APPROXIMATE CONFIDENCE INTERVALS FOR EFFECT SIZES

JAMES ALGINA
University of Florida

H. J. KESELMAN
University of Manitoba

This article defines an approximate confidence interval for effect size in correlated (repeated measures) groups designs. The authors found that their method was much more accurate than the interval presented and acknowledged to be approximate by Bird. That is, the coverage probability over all the conditions investigated was very close to the theoretical .95 value. By contrast, Bird's interval could have coverage probability that was substantially below .95. In addition, the authors' interval was less likely than Bird's method to present an overly optimistic portrayal of the effect. They also examined the operating characteristics of the Bird interval for effect size in an independent groups design and found that, although it is fairly accurate in its approximation of coverage probability, the accuracy of the approximation does vary with the magnitude of the population effect size.

Keywords: *effect size; confidence interval; between-subjects design; within-subjects design*

Bird (2002) has made a very valuable contribution to the literature by presenting confidence interval procedures for effect sizes in completely randomized and correlated groups designs. His article follows up on the American Psychological Association's (APA) Task Force on Statistical Inference, where recommendations were made that authors should present effect sizes and confidence intervals along with tests of significance (Wilkinson & the APA Task Force on Statistical Inference, 1999).

Educational and Psychological Measurement, Vol. 63 No. 4, August 2003 537-553
DOI: 10.1177/0013164403256358

We support the recommendations of the Task Force and the effort made by Professor Bird to bring to the attention of researchers methods for implementing these recommendations. Indeed, Bird's (2002) contribution is quite significant in terms of the generality of the approach he has presented. Because we intended to present his methods to our students, we wanted to verify the operating characteristics of his methods. In particular, because Bird pointed out that the methods he presents are approximate we wanted to be able to comment to our students on the degree of accuracy of the methods he presented. Accordingly, we examined the procedures he presented.

Approximate Confidence Intervals for Effect Sizes

For the case of two independent groups, the approximate interval used in Bird (2002) is

$$ES \pm t_{(1-\alpha/2,N-2)}\left(\frac{1}{n_1}+\frac{1}{n_2}\right) \tag{1}$$

where

$$ES = \frac{\bar{Y}_1 - \bar{Y}_2}{S_{Pooled}} \tag{2}$$

$\bar{Y}_j (j = 1, 2)$ is a treatment level group mean, $n_j (n_1 + n_2 = N)$ is the sample size for the jth group, $t_{(1-\alpha/2,\ N-2)}$ is a critical value from Student's t distribution, and S_{Pooled} equals $\sqrt{MSE}$, a value obtained from an analysis of variance of the data. Equation 1 provides an approximate $100(1 - \alpha)\%$ confidence interval for the population effect size

$$\delta = \frac{\mu_1 - \mu_2}{\sigma} \tag{3}$$

where μ_j is the population mean for level j and σ is the population standard deviation, which is assumed to be equal for the two levels of the factor.

The interval in Equation 1 is symmetric, and the length of the interval is independent of the population effect size. As Bird notes, when the groups are independent, the correct (exact) procedure, under normality, for obtaining an interval for the effect size uses the noncentral t distribution (e.g., for a presentation of this interval see Cumming & Finch, 2001; Steiger & Fouladi, 1997). Figure 5 in Cumming and Finch (2001) illustrates that the variance and skew of the noncentral t distribution increase as the population effect size increases. The increase in skew is particularly notable for small degrees of freedom. Thus, the length of the exact interval would not be independent of

the population effect size interval and, for small degrees of freedom, the interval would not be symmetric when the population effect size is large. Therefore, we would expect the interval used by Bird to become less accurate as the population effect size increases.

