p-1)s_\mu^2 + p\sigma^2}.
\end{aligned} \tag{A3}$$

Combining Equations A1 through A3, we obtain

$$\rho^2 = \frac{\dfrac{\boldsymbol{\beta}'\mathbf{X}'\mathbf{Q}_1\mathbf{X}\boldsymbol{\beta}}{n}}{\dfrac{\boldsymbol{\beta}'\mathbf{X}'\mathbf{Q}_1\mathbf{X}\boldsymbol{\beta}}{n} + p\sigma^2} \tag{A4}$$

and

$$N_{\text{tot}}\frac{\rho^2}{1-\rho^2} = N_{\text{tot}}\frac{\boldsymbol{\beta}'\mathbf{X}'\mathbf{Q}_1\mathbf{X}\boldsymbol{\beta}}{np\sigma^2} = \frac{\boldsymbol{\beta}'\mathbf{X}'\mathbf{Q}_1\mathbf{X}\boldsymbol{\beta}}{\sigma^2} = \lambda, \tag{A5}$$

where $\lambda$ is as defined in Equation 34.

# Appendix B

## The Distribution of the *F* Statistic for $r^2_{\mathrm{contrast}}$

Assume the general linear model as described in Equations 29 and 30. For any full column rank matrix $\mathbf{A}$, define $\mathbf{P_A} = \mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'$, and $\mathbf{Q_A} = \mathbf{I} - \mathbf{P_A}$, with $\mathbf{I}$ a conformable identity matrix. Define $\mathbf{1}$ to be a column of 1s. Partition $\mathbf{X}$ as $\mathbf{X} = [\mathbf{1} \quad \mathbf{x}_1 \quad \mathbf{X}_2]$. $\mathbf{x}_1$ contains replications of the contrast weights for the contrast being evaluated, so that the *i*th value in $\mathbf{x}_1$ is the contrast weight for the group that $y_i$ is in, and $\mathbf{X}_2$ contains a set of columns that are the orthogonal complement of the contrast weights in $\mathbf{x}_1$. Thus, for example, if $\mathbf{x}_1$ contains contrast weights for evaluating linear trend, $\mathbf{X}_2$ would contain quadratic and cubic contrast weights (or some full rank transformation of them). The regression weight vector $\boldsymbol{\beta}$ is partitioned accordingly as

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \boldsymbol{\beta}_2 \end{bmatrix}. \qquad \text{(B1)}$$

Define $\boldsymbol{\mu}$ as a vector of the population means of the *p* groups, and $\mathbf{c}$ as the linear weights for the contrast of interest. In this case, the contrast of interest is

$$\psi = \mathbf{c}'\boldsymbol{\mu}, \qquad \text{(B2)}$$

and because $\mathbf{x}_1$ contains *n* replications of the elements of $\mathbf{c}$, and $E(\boldsymbol{\eta})$ contains *n* replications of the elements of $\boldsymbol{\mu}$, we have

$$\mathbf{c}'\mathbf{c} = \mathbf{x}_1'\mathbf{x}_1/n \qquad \text{(B3)}$$

and

$$\beta_1 = \frac{n\psi}{\mathbf{x}_1'\mathbf{x}_1}. \qquad \text{(B4)}$$

I first demonstrate that an *F* statistic may be constructed for $r^2_{\mathrm{contrast}}$. Rosenthal et al. (2000) defined $r^2_{\mathrm{contrast}}$ as the squared partial correlation between $\mathbf{y}$ and $\mathbf{x}_1$ with $\mathbf{X}_2$ partialed out. This sample statistic can be computed as the following ratio of quadratic forms in $\mathbf{y}$:

$$r^2_{\mathrm{contrast}} = \frac{\mathbf{y}'\mathbf{P}_{\mathbf{x}_1}\mathbf{y}}{\mathbf{y}'(\mathbf{I} - \mathbf{P}_{\mathbf{X}_2} - \mathbf{P}_{\mathbf{1}})\mathbf{y}}. \qquad \text{(B5)}$$

Consider the statistic

$$F_{\mathrm{contrast}} = \frac{r^2_{\mathrm{contrast}}}{(1 - r^2_{\mathrm{contrast}})/(N_{tot} - p)} = \frac{\mathbf{y}'\mathbf{P}_{\mathbf{x}_1}\mathbf{y}}{\mathbf{y}'(\mathbf{I} - \mathbf{P}_{\mathbf{x}_1} - \mathbf{P}_{\mathbf{X}_2} - \mathbf{P}_{\mathbf{1}})\mathbf{y}/(N_{\mathrm{tot}} - p)}. \qquad \text{(B6)}$$

