Psychological Methods
2008, Vol. 13, No. 2, 99–109

1082-989X/08/$12.00 DOI: 10.1037/1082-989X.13.2.99

Confidence Intervals for Standardized Linear Contrasts of Means

Douglas G. Bonett
Iowa State University

Most psychology journals now require authors to report a sample value of effect size along with hypothesis testing results. The sample effect size value can be misleading because it contains sampling error. Authors often incorrectly interpret the sample effect size as if it were the population effect size. A simple solution to this problem is to report a confidence interval for the population value of the effect size. Standardized linear contrasts of means are useful measures of effect size in a wide variety of research applications. New confidence intervals for standardized linear contrasts of means are developed and may be applied to between-subjects designs, within-subjects designs, or mixed designs. The proposed confidence interval methods are easy to compute, do not require equal population variances, and perform better than the currently available methods when the population variances are not equal.

Keywords: Cohen's *d*, effect size, heteroscedasticity, noncentral *t* distribution, standardized mean difference

In response to the recommendations of the Task Force on Statistical Inference convened by the Board of Scientific Affairs and the American Psychological Association (Wilkinson & the Task Force on Statistical Inference, 1999), most psychology journals now require authors to supplement their hypothesis testing results with a point estimate (i.e., sample value) of effect size to provide information regarding the magnitude of the effect that cannot be ascertained from the *p* value alone. A standardized difference in means, such as Cohen's *d* or Hedges's *g*, is a commonly used measure of effect size. Values of *d* equal to 0.2, 0.5, and 0.8 are suggested by Cohen (1988, p. 25) to represent small, medium, and large effects, respectively. In small samples, a point estimate of *d* must be large to obtain $p < .05$, a criterion for publication that is still used in many psychology journals. Psychology journals are now replete with studies that report a large effect size when, in fact, the true effect size could be close to zero. For instance, a researcher might report a statistical result from a two-group experiment as $t(18) = 2.15$, $p < .05$, $d = 0.96$, and claim that the treatment effect is "statistically significant" and "large." This conclusion is deceiving because the value $d = 0.96$ describes the sample and not the population of interest. If the researcher also had reported a 95% confidence interval for the population effect size in this study, it would have been revealed that the true effect size could be close to zero. The current practice of reporting only a point estimate of an effect size is actually a step backward from the goal of improving the quality of psychological research. This problem is easily rectified by simply reporting an effect size confidence interval in addition to an effect size point estimate. The reporting of an effect size confidence interval was a key recommendation of the Task Force for Statistical Inference, but there has been minimal compliance with this recommendation in psychology journals (Cumming et al., 2007).

Many interesting research questions may be expressed in terms of a linear contrast of k population means, $\psi = \sum_{j=1}^{k} c_j \mu_j$, where μ_j is a population mean and $\sum_{j=1}^{k} c_j = 0$. Excellent discussions of the use of linear contrasts in single-factor designs can be found in Keppel and Wickens (2004) and Maxwell and Delaney (2004). Linear contrasts are used in qualitative single-factor designs to assess pairwise comparisons, comparisons with a control, and differences between averages of means. In designs with quantitative factors, a slope parameter may be defined in terms of linear contrasts. In factorial designs, linear contrasts may be used to assess main effects, interaction effects, and simple effects (Bird, 2004).

A linear contrast of population means is a measure of effect size that is useful in applications in which the researcher knows enough about the metric of the response variable to assess the meaning and importance of the ψ value. In applications where information regarding the metric of the response variable is limited, a standardized linear contrast of means or a correlation measure such as ρ_{contrast} (Rosenthal, Rosnow, & Rubin, 2000, p. 62) may be easier to interpret than an unstandardized linear contrast of means. Standardized linear contrasts of means are considered here.

Correspondence concerning this article should be addressed to Douglas G. Bonett, Department of Statistics, Iowa State University, Ames, IA 50011. E–mail: dgbonett@iastate.edu

Unlike the confidence intervals for unstandardized linear contrast of means with Satterthwaite degrees of freedom (Snedecor & Cochran, 1980, p. 228), the currently available confidence intervals for standardized linear contrasts of means are not robust to minor violations of their equal variance assumption. Furthermore, some of the currently available methods are computationally intensive and require specialized computer software.

Currently Used Confidence Intervals

Kline (2004, pp. 171–179) provided a review of point and interval estimates for standardized linear contrasts of means. In the case of a between-subjects design with k independent samples, Kline (2004, p. 170) recommended the following estimate of a standardized linear contrast of means:

$$\hat{\delta}' = \sum_{j=1}^{k} c_j \hat{\mu}_j / \hat{\sigma}_p \tag{1}$$

where $\hat{\sigma}_p = \{[\sum_{j=1}^{k}(n_j - 1)\hat{\sigma}_j^2]/(\sum_{j=1}^{k} n_j - k)\}^{1/2}$, $\hat{\mu}_j$ is a sample mean, $\hat{\sigma}_j^2$ is an unbiased sample variance, and n_j is the sample size in group j. Note that $\hat{\sigma}_p^2$ is a pooled variance and equal to the analysis of variance mean square error for k independent samples. Equation 1 with $c_1 = 1$ and $c_2 = -1$ reduces to Hedges's g in the case of two independent samples. Equation 1 also may be used in a within-subjects design with k levels where k measurements are obtained from every member of a single group of size n. For within-subject designs, n_j is replaced with n in Equation 1. The unknown population parameter that is estimated by Equation 1 is denoted as δ'.

Textbook discussions of linear contrasts often focus on hypothesis testing applications where contrast coefficients such as $c_1 = 2$, $c_2 = -1$, and $c_3 = -1$ yield exactly the same test statistic and p value as $c_1 = 1$, $c_2 = -1/2$, and $c_3 = -1/2$, but ψ would be twice as large using the first set of coefficients and would not be an interesting parameter to estimate. As a general rule for qualitative factors, the absolute values of contrast coefficients should sum to 2 for pairwise comparisons, comparisons with a control, and differences between averages of means. This rule also applies to contrast coefficients for main effects and simple main effects in factorial designs.