For the case of two dependent groups, the interval used in Bird (2002) is

$$\frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{\frac{S_1^2 + S_2^2}{2}}} \pm t_{(1-\alpha/2,\,n-1)}\sqrt{\frac{2(S_1^2 + S_2^2 - 2S_{12})}{n(S_1^2 + S_2^2)}} \tag{4}$$

where S_1^2 and S_2^2 are the variances of the scores for Levels 1 and 2 of the repeated measures variable, and S_{12} is the covariance between the scores at Levels 1 and 2. If $\sigma_1^2 = \sigma_2^2$, Equation 4 is an approximate confidence interval for the population effect size defined in Equation 3. Note that if $\sigma_1^2 \neq \sigma_2^2$, then Equation 4 is an approximate confidence interval for

$$\delta = \frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma_1^2 + \sigma_2^2}{2}}} \tag{5}$$

For the interval in Equation 4 to be correct for the effect size in Equation 3 or 5, the expression to the right of $t_{(1-\alpha/2,\,n-1)}$ should be the standard error of the estimated effect size and the square root of a multiple of a chi-square distributed variable. Both of these requirements would be met if the divisors on the left- and right-hand side of Equation 4 were fixed constants, but because S_1^2 and S_2^2 are variables, neither requirement is met and so we can expect the interval to be inaccurate to some degree.

As indicated in our introduction, we are strongly committed to the practice of constructing effect sizes and intervals for them and thus very much appreciate the contribution made by Bird to this literature. To help promote their use, we thought, however, it would be necessary to evaluate how accurate the approximations are.

Study 1

Method

In Study 1, we investigated coverage probability for Bird's approximate interval for a between-subjects design (see Equation 1). Sample size and population effect size were manipulated. Sample size was varied from 10 to 80 per group in steps of 10. The sample size covers a range from a small study to a fairly substantial study. Population effect size (PES) was varied from .00 to

1.60 in steps of .20 and covers the range from a null effect size to a very large effect size. The simulation was conducted by using the following steps:

1. For the first treatment, generate the mean $\overline{Y}_1$ as a normal random deviate with mean zero and variance σ^2/n and generate the variance S_1^2 as a multiple of a chi-square pseudo-random variable: $\sigma^2\chi^2_{n-1} / (n-1)$. In all cases, and without loss of generalizability, $\sigma^2 = 1$.
2. For the second treatment, generate the mean by using $\overline{Y}_2 = \overline{X} + \delta$ where $\overline{X}$ was a normal random deviate with mean zero and variance σ^2/n and δ is the population effect size. The mean $\overline{X}$ was generated independently of $\overline{Y}_1$ and therefore $\overline{Y}_1$ and $\overline{Y}_2$ were independent. Then generate S_2^2 as a multiple of a chi-square pseudo-random variable: $\sigma^2\chi^2_{n-1} / (n-1)$. In all cases, and without loss of generalizability, $\sigma^2 = 1$. The chi-square pseudo-random variable in Step 2 was generated independently of that in Step 1. Therefore, S_1^2 and S_2^2 were independently distributed.
3. Calculate the approximate confidence interval reported in Bird (2002).
4. Estimate coverage probability, proportion above, and proportion below as the proportion of 10,000 replications in which the interval in Equation 1 contained the population effect size, was entirely above, or was entirely below the population effect size.

Results

A preliminary analysis of our data indicated that there was very little effect due to sample size. For example, when the population effect size was .50, the coverage probability ranged from .939 to .943, and the variation was unrelated to the sample size. Accordingly, the values for coverage probability, proportion of intervals above the population effect size (% Above) and proportion of intervals below the population effect size (% Below) in Table 1 are averages taken over the sample size cases.

Table 1 contains our results for the approximate effect size interval due to Bird (2002) for the independent two-groups design. As we speculated in the introduction, the coverage probability for the effect size became increasingly inaccurate as the population effect size increased in value. Specifically, when PES = 0, the coverage probability was .95 (as expected), whereas for the largest PES considered (1.6), coverage probability equaled .911. That is, the empirical confidence coefficient was smaller than desired. One should also note from Table 1 that for all cases investigated, the approximate interval was above the true effect size more often than it was below the true effect size. This implies an unduly optimistic impression of the magnitude that the effect size can assume, that is, overestimating the true effect size. Although we have not reported our results by sample size, it should be noted that the tendency to exhibit optimistic results increases as the sample size decreases because the effect size estimator tends to be positively biased for small sample sizes (e.g., see Hedges, 1981).