This statistic is a ratio of two quadratic forms, in the general form

$$\frac{\mathbf{y}'\mathbf{A}\mathbf{y}/a\sigma^2}{\mathbf{y}'\mathbf{B}\mathbf{y}/b\sigma^2}, \qquad \text{(B7)}$$

where $a = 1$, and $b = N_{\mathrm{tot}} - p$. From Searle (1987, pp. 233–234), $F_{\mathrm{contrast}}$ has a noncentral *F* distribution with *a* and *b* degrees of freedom and noncentrality parameter

$$\lambda_{\mathrm{contrast}} = \boldsymbol{\beta}'\mathbf{X}'\mathbf{A}\mathbf{X}\boldsymbol{\beta}/\sigma^2 \qquad \text{(B8)}$$

if $\mathbf{A}\sigma^2$ is idempotent, $\mathbf{B}\sigma^2$ is idempotent, $\mathbf{AB} = \mathbf{0}$, and *a* and *b* are the ranks of $\mathbf{A}$ and $\mathbf{B}$, respectively. These four properties are easily established by substitution and the fact that $\mathbf{x}_1$, $\mathbf{X}_2$, and $\mathbf{1}$ are pairwise orthogonal. The orthogonality implies that the noncentral *F* distribution has a noncentrality parameter equal to

$$\lambda_{\mathrm{contrast}} = \frac{\boldsymbol{\beta}'\mathbf{X}'\mathbf{P}_{\mathbf{x}_1}\mathbf{X}\boldsymbol{\beta}}{\sigma^2} = \frac{\beta_1^2\mathbf{x}_1'\mathbf{x}_1}{\sigma^2}, \qquad \text{(B9)}$$

and the ranks of the $\mathbf{A}$ and $\mathbf{B}$ are 1 and $N_{\mathrm{tot}} - p$, respectively. Next, we derive the relationship between $\lambda_{\mathrm{contrast}}$ and the population equivalent of $r^2_{\mathrm{contrast}}$.

We may write $r^2_{\mathrm{contrast}}$ as follows:

$$\begin{aligned} r^2_{\mathrm{contrast}} &= \frac{F_{\mathrm{contrast}}}{F_{\mathrm{contrast}} + p(n-1)} \\ &= \frac{n\hat{\psi}^2/MS_{\mathrm{error}}(\mathbf{c}'\mathbf{c})}{n\hat{\psi}^2/MS_{\mathrm{error}}(\mathbf{c}'\mathbf{c}) + p(n-1)} \\ &= \frac{\hat{\psi}^2}{\hat{\psi}^2 + (\mathbf{c}'\mathbf{c})MS_{\mathrm{error}}p(n-1)/n}. \end{aligned} \qquad \text{(B10)}$$

We define $\rho^2_{\mathrm{contrast}}$ as

$$\begin{aligned} \rho^2_{\mathrm{contrast}} &= \mathrm{plim}(r^2_{\mathrm{contrast}}) \\ &= \frac{\mathrm{plim}(\hat{\psi}^2)}{\mathrm{plim}(\hat{\psi}^2) + p(\mathbf{c}'\mathbf{c})\lim_{n\to\infty}\left(\dfrac{n-1}{n}\right)\mathrm{plim}(MS_{\mathrm{error}})} \\ &= \frac{\psi^2}{\psi^2 + p(\mathbf{c}'\mathbf{c})\sigma^2}. \end{aligned} \qquad \text{(B11)}$$

Hence,

$$\frac{\rho^2_{\mathrm{contrast}}}{1 - \rho^2_{\mathrm{contrast}}} = \frac{\psi^2}{p(\mathbf{c}'\mathbf{c})\sigma^2}, \qquad \text{(B12)}$$

*Appendix continues*

which, after substitution of Equations B3 and B4, becomes

$$\frac{\rho^2_{\text{contrast}}}{1-\rho^2_{\text{contrast}}} = \frac{(\mathbf{x}_1'\mathbf{x}_1)^2\beta_1^2/n^2}{(\mathbf{x}_1'\mathbf{x}_1)p\sigma^2/n}$$

$$= \frac{(\mathbf{x}_1'\mathbf{x}_1)\beta_1^2}{np\sigma^2}$$

$$= \frac{(\mathbf{x}_1'\mathbf{x}_1)\beta_1^2}{N_{\text{tot}}\sigma^2}. \qquad \text{(B13)}$$

Recalling the result of Equation B9 for $\lambda_{\text{contrast}}$, we have thus shown that

$$\lambda_{\text{contrast}} = \frac{(\mathbf{x}_1'\mathbf{x}_1)\beta_1^2}{\sigma^2} = N_{\text{tot}}\frac{\rho^2_{\text{contrast}}}{1-\rho^2_{\text{contrast}}}. \qquad \text{(B14)}$$

Received February 23, 2001
Revision received December 1, 2003
Accepted January 8, 2004 ■