The following approximate $100(1 - \alpha)\%$ confidence interval for δ' was suggested by Bird (2002) and may be applied in between-subjects or within-subjects designs:

$$\left\{ \sum_{j=1}^{k} c_j \hat{\mu}_j \pm t_{\alpha/2;df} \left[\text{var}\left(\sum_{j=1}^{k} c_j \hat{\mu}_j \right) \right]^{1/2} \right\} / \hat{\sigma}_p \tag{2}$$

where $t_{\alpha/2;df}$ is a two-tailed critical t value and $\hat{\sigma}_p$ is given in Equation 1. For between-subjects designs, $df = \sum_{j=1}^{k} n_j - k$ and $\text{var}(\sum_{j=1}^{k} c_j \hat{\mu}_j) = \hat{\sigma}_p^2 \sum_{j=1}^{k} c_j^2 / n_j$. For within-subjects designs, $df = n - 1$ and $\text{var}(\sum_{j=1}^{k} c_j \hat{\mu}_j) = \hat{\sigma}_y^2 / n$ where $y_i = \sum_{j=1}^{k} c_j y_{ij}$ and y_{ij} is the jth score for participant i. Although Equation 2 may be used to assess a standardized difference between two means, better methods have been recommended for this important special case.

Viechtbauer (2007) examined the small-sample performance of several confidence intervals for a standardized difference between two means. For the case of two independent samples, Viechtbauer (2007) found that the following approximate $100(1 - \alpha)\%$ confidence interval for δ' was one of the best:

$$g \pm z_{\alpha/2}[\text{var}(g)]^{1/2}, \tag{3}$$

where $g = (\hat{\mu}_1 - \hat{\mu}_2)/\hat{\sigma}_p$, $\text{var}(g) = g^2/(2n_1 + 2n_2 - 4) + 1/n_1 + 1/n_2$, and $z_{\alpha/2}$ is a two-tailed critical z value. Viechtbauer (2007) found that Equation 3 captures δ' with probability very close to $1 - \alpha$ in small samples assuming the two population variances are equal. For a within-subject design with $k = 2$ levels, Viechtbauer (2007) found that the following approximate $100(1 - \alpha)\%$ confidence interval for $(\mu_1 - \mu_2)/\sigma_1$ was one of the best:

$$d \pm z_{\alpha/2}[\text{var}(d)]^{1/2}, \tag{4}$$

where $d = (\hat{\mu}_1 - \hat{\mu}_2)/\hat{\sigma}_1$, $\text{var}(d) = d^2/(2n - 2) + 2(1 - \hat{\rho}_{12})/n$, $\hat{\rho}_{12}$ is the sample product–moment correlation between the n pairs of observations, and $\hat{\sigma}_1$ is the sample standard deviation in the pretest or control condition. Algina and Keselman (2003) also examined within-subjects designs with $k = 2$ levels and proposed an approximate confidence interval for $\delta = (\mu_1 - \mu_2)/[(\sigma_1^2 + \sigma_2^2)/2]^{1/2}$ that performs well under the conditions they considered.

For the case of two independent samples, Hedges and Olkin (1985, p. 91) explained how an exact confidence interval for δ' may be obtained from an exact confidence interval for the noncentrality parameter of a noncentral t distribution. This approach generalizes to a standardized linear contrast of means by noting that

$$\delta' = \Delta \left(\sum_{j=1}^{k} c_j^2 / n_j \right)^{1/2} \tag{5}$$

(Steiger, 2004), where Δ is the noncentrality parameter. Multiplying the endpoints of an exact $100(1 - \alpha)\%$ confidence interval for Δ by $(\sum_{j=1}^{k} c_j^2/n_j)^{1/2}$ gives an exact $100(1 - \alpha)\%$ confidence interval for δ'. Obtaining an exact confidence interval for Δ is computationally intensive and requires specialized programs such as those described in Kline (2004) and Bird (2004). An exact confidence interval for δ' is currently unavailable for within-subject designs. The exact confidence intervals are exact only under an assumption of normality and equal population variances.

It is common practice to perform a statistical test for homogeneity of variances and conclude that the equal variance assumption has been satisfied if the null hypothesis is not rejected. Failure to reject the null hypothesis of equal population variances should not be taken as evidence that the population variances are equal. The real question concerns the degree to which the population variances differ, and this question may be answered by computing simultaneous confidence intervals for all pairwise ratios of population variances. However, large sample sizes are needed to obtain acceptably narrow simultaneous confidence intervals. Given the difficulty of accurately assessing the degree to which population variances differ, a confidence interval for δ that does not assume equal population variances would be desirable. Such a confidence interval is described in the following section.

Proposed Confidence Interval

The following alternative estimate of a standardized linear contrast of means is recommended here for both between-subjects and within-subjects designs:

$$\hat{\delta} = \sum_{j=1}^{k} c_j \hat{\mu}_j / \hat{\sigma} \tag{6}$$

where $\hat{\sigma} = (k^{-1}\sum_{j=1}^{k}\hat{\sigma}_j^2)^{1/2}$. Equations 1 and 6 are identical in within-subjects designs and in between-subjects designs with equal sample sizes. The unknown population parameter that is estimated by Equation 6 is denoted as δ.

A confidence interval for δ is developed here by first deriving the variance of $\hat{\delta}$. The derivation is given in the Appendix. For between-subjects designs, the variance of $\hat{\delta}$ may be estimated as

$$\text{var}(\hat{\delta}) = (\hat{\delta}^2/k^2\hat{\sigma}^4)\sum_{j=1}^{k}\hat{\sigma}_j^4/2df_j + \left(\sum_{j=1}^{k} c_j^2\hat{\sigma}_j^2/df_j\right)/\hat{\sigma}^2, \tag{7}$$

where $df_j = n_j - 1$, and Equation 7 specializes to

$$\text{var}(\hat{\delta}) = \hat{\delta}^2(\hat{\sigma}_1^4/df_1 + \hat{\sigma}_2^4/df_2)/8\hat{\sigma}^4 + \hat{\sigma}_1^2/\hat{\sigma}^2 df_1 + \hat{\sigma}_2^2/\hat{\sigma}^2 df_2 \tag{8}$$

for a standardized difference between two means. For within-subjects designs,

$$\text{var}(\hat{\delta}) = (\hat{\delta}^2/2k^2\hat{\sigma}^4 df)\left(\sum_{j=1}^{k}\hat{\sigma}_j^4 + 2\sum_{r=1}^{k}\sum_{t=r+1}^{k}\hat{\rho}_{rt}^2\hat{\sigma}_r^2\hat{\sigma}_t^2\right) + \left(\sum_{j=1}^{k} c_j^2\hat{\sigma}_j^2 + 2\sum_{r=1}^{k}\sum_{t=r+1}^{k} c_r c_t\hat{\rho}_{rt}\hat{\sigma}_r\hat{\sigma}_t\right)/\hat{\sigma}^2 df, \tag{9}$$

where $df = n - 1$, and Equation 9 specializes to

$$\text{var}(\hat{\delta}) = \hat{\delta}^2(\hat{\sigma}_1^4 + \hat{\sigma}_2^4 + 2\hat{\rho}_{12}^2\hat{\sigma}_1^2\hat{\sigma}_2^2)/8\hat{\sigma}^4 df + (\hat{\sigma}_1^2 + \hat{\sigma}_2^2 - 2\hat{\rho}_{12}\hat{\sigma}_1\hat{\sigma}_2)/\hat{\sigma}^2 df \tag{10}$$

for a standardized difference between two means.