Table 1
Average Estimated Coverage Probability for the Independent Groups Pairwise Comparison

PES	Coverage Probability	% Above	% Below
0.0	.950	.025	.025
0.2	.950	.028	.023
0.4	.948	.030	.023
0.6	.945	.033	.022
0.8	.941	.036	.023
1.0	.935	.041	.024
1.2	.928	.047	.026
1.4	.921	.051	.028
1.6	.911	.059	.030

Note. PES = population effect size; % Above/% Below = proportion above/below of 10,000 replications in which the interval contained the population effect size, was entirely above, or was entirely below the population effect size.

We replicated our simulation for a design with four independent groups. That is, the confidence interval for the effect size for comparing Groups 1 and 2 was the following:

$$ES \pm t_{(1-\alpha/2, N-4)}\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \tag{6}$$

where $N = n_1 + \mathrm{L} + n_4$. The results suggest that when the effect size is large (the condition in which the approximate effect size performed most poorly for a design with two groups), coverage probability improves as the number of independent groups in the study increases. For example, with a population effect size of 1.6, probability coverage was .929 in contrast to .911 when there were two groups.

Study 2

Method

In Study 2, we investigated coverage probability for Bird's approximate interval for a within-subjects design. For the within-subjects design, sample size, population effect size, and correlation between the two repeated measures were manipulated. Sample size and population effect size were varied over the levels used in the simulations of the between-subjects design. The correlation was varied from .00 to .80 in steps of .20 and covers the range from no correlation to a substantial correlation. The simulation was conducted by using the following steps:

1. For the first treatment, generate the mean $\bar{Y}_1$ as a normal random deviate with mean zero and variance σ^2/n.
2. For the second treatment, generate $\bar{\varepsilon}_1$ as a normal random deviate with mean zero and variance σ^2/n and generate

$$\bar{Y}_2 = \rho\bar{Y}_1 + \sqrt{(1-\rho^2)}\bar{\varepsilon}_1 + \delta.$$

   With this generation method, $\bar{Y}_1$ and $\bar{Y}_2$ are means for a sample of size n for two bivariate normal variables, each with variance of σ^2 and a correlation of ρ. As in the between-subjects design, $\sigma^2 = 1$. This specification, in and of itself, does not entail any loss of generality. However, the approximate procedure may have different operating characteristics if $\sigma_1^2 \neq \sigma_2^2$.
3. Bartlett's decomposition (Browne, 1969) was used to simulate the sample variances and the covariance. Specifically, let **T** be a (2×2) lower triangular matrix. The jth diagonal element of **T** is distributed as a chi variable with degrees of freedom $n - j + 1$. The nonzero off-diagonal element is distributed as a standard normal random variable. All random variables are independently distributed. Furthermore, let Σ be the population covariance matrix and let **A** be a Cholesky factor of Σ. Then, the (2×2) sample covariance matrix **S** was calculated by using $\mathbf{S} = \frac{\mathbf{ATT'A'}}{n}$. Because we specified $\sigma_1^2 = \sigma_2^2 = 1$, the population covariance matrix Σ had ones on the diagonal and ρ on the off-diagonal. However, the matrix **S** that results from the generation method does not have ones on the diagonal and is therefore a covariance matrix. In fact, it is a covariance matrix for a sample from a bivariate normal distribution with covariance matrix Σ.
4. Calculate the approximate confidence interval reported in Bird (2002).
5. For each interval, estimate coverage probability, proportion above, and proportion below as the proportion of 10,000 replications in which the interval in Equation 4 contained the population effect size, was entirely above, or was entirely below the population effect size.