The following approximate 100(1 − α)% confidence interval for δ is proposed:

$$\hat{\delta} \pm z_{\alpha/2}[\text{var}(\hat{\delta})]^{1/2}, \tag{11}$$

where var($\hat{\delta}$) is given by Equation 7 or 8 for between-subjects designs and Equation 9 or 10 for within-subjects designs. To obtain simultaneous Bonferroni confidence intervals for v linear contrasts, replace α with $\alpha^* = \alpha/v$ in $z_{\alpha/2}$.

Equation 11 is an approximate large-sample confidence interval. Computer simulation results to assess its small-sample performance in comparison with Equations 2–4 are reported in the following section.

Simulation Studies

The Monte Carlo method was used to examine the small-sample performance of Equations 2–4 and 11. The coverage probabilities at α = .05 were estimated from 100,000 Monte Carlo replications for each condition. The simulation programs were written in GAUSS and executed on a Pentium 4 computer.

Between-Subjects Designs

The two-group case was examined first. The sample data were randomly generated from independent normal distributions with $\boldsymbol{\delta}$ = [0.25 0.50 1.0 2.0], $\boldsymbol{\sigma}$ = [1 1] for the homoscedastic case and $\boldsymbol{\sigma}$ = [1 1.5] for the heteroscedastic case. The probabilities of Equation 11 capturing δ and Equation 3 capturing δ′ were estimated for each condition. The results are summarized in Table 1. The results indicate that the performance of Equation 11 is similar to the performance of Equation 3 under homoscedasticity or equal sample sizes, but Equation 11 has a coverage probability closer to .95 than Equation 3 under mild heteroscedasticity combined with unequal sample sizes. The performance of Equation 3 is unacceptable with unequal sample sizes and a degree of heteroscedasticity that would be very difficult to detect in small or moderate sample sizes using standard diagnostic tools.

Equation 3 has nearly exact coverage probability when the sample sizes are equal and the population variances are approximately equal. In these cases, the average width of Equation 3 may meaningfully be compared with the average width of Equation 11. Table 2 gives the average 95% confidence interval widths for Equations 3 and 11 under conditions of equal sample sizes and equal population variances. As can be seen in Table 2, the average width of Equation 3 is slightly less than the average width of Equa-

Table 1
95% Coverage Probabilities for Two Approximate Confidence Intervals: Two Independent Samples

			$\boldsymbol{\sigma}$ = [1 1]		$\boldsymbol{\sigma}$ = [1 1.5]	
n_1	n_2	δ	Eq. 3	Eq. 11	Eq. 3	Eq. 11
10	10	0.25	.948	.958	.946	.959
10	10	0.50	.946	.958	.946	.959
10	10	1.00	.948	.959	.946	.958
10	10	2.00	.952	.960	.945	.957
20	10	0.25	.949	.954	.913	.952
20	10	0.50	.948	.954	.913	.952
20	10	1.00	.950	.954	.916	.953
20	10	2.00	.950	.956	.919	.950
10	20	0.25	.948	.954	.972	.956
10	20	0.50	.948	.954	.970	.956
10	20	1.00	.949	.954	.968	.957
10	20	2.00	.950	.956	.963	.958
30	10	0.25	.949	.950	.893	.950
30	10	0.50	.950	.951	.892	.950
30	10	1.00	.949	.951	.894	.948
30	10	2.00	.950	.951	.905	.946
10	30	0.25	.948	.949	.981	.954
10	30	0.50	.949	.950	.980	.954
10	30	0.50	.949	.950	.980	.954
10	30	1.00	.950	.951	.979	.954
10	30	2.00	.949	.950	.973	.955

tion 11. It also should be noted that the average width of the exact confidence interval based on the noncentral *t* distribution is nearly identical to the average width of Equation 3 when the sample sizes are equal. In practical terms, Equation 11 requires one additional member per group compared with Equation 3 to achieve about the same interval width as Equation 3. The additional member per group may be viewed as the cost of relaxing the homoscedasticity assumption. Recognize also that Equation 3 will have a coverage probability of less than $1 - \alpha$, even with equal sample sizes when the population variances are not approximately equal, and Equation 3 will then have an interval width that is misleadingly more narrow than that of Equation 11.

Table 2
Comparison of 95% Confidence Interval Widths Under Homoscedasticity

n_1	n_2	δ	Eq. 3	Eq. 11
10	10	0.25	1.79	1.88
10	10	0.50	1.81	1.91
10	10	1.00	1.91	2.01
10	10	2.00	2.24	2.36
20	20	0.25	1.25	1.29
20	20	0.50	1.27	1.30
20	20	1.00	1.33	1.36
20	20	2.00	1.55	1.59
30	30	0.25	1.02	1.04
30	30	0.50	1.03	1.05
30	30	1.00	1.08	1.10
30	30	2.00	1.26	1.28

Sample data were randomly generated from homoscedastic and heteroscedastic normal distributions with $\boldsymbol{\delta}$ = [0.25 0.50 1.0 2.0] to assess the performance of Equations 2 and 11 in three-group and four-group designs. In the heteroscedastic conditions, $\boldsymbol{\sigma}$ = [1 1 1.5] for $k = 3$ and $\boldsymbol{\sigma}$ = [1 1 1 1.5] for $k = 4$. The results of the computer simulation are summarized in Table 3 for $\boldsymbol{c}$ = [1 −0.5 −0.5] and in Table 4 for $\boldsymbol{c}$ = [0.5 0.5 −0.5 −0.5]. The results indicate that the performance of Equation 2 is unacceptable unless the population variances are equal and $\delta \leq 0.5$. Equation 11 performed well under all of the conditions examined. In between-subjects designs, it is easy to show that Equation 2 assumes a zero effect size (Viechtbauer, 2007), which explains its poor performance when $\delta > 0.5$.

Within-Subjects Designs

The paired-sample ($k = 2$) case was examined first. Sample data were randomly generated from homoscedastic and heteroscedastic bivariate normal distributions with $\boldsymbol{\delta}$ = [0.25 0.50 1.0 2.0], $\boldsymbol{\rho}$ = [0.3 0.6 0.9], $\boldsymbol{\sigma}$ = [1 1], and $\boldsymbol{\sigma}$ = [1 1.5]. The probabilities of Equation 11 capturing δ and Equation 4 capturing $(\mu_1 - \mu_2)/\sigma_1$ were estimated for each condition. The results are summarized in Table 5 and indicate that the performance of Equation 11 is superior to the performance of Equation 4 in almost every condition. The performance of Equation 4 is unacceptable with a degree of heteroscedasticity that would be very difficult to assess in small or moderate sample sizes using standard diagnostic tools.

Table 3
95% Coverage Probabilities for Two Approximate Confidence Intervals: Three Independent Samples With $\mathbf{c}$ *= [1 −0.5 −0.5]*

				$\boldsymbol{\sigma}$ = [1 1 1]		$\boldsymbol{\sigma}$ = [1 1 1.5]	
n_1	n_2	n_3	δ	Eq. 2	Eq. 11	Eq. 2	Eq. 11
10	10	10	0.25	.949	.956	.963	.959
10	10	10	0.50	.946	.956	.960	.959
10	10	10	1.00	.936	.956	.948	.959
10	10	10	2.00	.895	.960	.902	.959
10	10	30	0.25	.949	.949	.986	.951
10	10	30	0.50	.948	.950	.984	.951
10	10	30	1.00	.939	.951	.979	.953
10	10	30	2.00	.914	.953	.951	.954
30	10	10	0.25	.948	.958	.936	.957
30	10	10	0.50	.946	.958	.933	.957
30	10	10	1.00	.934	.956	.921	.956
30	10	10	2.00	.892	.957	.870	.954

Table 4
95% Coverage Probabilities for Two Approximate Confidence Intervals: Four Independent Samples With $\mathbf{c} = [0.5\ 0.5\ -0.5\ -0.5]$

					$\boldsymbol{\sigma} = [1\ 1\ 1\ 1]$		$\boldsymbol{\sigma} = [1\ 1\ 1\ 1.5]$	
n_1	n_2	n_3	n_4	δ	Eq. 2	Eq. 11	Eq. 2	Eq. 11
10	10	10	10	0.25	.949	.958	.947	.959
10	10	10	10	0.50	.945	.959	.944	.958
10	10	10	10	1.00	.932	.959	.929	.958
10	10	10	10	2.00	.885	.961	.878	.960
10	10	10	30	0.25	.949	.955	.980	.955
10	10	10	30	0.50	.948	.956	.977	.954
10	10	10	30	1.00	.938	.956	.970	.958
10	10	10	30	2.00	.900	.957	.929	.958
30	10	10	10	0.25	.949	.955	.932	.955
30	10	10	10	0.50	.948	.956	.931	.956
30	10	10	10	1.00	.938	.956	.919	.955
30	10	10	10	2.00	.900	.957	.876	.955

Sample data were randomly generated from multivariate homoscedastic and heteroscedastic normal distributions with $\boldsymbol{\delta} = [0.25\ 0.50\ 1.0\ 2.0]$ so that the performance of Equations 2 and 11 for $k = 3$ and $k = 4$ within-subject levels could be assessed. In the heteroscedastic conditions, $\boldsymbol{\sigma} = [1\ 1\ 1.5]$ for $k = 3$ and $\boldsymbol{\sigma} = [1\ 1\ 1\ 1.5]$ for $k = 4$. In a preliminary investigation, it was found that Equations 2 and 11 were affected primarily by the average magnitude of the correlations rather than any specific pattern of correlations. For this reason, the simulations could be simplified by using multivariate normal distributions with a common correlation of 0.3, 0.6, or 0.9. The results are summarized in Table 6 for $\boldsymbol{c} = [1\ -0.5\ -0.5]$ and in Table 7 for $\boldsymbol{c} = [0.5\ 0.5\ -0.5\ -0.5]$. The results indicate that the performance of Equation 2 is unacceptable unless the population variances are equal and $\delta \leq 0.25$. The negative effect of a nonzero effect size on Equation 2 is more severe with larger correlations. Equation 11 performed well under all conditions examined.

The pattern of results in Tables 1–7 for $\alpha = .05$ also were found to hold for $\alpha = .10$ and $\alpha = .01$. The proposed confidence interval (Equation 11) also performs very well under more extreme levels of heteroscedasticity such as $\sigma_{max}/\sigma_{min} = 2$, 4, and 8. However, with extreme heteroscedasticity, the meaningfulness of δ as a standardized measure of effect size might be called into question.

Table 5
95% Coverage Probabilities for Two Approximate Confidence Intervals: Two Within-Subjects Levels

			$\boldsymbol{\sigma} = [1\ 1]$		$\boldsymbol{\sigma} = [1\ 1.5]$	
n	ρ	δ	Eq. 4	Eq. 11	Eq. 4	Eq. 11
10	0.3	0.25	.930	.948	.859	.948
10	0.3	0.50	.932	.949	.864	.951
10	0.3	1.00	.938	.952	.884	.952
10	0.3	2.00	.948	.958	.915	.957
10	0.6	0.25	.929	.950	.844	.950
10	0.6	0.50	.932	.951	.858	.951
10	0.6	1.00	.939	.952	.883	.953
10	0.6	2.00	.948	.957	.921	.956
10	0.9	0.25	.930	.950	.783	.953
10	0.9	0.50	.934	.949	.829	.954
10	0.9	1.00	.946	.951	.891	.954
10	0.9	2.00	.952	.951	.935	.953
30	0.3	0.25	.944	.949	.876	.950
30	0.3	0.50	.944	.950	.872	.949
30	0.3	1.00	.946	.951	.889	.950
30	0.3	2.00	.950	.953	.917	.952
30	0.6	0.25	.943	.950	.853	.950
30	0.6	0.50	.944	.952	.864	.951
30	0.6	1.00	.946	.951	.889	.950
30	0.6	2.00	.949	.952	.922	.951
30	0.9	0.25	.943	.950	.784	.950
30	0.9	0.50	.945	.950	.827	.952
30	0.9	1.00	.949	.951	.890	.951
30	0.9	2.00	.950	.950	.933	.951