Results

Table 2 contains average values of coverage probability, proportion of intervals above the population effect size, and proportion of intervals below for the correlated groups design. Results for various levels of PES are averaged over the population correlation among the levels of the repeated measures variable (RHO) and sample size. Results for various levels of RHO are averaged over PES and sample size. Again, as we speculated, the intervals become increasingly inexact as PES and RHO increase in value. The PES coverage probability ranged from .950 when PES was 0 to .845 when PES equaled 1.6. On the other hand, the empirical values ranged from .937 to .854 when RHO was increased from 0 to 0.8. In both cases, a larger proportion of intervals were found to be above the true population effect size than below. That is, again, researchers will have an overly optimistic estimate of the magnitude of the effect size, claiming value limits for the effect size that are larger than truly exist.

Table 2
Average Estimated Coverage Probability for the Correlated Groups Pairwise Comparison: Marginal Effects of Population Effect Size (PES) and the Correlation Between Two Repeated Measures (RHO)

	Coverage Probability	% Above	% Below
PES			
0.0	.950	.025	.025
0.2	.948	.028	.025
0.4	.943	.031	.026
0.6	.934	.037	.030
0.8	.922	.045	.034
1.0	.905	.055	.040
1.2	.886	.065	.048
1.4	.866	.077	.057
1.6	.845	.090	.065
RHO			
0.0	.937	.038	.025
0.2	.933	.039	.028
0.4	.924	.043	.033
0.6	.906	.052	.042
0.8	.854	.079	.067

Note. % Above/% Below = proportion above/below of 10,000 replications in which the interval contained the population effect size, was entirely above, or was entirely below the population effect size.

Table 3 presents our empirical findings for combinations of RHO and four values of PES (0.2, 0.8, 1.2, and 1.6). The reported values are averaged over sample size. Again, one can see severe depression in coverage probability (the intervals are too narrow) as RHO and PES increased in value. Indeed, for RHO = 0.8 and PES = 1.6, coverage probability was .703! That is, a .247 discrepancy between what is presumed and what occurs as RHO and PES diverge from zero. Not surprisingly, once again, these intervals frequently present an overly optimistic picture of the magnitude of effect as opposed to an underestimate of the true value.

The discrepancy between the expected coefficient value ($100\,[1 - \alpha]$) and what can actually result was much larger in our correlated groups (repeated measures) analyses of Bird's (2002) approximate intervals for effect size. That is, his interval worked reasonably well in our independent groups design but not nearly as well in the correlated groups design we investigated. Accordingly, in the following section, we define an approximate interval that we found provides better interval coverage in a two-level correlated groups design.

Table 3
Average Estimated Coverage Probability for the Correlated Groups Pairwise Comparison: Joint Effects of Population Effect Size (PES) and the Correlation Between Two Repeated Measures (RHO)

RHO	PES	Coverage Probability	% Above	% Below
0.0	0.4	.948	.028	.024
0.0	0.8	.942	.035	.023
0.0	1.2	.929	.044	.026
0.0	1.6	.914	.056	.030
0.4	0.4	.946	.030	.025
0.4	0.8	.933	.039	.028
0.4	1.2	.909	.055	.037
0.4	1.6	.878	.073	.049
0.8	0.4	.931	.037	.032
0.8	0.8	.875	.067	.058
0.8	1.2	.792	.113	.095
0.8	1.6	.703	.163	.135

Note. % Above/% Below = proportion above/below of 10,000 replications in which the interval contained the population effect size, was entirely above, or was entirely below the population effect size.

A New Approximate Interval for Correlated Means

For dependent samples from a bivariate normal distribution, the test statistic

$$t = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{\frac{S_1^2 + S_2^2 - 2S_{12}}{n}}} \tag{7}$$

is distributed as a noncentral t with degrees of freedom $n - 1$. The noncentrality parameter is

$$\lambda = \frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma_1^2 + \sigma_2^2 + 2\sigma_{12}}{n}}}. \tag{8}$$

With the population effect size δ defined by Equation 3 for equal variances or by Equation 5 for unequal variances, an approximate $100(1 - \alpha)\%$ confidence interval for δ can be constructed from the $100(1 - \alpha)\%$ confidence interval for λ. The latter can be found by using the following steps (e.g., see Steiger & Fouladi, 1997):

1. Use t in Equation 7 to estimate λ.

2. Find the noncentral *t* distribution such that *t* is at the $\alpha/2$ percentile of the distribution. This is the upper limit of the confidence interval for λ.
3. Find the noncentral *t* distribution such that *t* is at the $1 - \alpha/2$ percentile of the distribution. This is the lower limit of the confidence interval for λ.