Table 6
95% Coverage Probabilities for Two Approximate Confidence Intervals: Three Within-Subjects Levels With $\mathbf{c} = [1\ -0.5\ -0.5]$

			$\boldsymbol{\sigma} = [1\ 1\ 1]$		$\boldsymbol{\sigma} = [1\ 1\ 1.5]$	
n	ρ	δ	Eq. 2	Eq. 11	Eq. 2	Eq. 11
10	0.3	0.25	.948	.944	.949	.945
10	0.3	0.50	.945	.947	.944	.947
10	0.3	1.00	.930	.951	.923	.951
10	0.3	2.00	.869	.961	.848	.960
10	0.6	0.25	.947	.947	.947	.948
10	0.6	0.50	.937	.947	.935	.947
10	0.6	1.00	.899	.951	.891	.952
10	0.6	2.00	.768	.957	.743	.956
10	0.9	0.25	.930	.948	.934	.951
10	0.9	0.50	.873	.949	.882	.952
10	0.9	1.00	.701	.950	.718	.952
10	0.9	2.00	.436	.950	.455	.950
30	0.3	0.25	.950	.948	.949	.949
30	0.3	0.50	.946	.949	.942	.947
30	0.3	1.00	.929	.950	.922	.951
30	0.3	2.00	.869	.953	.841	.953
30	0.6	0.25	.948	.949	.947	.949
30	0.6	0.50	.937	.949	.933	.949
30	0.6	1.00	.895	.951	.886	.950
30	0.6	2.00	.755	.952	.729	.950
30	0.9	0.25	.930	.949	.932	.951
30	0.9	0.50	.868	.949	.875	.950
30	0.9	1.00	.685	.950	.701	.950
30	0.9	2.00	.424	.951	.443	.950

Table 7
95% Coverage Probabilities for Two Approximate Confidence Intervals: Four Within-Subjects Levels With $\mathbf{c} = [0.5\ 0.5\ -0.5\ -0.5]$

			$\boldsymbol{\sigma} = [1\ 1\ 1\ 1]$		$\boldsymbol{\sigma} = [1\ 1\ 1\ 1.5]$	
n	ρ	δ	Eq. 2	Eq. 11	Eq. 2	Eq. 11
10	0.3	0.25	.948	.942	.948	.943
10	0.3	0.50	.944	.945	.943	.945
10	0.3	1.00	.924	.952	.923	.952
10	0.3	2.00	.853	.965	.847	.963
10	0.6	0.25	.946	.946	.945	.945
10	0.6	0.50	.932	.946	.933	.947
10	0.6	1.00	.882	.953	.883	.952
10	0.6	2.00	.723	.958	.726	.957
10	0.9	0.25	.922	.948	.931	.951
10	0.9	0.50	.841	.949	.872	.950
10	0.9	1.00	.631	.951	.694	.951
10	0.9	2.00	.369	.949	.432	.949
30	0.3	0.25	.949	.947	.949	.948
30	0.3	0.50	.944	.948	.943	.948
30	0.3	1.00	.924	.950	.922	.950
30	0.3	2.00	.849	.954	.840	.955
30	0.6	0.25	.946	.948	.947	.950
30	0.6	0.50	.932	.948	.933	.949
30	0.6	1.00	.876	.950	.877	.951
30	0.6	2.00	.706	.951	.710	.950
30	0.9	0.25	.919	.949	.952	.950
30	0.9	0.50	.832	.950	.899	.950
30	0.9	1.00	.611	.950	.727	.950
30	0.9	2.00	.359	.949	.456	.950

Alternate Standardizers

Olejnik and Algina (2000) referred to the denominator of Equation 1 and Equation 6 as a *standardizer*. Other standardizers may be considered. The general method of approximating var($\hat{\delta}$) described in the Appendix may be applied to standardized linear contrasts of means using different standardizers. For instance, suppose one wanted to find the variance of $(\sum_{j=1}^{k} c_j\hat{\mu}_j)/\hat{\sigma}_1$ where $\hat{\sigma}_1$ is the estimated standard deviation in a control condition. Applying the method described in the Appendix gives

$$\text{var}(\hat{\delta}) = \hat{\delta}^2/2df_1 + \left(\sum_{j=1}^{k} c_j^2\hat{\sigma}_j^2/df_j\right)/\hat{\sigma}_1^2 \quad (12)$$

for between-subjects designs and

$$\text{var}(\hat{\delta}) = \hat{\delta}^2/2df_1 + \left(\sum_{j=1}^{k} c_j^2\hat{\sigma}_j^2 + 2\sum_{r=1}^{k}\sum_{t=r+1}^{k} c_r c_t\hat{\rho}_{rt}\hat{\sigma}_r\hat{\sigma}_t\right)/\hat{\sigma}_1^2 df \quad (13)$$

for within-subjects designs.

Olejnik and Algina (2000) also described a standardizer that uses only those variances associated with the nonzero contrast coefficients, which may be expressed as $\hat{\sigma} = (b^{-1}\sum_{j=1}^{k} h_j\hat{\sigma}_j^2)^{1/2}$, where $b = \sum_{j=1}^{k} h_j$, $h_j = 0$ if $c_j = 0$, and $h_j = 1$ if $c_j \neq 0$. Applying the method described in the Appendix gives

$$\text{var}(\hat{\delta}) = (\hat{\delta}^2/b^2\hat{\sigma}^4)\sum_{j=1}^{k} h_j\hat{\sigma}_j^4/2df_j + \left(\sum_{j=1}^{k} c_j^2\hat{\sigma}_j^2/df_j\right)/\hat{\sigma}^2 \quad (14)$$

for between-subjects designs and

$$\text{var}(\hat{\delta}) = (\hat{\delta}^2/2b^2\hat{\sigma}^4 df)\left(\sum_{j=1}^{k} h_j\hat{\sigma}_j^4 + 2\sum_{r=1}^{k}\sum_{t=r+1}^{k} h_r h_t\hat{\rho}_{rt}^2\hat{\sigma}_r^2\hat{\sigma}_t^2\right) + \left(\sum_{j=1}^{k} c_j^2\hat{\sigma}_j^2 + 2\sum_{r=1}^{k}\sum_{t=r+1}^{k} c_r c_t\hat{\rho}_{rt}\hat{\sigma}_r\hat{\sigma}_t\right)/\hat{\sigma}^2 df \quad (15)$$

for within-subjects designs.

In Block × Treatment designs in which there is a substantial difference in within-group variability across blocks, a standardized Block × Treatment interaction effect might be most meaningful if the differences among treatments at block i are standardized using only those variances from block i. For instance, in a Sex × Treatment design where males are believed to be more variable than females, an alternative way to describe the Sex × Treatment interaction is to define a standardized interaction effect as $\delta = (\mu_1 - \mu_2)/\sigma_M - (\mu_3 - \mu_4)/\sigma_F$ where $\sigma_M = [(\sigma_1^2 + \sigma_2^2)/2]^{1/2}$ and $\sigma_F = [(\sigma_3^2 + \sigma_4^2)/2]^{1/2}$. Applying the method described in the Appendix gives

$$\text{var}(\hat{\delta}) = \hat{\delta}_M^2(\hat{\sigma}_1^4/df_1 + \hat{\sigma}_2^4/df_2)/8\hat{\sigma}_M^4 + \hat{\sigma}_1^2/\hat{\sigma}_M^2 df_1 + \hat{\sigma}_2^2/\hat{\sigma}_M^2 df_2 + \hat{\delta}_F^2(\hat{\sigma}_3^4/df_3 + \hat{\sigma}_4^4/df_4)/8\hat{\sigma}_F^4 + \hat{\sigma}_3^2/\hat{\sigma}_F^2 df_3 + \hat{\sigma}_4^2/\hat{\sigma}_F^2 df_4, \quad (16)$$

where $\hat{\delta}_M$ is an estimate of $(\mu_1 - \mu_2)/\sigma_M$ and $\hat{\delta}_F$ is an estimate of $(\mu_3 - \mu_4)/\sigma_F$.

In paired-samples designs, Glass, McGaw, and Smith (1981) suggested that the standard deviation of the difference scores is an appropriate standardizer when a gain score is the response variable of interest. A generalization of this idea leads to an alternative estimate of effect size for dependent samples: $\hat{\delta}_c = \sum_{j=1}^{k} c_j\hat{\mu}_j/\hat{\sigma}_y$, where $\hat{\sigma}_y^2$ was defined in the context of Equation 2. Using the approach described in the Appendix gives

$$\text{var}(\hat{\delta}_c) = (\hat{\delta}_c^2/2 + 1)/df. \quad (17)$$

It is important to note that $\hat{\delta}_c$ is not comparable with $\hat{\delta}$ and the use of $\hat{\delta}_c$ might be justified only in very special circumstances.

The most popular standardizer is σ_p, which is used in δ'. The use of this standardizer is problematic when the

samples sizes are not equal. Its weakness can be seen most easily in the case of $k = 2$ independent samples where $\sigma_p^2 = [(n_1 - 1)\sigma_1^2 + (n_2 - 1)\sigma_2^2]/(n_1 + n_2 - 2) = (\sigma_1^2 + m\sigma_2^2)/(1 + m)$ and $m = (n_2 - 1)/(n_1 - 1)$. When $\sigma_1^2 \neq \sigma_2^2$, σ_p depends on the ratio of sample sizes, and estimates of δ' from different studies of the same design will not be comparable unless the studies use the same sample size ratios.

The variance derivations for different standardizers presented in this section are intended primarily to illustrate the flexibility of the general method described in the Appendix. The square root of an unweighted average of all variances is perhaps the most useful standardizer unless the population variances are believed to differ considerably. A standardizer that is based on only a pretest or a control group variance or only those variances corresponding to nonzero contrast coefficients might be preferred in applications where the population variances are believed to differ considerably. Although the choice of a standardizer should be made primarily in terms of effect size interpretability, there are secondary issues that also should be considered. For instance, with equal sample sizes and equal sample variances, $\text{var}(\hat{\delta})$ is larger in Equations 12–15 than in Equations 7 and 9 and hence the confidence interval based on Equation 11 tends to be wider with these alternative standardizers. Furthermore, additional simulation results suggest that n_1 should be at least 30 when $\hat{\sigma}_1$ is used as the standardizer. When $\hat{\sigma}_1$ is used as the standardizer, $\text{var}(\hat{\delta})$ contains ratios of independent variance estimates ($\hat{\sigma}_j^2/\hat{\sigma}_1^2$), which exhibit greater sampling variability than do the variance ratios ($\hat{\sigma}_j^4/\hat{\sigma}^4$ and $\hat{\sigma}_j^2/\hat{\sigma}^2$) in Equations 7 and 9, which are positively correlated and hence have smaller sampling variability. It is also interesting that the use of different standardizers in a Block × Treatment design could have important applications in meta-analysis where the blocks represent different studies. This approach would mimic currently used meta-analysis methods that combine standardized differences in which the standardizer varies across studies.

Mixed Designs

The method of approximating $\text{var}(\hat{\delta})$ described in the Appendix also may be applied to designs that have both between-subject and within-subject factors. These designs are often referred to as *mixed designs* or *split-plot designs*. Suppose k means have been estimated in a design that contains between-subjects factors, within-subject factors, or both. The variance of $\hat{\delta}$ may be expressed in the following general matrix form:

$$\text{var}(\hat{\delta}) = \hat{\delta}^2\mathbf{1}'\boldsymbol{V}\mathbf{1}/4r^2\hat{\sigma}^4 + \boldsymbol{c}'\boldsymbol{M}\boldsymbol{c}/\hat{\sigma}^2, \quad (18)$$

where $\hat{\sigma}$ is the standardizer, $\boldsymbol{V}$ is a $k \times k$ matrix of estimated variances and covariances among the k sample variances, $\boldsymbol{M}$ is a $k \times k$ matrix of estimated variances and covariances among the k sample means, $\mathbf{1}$ is a $r \times 1$ vector of ones, $\boldsymbol{c}$ is a $k \times 1$ vector of contrast coefficients, and r is the number of variances used in the standardizer. The easiest way to implement Equation 18 using currently available software is to compute sample means, variances, and covariances from a statistical package; compute $\hat{\delta}$ and $\hat{\sigma}$ by hand; and then use a matrix-based language to compute the matrix multiplications $\mathbf{1}'\boldsymbol{V}\mathbf{1}$ and $\boldsymbol{c}'\boldsymbol{M}\boldsymbol{c}$. The matrix multiplication, which is simple but tedious, also could be performed by hand if k is not too large.