The interval for λ can be converted into an approximate interval for δ by multiplying each limit by

$$\sqrt{\frac{2(S_1^2 + S_2^2 - 2S_{12})}{n(S_1^2 + S_2^2)}}. \tag{9}$$

Steps 2 and 3 for constructing the interval for λ can easily be implemented in any computer program that can evaluate percentiles of the noncentral *t* distribution. SAS Institute (1999) is particularly convenient to use because it contains a function TNONCT that, given *t*, degrees of freedom, and a percentile returns the noncentrality parameter such that *t* is at the provided percentile of the noncentral *t* distribution with the provided degrees of freedom.

Results for coverage probability, proportion of intervals above the true population effect size, and the proportion of intervals below the true population effect size are presented in Table 4 for the interval constructed by using Equation 9. Results for various levels of PES are averaged over RHO and sample size. Results for various levels of RHO are averaged over PES and sample size. The coverage probability results in Table 4 are much closer to their theoretical value (i.e., .95) than those reported in Table 2. That is, our approximate interval was more accurate. The range of coverage probability was .951 to .971 when PES ranged from 0 to 1.6 and the range was .958 to .957 when RHO ranged from 0 to 0.8. For the combinations of RHO and PES that we tabled, the range was .951 to .972 (see Table 5). Correspondingly, the proportions above and below values in Table 3 were typically larger than those reported in Table 5. In addition, the percentage of intervals below the population effect size was larger than the percentage above with our method as compared to Bird's (2002) approach, where the reverse was true. Thus, with our method, when a mistake is made, it is more likely to be an underestimate of the true effect size, not an overestimate, as with Bird's method.

We replicated our simulation for a design with four levels of the within-subjects factor. Bird's (2002) version of the approximate confidence interval for the effect size for comparing Groups 1 and 2 was the following:

$$\frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{\frac{S_1^2 + S_2^2 + S_3^2 + S_4^2}{4}}} \pm t_{(1-\alpha/2,\, n-1)} \sqrt{\frac{4(S_1^2 + S_2^2 - 2S_{12})}{n(S_1^2 + S_2^2 + S_3^2 + S_4^2)}}. \tag{10}$$

Table 4
Average Estimated Coverage Probability (New Method) for the Correlated Groups Pairwise Comparison: Marginal Effects of Population Effect Size (PES) and the Correlation Between Two Repeated Measures (RHO)

	Coverage Probability	% Above	% Below
PES			
0.0	.951	.025	.024
0.2	.953	.021	.027
0.4	.951	.019	.029
0.6	.955	.016	.029
0.8	.959	.013	.028
1.0	.959	.012	.029
1.2	.963	.012	.025
1.4	.966	.011	.024
1.6	.971	.009	.020
RHO			
0.0	.958	.017	.025
0.2	.960	.015	.025
0.4	.959	.015	.026
0.6	.959	.014	.027
0.8	.957	.016	.027

Note. % Above/% Below = proportion above/below of 10,000 replications in which the interval contained the population effect size, was entirely above, or was entirely below the population effect size.

Table 5
Average Estimated Coverage Probability (New Method) for the Correlated Groups Pairwise Comparison: Joint Effects of the Correlation Between Two Repeated Measures (RHO) and Population Effect Size (PES)

RHO	PES	Coverage Probability	% Above	% Below
0.0	0.4	.951	.021	.027
0.0	0.8	.959	.015	.026
0.0	1.2	.960	.014	.026
0.0	1.6	.972	.008	.020
0.4	0.4	.951	.020	.029
0.4	0.8	.958	.014	.028
0.4	1.2	.966	.010	.024
0.4	1.6	.971	.010	.019
0.8	0.4	.951	.017	.032
0.8	0.8	.955	.014	.031
0.8	1.2	.959	.013	.029
0.8	1.6	.969	.011	.020

Note. % Above/% Below = proportion above/below of 10,000 replications in which the interval contained the population effect size, was entirely above, or was entirely below the population effect size.

In the new approximate confidence interval, the confidence interval for the noncentrality parameter was obtained by using the same procedure that was used when there were just two levels. That is, the confidence limits for the new approximate interval were obtained by multiplying each limit of the confidence interval for the noncentrality parameter by

$$\sqrt{\frac{4(S_1^2 + S_2^2 - 2S_{12})}{n(S_1^2 + S_2^2 + S_3^2 + S_4^2)}}. \tag{11}$$

The results indicated that the interval used by Bird (2002) continued to perform poorly in the same conditions in which it performed poorly for the design with two within-subjects levels. For example, for RHO = 0.8 and PES = 1.6, coverage probability was .717 in the new study and .703 in the first study. The new interval seemed to yield somewhat inflated coverage (.975) probability when Equation 11 was used; consequently, we recommend using Equation 9.

Examples

Although Bird's (2002) approximate interval works fairly well for an independent samples design, the interval is too narrow when the population effect size is large and can provide an optimistic assessment of the effect size, a tendency that increases as the sample size decreases and as the effect size increases. (It should also be noted that unreported results comparing the approximate and exact interval for the independent samples case, with two samples, showed that when both intervals had the correct coverage probability, the exact interval tended to be shorter, on average, than was the approximate interval.) In Appendix A, we present a SAS/IML (SAS Institute, 1999) program that calculates exact confidence intervals when the samples are independent. It is set up to provide results for Bird's first example, in which there were three independent groups with means $\bar{Y}_1 = 22.467$, $\bar{Y}_2 = 24.933$, and $\bar{Y}_3 = 32.000$. It should be noted that, following Bird's example, we apply the Bonferroni principle and construct confidence intervals intended to control the familywise confidence level. However, the program contains a switch that allows the user to construct confidence intervals intended to control the per comparison confidence level. Also following Bird, we assume the variances are equal for the three levels and use the square root of the variance pooled over the three groups to standardize the contrasts and compute the exact confidence intervals. However, the program contains a switch that allows the user to pool the variances over only those levels involved in the contrast for which an effect size is to be estimated.

Bird reported confidence intervals for effect sizes for two contrasts. The first compared the average of the first two means to the third mean. The sec-

Table 6
Comparison of Independent Groups' Confidence Intervals

Contrast	Method	Effect Size	Lower Limit	Upper Limit
$\hat{\psi}_1 = (\bar{Y}_1 + \bar{Y}_2)/2 - \bar{Y}_3$	Bird	–1.12	–1.63	–0.61
	Exact	–1.12	–1.90	–0.67
$\hat{\psi}_1 = \bar{Y}_1 - \bar{Y}_2$	Bird	0.33	–0.92	0.26
	Exact	0.33	–0.91	0.25

Table 7
Comparison of Correlated Groups' Confidence Intervals

Contrast	Method	Effect Size	Lower Limit	Upper Limit
$\hat{\psi}_1 = \bar{Y}_1 - \bar{Y}_2$	Bird	1.39	1.00	1.78
	New	1.39	0.82	1.95
$\hat{\psi}_2 = \bar{Y}_1 - \bar{Y}_3$	Bird	1.30	0.94	1.65
	New	1.30	0.76	1.83
$\hat{\psi}_3 = \bar{Y}_2 - \bar{Y}_3$	Bird	–0.09	–0.36	0.18
	New	–0.09	–0.35	0.16

ond contrast compared the first and second means. Results comparing Bird's intervals to exact intervals produced using the noncentral *t* distribution are presented in Table 6. As will occur when the estimated effect size is small, the intervals for the second contrast are in close agreement. And as will occur when the estimated effect size is large, the intervals for the first contrast exhibit less agreement. The exact interval is wider than the approximate interval.