To illustrate the application of Equation 18 to a mixed design, consider a two-group pretest–posttest design where μ_1 and μ_2 are the pretest and posttest means under Treatment 1 and μ_3 and μ_4 are the pretest and posttest means under Treatment 2. To obtain $\text{var}(\hat{\delta})$ for $\hat{\delta} = [(\hat{\mu}_1 - \hat{\mu}_2) - (\hat{\mu}_3 - \hat{\mu}_4)]/\hat{\sigma}$ where $\hat{\sigma}$ is computed from all four sample variances, set $r = 4$ and $\boldsymbol{c}' = [1\ -1\ -1\ 1]$. Substituting estimates of the variances and covariances of the sample means and variances into $\boldsymbol{M}$ and $\boldsymbol{V}$ gives

$$\text{var}(\hat{\delta}) = \hat{\delta}^2[(\hat{\sigma}_1^4 + \hat{\sigma}_2^4 + 2\hat{\rho}_{12}^2\hat{\sigma}_1^2\hat{\sigma}_2^2)/df_1 + (\hat{\sigma}_3^4 + \hat{\sigma}_4^4 + 2\hat{\rho}_{34}^2\hat{\sigma}_3^2\hat{\sigma}_4^2)/df_2]/32\hat{\sigma}^4 + [(\hat{\sigma}_1^2 + \hat{\sigma}_2^2 - 2\hat{\rho}_{12}\hat{\sigma}_1\hat{\sigma}_2)/df_1 + (\hat{\sigma}_3^2 + \hat{\sigma}_4^2 - 2\hat{\rho}_{34}\hat{\sigma}_3\hat{\sigma}_4)/df_2]/\hat{\sigma}^2. \quad (19)$$

Computational Examples

Five simplified examples are presented in this section to help illustrate the computation of some basic equations. After the sample means, variances, and correlations (for within-subjects designs) have been computed from some statistical package, the remaining computations illustrated in these examples may then performed on a hand calculator. Readers interested in how to interpret a confidence interval are referred to Bonett and Wright (2007).

Example 1

There are $n_1 = 55$ participants in Group 1 and $n_2 = 60$ participants in Group 2. The sample means are $\hat{\mu}_1 = 4.4$ and $\hat{\mu}_2 = 3.5$. The sample variances are $\hat{\sigma}_1^2 = 1.9$ and $\hat{\sigma}_2^2 = 2.6$. The point estimate of δ is $\hat{\delta} = (4.4 - 3.5)/[(1.9 + 2.6)/2]^{1/2} = 0.9/1.5 = 0.6$. Applying Equation 8, the estimated variance of $\hat{\delta}$ is

$$\text{var}(\hat{\delta}) = [0.6^2(3.61/432 + 6.76/472)/5.06 + 1.9/121.5 + 2.6/132.75] = 0.0368$$

and the 95% confidence interval for δ is $0.6 \pm 1.96(0.0368)^{1/2} = (0.22, 0.98)$.

Example 2

In one group of 60 participants, each participant is measured under two different treatment conditions. The sample

means are $\hat{\mu}_1 = 26$ and $\hat{\mu}_2 = 22$. The sample variances are $\hat{\sigma}_1^2 = 30$ and $\hat{\sigma}_2^2 = 20$. The sample product–moment correlation between the 60 pairs of measurements is 0.7. The point estimate of δ is $\hat{\delta} = (26 - 22)/[(30 + 20)/2]^{1/2} = 4/5 = 0.8$. Applying Equation 10, the estimated variance of $\hat{\delta}$ is

$$\text{var}(\hat{\delta}) = 0.8^2(900 + 400 + 588)/295000 + (30 + 20 - 34.29)/1475 = 0.0148$$

and the 95% confidence interval for δ is $0.8 \pm 1.96(0.0148)^{1/2} = (0.56, 1.04)$.

Example 3

A 2×2 factorial experiment has 40 participants in each of the $k = 4$ groups. The sample means are 45, 48, 37, and 42 and the sample variances are 62, 68, 74, and 52 for conditions a_1b_1, a_1b_2, a_2b_1, and a_2b_2, respectively. The point estimate of δ for the main effect of Factor A is $\hat{\delta} = [(0.5)45 + (0.5)48 + (-0.5)37 + (-0.5)42]/[(45 + 68 + 74 + 52)/4]^{1/2} = 7/8 = 0.875$. Applying Equation 7, the estimated variance of $\hat{\delta}$ is

$$\text{var}(\hat{\delta}) = (0.8^2/65536)(3844/78 + 4624/78 + 5476/78 + 2702/78) + [(0.25)62/39 + (0.25)68/39 + (0.25)74/39 + (0.25)52/39]/64 = 0.028$$

and the 95% confidence interval for δ is $0.875 \pm 1.96(0.028)^{1/2} = (0.55, 1.20)$.

Example 4

Fifty participants are each measured at $k = 3$ time periods. The sample means are 19, 15, and 17 and the sample variances are 41, 35, and 32 at Times 1, 2 and 3, respectively. The sample product–moment correlations are 0.90, 0.90, and 0.80 between Time 1 and Time 2, Time 2 and Time 3, and Time 1 and Time 3, respectively. The point estimate of δ for $\boldsymbol{c}' = [1 \ -0.5 \ -0.5]$ is $\hat{\delta} = [41 + (-0.5)35 + (-0.5)32]/[(41 + 35 + 32)/3]^{1/2} = 3/6 = 0.5$. Applying Equation 9, the estimated variance of $\hat{\delta}$ is

$$\text{var}(\hat{\delta}) = (0.5^2/1143072)[1681 + 1225 + 1024 + 2(0.9)1435 + 2(0.8)1312 + 2(0.9)1120] + [41 + (0.25)35 + (0.25)32 + 2(-0.5)34.09 + 2(-0.5)28.98 + 2(0.25)29.65]/1764 = 0.0077$$

and the 95% confidence interval for δ is $0.5 \pm 1.96(0.0077)^{1/2} = (0.33, 0.67)$.