The SAS/IML (SAS Institute, 1999) program in Appendix B calculates the new approximate confidence interval when the samples are dependent. It is set up to provide results for Bird's third example, in which there were three dependent groups with means $\bar{Y}_1 = 32.500$, $\bar{Y}_2 = 22.467$, and $\bar{Y}_3 = 23.133$. Again, following Bird's example, in the program we use the Bonferroni principle to construct the confidence intervals. We also use the square root of the average of the variances for the repeated measures to standardize the contrasts and compute the approximate confidence interval. Switches are provided in the program to allow one to use an unadjusted confidence level and the variances for only those levels involved in the contrast for which an effect size is to be estimated.

Bird reported approximate intervals for pairwise comparisons. Results comparing Bird's approximate interval to our approximate interval are presented in Table 7. We see that the intervals are in closer agreement when the estimated effect size is small, but not when the estimated effect sizes are large.

Discussion

Because theoretical considerations lead us to believe that the approximate intervals of effect size proposed by Bird (2002) for independent and correlated groups designs could become inaccurate (very approximate) as the size of the population effect (and the correlation between treatment levels in correlated groups designs) increases in value, we conducted an investigation into the operating characteristics of his intervals. Our empirical results confirmed our hypothesis, particularly for the correlated groups design. That is, coverage probability diverged from the theoretical value (i.e., .95), and the divergence could be described as substantial in correlated groups designs. For the correlated samples design, the approximation of the theoretical value can be very off, for example, .703 rather than .95. The consequence of this distortion is that applied researchers can form very inaccurate assessments of the confidence limits. Specifically, they can think that intervals are much shorter than they should be. Furthermore, the intervals are more likely to be above the true effect size than below it, thus giving an optimistic picture of magnitude of effect.

Based on our hypothesis and subsequent empirical findings, we suggested another approximate interval for the dependent samples case that we believed would provide better approximate coverage. The findings we reported for this new approximate interval for effect sizes in correlated (repeated measures) designs support our belief that the new interval would provide better coverage probability. Indeed, over all the cases we investigated, the average coverage probability for our method was .959! Moreover, when a mistake is made, our approach tends to underestimate the magnitude of effect rather than overestimate the possible range of values for effect size. Thus, our method will provide researchers with a fairly accurate assessment of the limits of the interval and will not present a too optimistic picture of the actual effect size. When a mistake might be made, we believe it is preferable to underestimate the magnitude of effect than to overestimate it. That is, by analogy to significance testing, the error of falsely claiming larger effects than truly exist could be more damaging to scientific inquiry than the error of underestimating the magnitude of effects. Naturally, the nature of the research could reverse this view.

Appendix A
SAS/IML (SAS Institute, 1999) Program for Confidence Intervals on the Effect Size for an Independent Samples Design

This program is used with between-subjects designs. It computes confidence intervals for effect size estimates. To use the program one inputs at the top of the program:

m—a vector of means
sd—a vector of standard deviations
n—a vector of sample size
prob—the confidence level prior to the Bonferroni adjustment

adjust—the number of contrasts if a Bonferroni adjustment to the confidence level is requested. Otherwise adjust=1

In addition one inputs at the bottom of the program:

c—a vector of contrast weights

Multiple contrasts can be entered. After each, type the code run ci;

```
proc iml;
m={22.467 24.933 32.000};
sd={7.001 8.288 6.938};
v=sd##2;
n={30 30 30};
cl=.95;
adjust=2;
prob=cl;
df=(n-j(1,ncol(n),1));
pdf=df[,+];
temp= df#v;
pvar=temp[,+]/pdf;
nn=diag(n);
ni=inv(nn);
print ‘Vector of means:’;
print m;
print ‘Vector of standard deviations:’;
print sd;
print ‘Vector of sample sizes:’;
print n;
print ‘Confidence level before Bonferroni adjustment:’;
print cl;
cl=1-(1-prob)/adjust;
print ‘Confidence level with Bonferroni adjustment:’;
print cl;
print ‘Pooled df:’;
print pdf;
print ‘Pooled variance:’;
print pvar;
start CI;
es=m*c’/sqrt(pvar);
nchat=es/(sqrt(c*ni*c’));
```

```
ncu=TNONCT(nchat,pdf,(1-prob)/(2*adjust));
ncl=TNONCT(nchat,pdf,1-(1-prob)/(2*adjust));
ll=(sqrt(c*ni*c'))*ncl;
ul=(sqrt(c*ni*c'))*ncu;
print 'Effect size:';
print es;
print 'Estimated noncentrality parameter';
print nchat;
print 'll is the lower limit of the CI and ul is the upper limit';
print ll ul;
finish;
c={.5 .5 -1};
run ci;
c={1 -1 0};
run ci;
quit;
```

Note: Programs listed in this appendix can be downloaded at plaza.ufl.edu/algina/index.programs.html

Appendix B
SAS/IML (SAS Institute, 1999) Program for Confidence Intervals on the Effect Size for a Dependent Samples Design

This program is used with within-subjects designs. It computes confidence intervals for effect size estimates. To use the program one inputs at the top of the program:

m—a vector of means

v—a covariance matrix in lower diagonal form, with periods for the upper elements

n—the sample size

prob—the confidence level prior to the Bonferroni adjustment

adjust—the number of contrasts if a Bonferroni adjustment to the confidence level is requested. Otherwise adjust=1

Bird—a switch that uses the variances of all variables to calculate the denominator of the effect size as suggested by K. Bird

(Bird=1). Our suggestion is to use the variance of those variables involved in the contrast to calculate the denominator of the effect size (Bird=0)

In addition one inputs at the bottom of the program:

c—a vector of contrast weights

Multiple contrasts can be entered. After each, type the code run ci;

```
proc iml;
m={32.500 22.467 23.133};
v={58.121 . .,
35.517 49.016 .,
36.690 40.384 49.430};
do ii = 1 to nrow(v)-1;
do jj = ii+1 to nrow(v);
```

```
v[ii,jj]=v[jj,ii];
end;
end;
n=30 ;
Bird=1;
df=n-1;
cl=.95;
adjust=3;
prob=cl;
print ‘Vector of means:’;
print m;
print ‘Covariance matrix:’;
print v;
print ‘Sample size:’;

print n;
print ‘Confidence level before Bonferroni adjustment:’;
print cl;
cl=1-(1-prob)/adjust;
print ‘Confidence level with Bonferroni adjustment:’;
print cl;
start CI;
pvar=0;
count=0;
if bird=0 then do;
do mm=1 to nrow(v);
if c[1,mm]^=0 then do;
pvar=pvar+v[mm,mm];
count=count+1;
end;
end;
end;
if bird=1 then do;
do mm=1 to nrow(v);
pvar=pvar+v[mm,mm];
count=count+1;
end;
end;
pvar=pvar/count;
es=m*c’/(sqrt(pvar));
se=(sqrt(c*v*c’/n);
nchat=m*c’/se;
ncu=TNONCT(nchat,df,(1-prob)/(2*adjust));
ncl=TNONCT(nchat,df,1-(1-prob)/(2*adjust));
ll=se*ncl/(sqrt(pvar));
ul=se*ncu/(sqrt(pvar));
print ‘Contrast vector’;
```

```
print c;
print 'Effect size:';
print es;
print 'Estimated noncentrality parameter';
print nchat;
print 'll is the lower limit of the CI and ul is the upper limit';
print ll ul;
finish;
c={1 -1 0};
run ci;
c={1 0 -1};
run ci;
c={0 1 -1};
run ci;
quit;
```

Note: Programs listed in this appendix can be downloaded at plaza.ufl.edu/algina/index.programs.html

References

Bird, K. D. (2002). Confidence intervals for effect sizes in analysis of variance. *Educational and Psychological Measurement, 62,* 197-226.

Browne, M. W. (1969). *Precision of prediction* (Research Bulletin No. 69-69). Princeton, NJ: Educational Testing Service.

Cumming, G., & Finch, S. (2001). A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement, 61,* 532-574.

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

SAS Institute. (1999). *SAS/IML user's guide, version 8.* Cary, NC: Author.

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

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