Example 5

In one group, $n_1 = 25$ participants are measured under Treatment A and then again under Treatment B. The sample means are $\hat{\mu}_A = 13$ and $\hat{\mu}_B = 8$. The sample variances are $\hat{\sigma}_A^2 = 7$ and $\hat{\sigma}_B^2 = 11$. The sample product–moment correlation between the 25 pairs of measurements is 0.8. In a second group, $n_2 = 26$ participants are measured under Treatment B and then again under Treatment A. The sample means are $\hat{\mu}_A = 11$ and $\hat{\mu}_B = 7$. The sample variances are $\hat{\sigma}_A^2 = 8$ and $\hat{\sigma}_B^2 = 10$. The sample product–moment correlation between the 26 pairs of measurements is 0.9.The point estimate of δ is $\hat{\delta} = [(13 + 11)/2 - (8 + 7)/2]/[(7 + 11 + 8 + 10)/4]^{1/2} = 4.5/3 = 1.5$. Applying Equation 18 with $r = 4$, $\boldsymbol{c}' = [0.5 \ -0.5 \ 0.5 \ -0.5]$, and $\hat{\sigma} = 3$, the estimated variance of $\hat{\delta}$ is

$$\text{var}(\hat{\delta}) = 1.5^2[2(49 + 121 + 98.6)/24 + 2(64 + 100 + 129.6)/25]/576 + (1/4)[(7 + 11 - 14.04)/24 + (8 + 10 - 16.1)/25]/9 = 0.186$$

and the 95% confidence interval for δ is $1.5 \pm 1.96(0.186)^{1/2} = (0.66, 2.34)$.

The Normality Assumption

The normality assumption has two important implications for the analysis of standardized linear contrast of means. One implication concerns the effect of nonnormality on the performance of the confidence interval. Unlike confidence intervals for unstandardized means, the confidence intervals considered here do not become more resistant to nonnormality as the sample size increases. The confidence intervals considered here will not have coverage probabilities close to $1 - \alpha$ unless the population distribution of the response variable is at most mildly nonnormal. In this respect, the confidence intervals considered here have limitations similar to the normal-theory confidence intervals for Pearson correlations, variances, and the parameters of structural equation models. A second and more important implication concerns the interpretation of δ under nonnormality. The usefulness of δ as a measure of effect size depends on the researcher's ability to interpret its magnitude. For instance, in a two-group (treatment and control) experiment where the response variable has an approximate normal distribution, the researcher can visualize the distribution of a treated population shifted to the left or to the right δ standard deviations from the control population. The degree of separation is easily visualized when one recalls the fact that the point of inflection in a standard normal distribution (i.e., the point where the curve changes from concave down to concave up) is one standard deviation from the mean. With a picture of the standard normal curve in mind, the researcher can then easily visualize distributions that are separated by 0.25, 0.5, 1.0, and 2.0 standard deviations.

Multi-item scales that produce scores within a fixed interval, which are common in psychology (Murphy, Plake, &

Spies, 2006; Shaw & Wright, 1967), tend to have distributions that do not exhibit extreme departures from the normal distribution. Nevertheless, the normality assumption must be taken very seriously when applying any of the confidence intervals considered here. Normalizing transformations (Sokal & Rohlf, 1995, pp. 409–422) are often able to convert a highly nonnormal distribution into a mildly nonnormal distribution. Researchers are sometimes reluctant to use normalizing transformations when computing confidence intervals for means because the mean of the transformed variable is difficult to interpret. However, this should not be a concern when using a standardized measure of effect size where the metric of the response variable need not be well understood.

It is tempting to consider robust versions of δ that replace sample means and variances with robust estimates of location and dispersion. For instance, Algina, Keselman, and Penfield (2005) proposed a robust version of δ′ that replaces means with trimmed means and replaces variances with Winsorized variances. This approach addresses only one of the two problems of nonnormality discussed above. Although a bootstrap confidence interval for a robust version of δ′ performs reasonably well under nonnormality, numerical values of the robust effect size measure are difficult to interpret in terms of visualizing the degree of separation among distributions. With nonnormal distributions, the magnitude of the difference between two trimmed means as well as the magnitude of a Winsorized standard deviation depends on the specific shapes of the distributions, and the specific shapes of the distributions will almost never be known to the researcher. Consequently, a robust version of δ′ will not serve its intended purpose of providing a meaningful description of effect size because researchers will not be able to visualize the amount of separation in the distributions represented by various values of the robust effect size measure. It also should be acknowledged that many psychological measurements are bounded scales with scores restricted within a fixed range of values. Bounded scales cannot obtain the types of extreme scores that motivated Algina et al. (2005) to propose a robust version of δ′.

Unstandardized Effect Sizes

A standardized measure of effect size, such as Equation 6, may not always the best way to describe the size of an effect. If the response variable is moderately nonnormal but has a metric that is well understood, unstandardized linear contrasts of means would be preferred to standardized linear contrasts of means. Confidence intervals for linear contrasts of unstandardized means are robust to moderate amounts of nonnormality and become increasingly robust as the sample size increases. Furthermore, a confidence interval for a linear contrast of unstandardized means using Satterthwaite degrees of freedom (Snedecor & Cochran, 1980, p. 228) does not require equal population variances. If the response variable is highly skewed or leptokurtic and has a metric that is well understood, a linear contrast of medians (Bonett & Price, 2002) would be preferred to standardized or unstandardized linear contrasts of means. If the response variable is dichotomous, the confidence interval for a linear contrast of proportions developed by Price and Bonett (2004) is recommended.

When researchers develop new scales, they often perform only the most rudimentary psychometric analyses and do not provide adequate evidence of what Messick (1995) referred to as *consequential validity*. Evidence of consequential validity provides information to help understand the metric of the response variable. In the absence of adequate consequential validity, standardized measures of effect size are usually preferred to unstandardized measures. The number of psychological measurements in current use is so large and the effort required to obtain adequate consequential validity evidence for these measures is so great, it seems that there may always be a need for standardized measures of effect size in psychological research.

Concluding Remarks

Fifty years ago, the eminent statisticians George W. Cochran and Gertrude M. Cox made the following statement in their now classic text:

> In many experiments, it seems obvious that the different treatments must produce some difference, however small, in effect. Thus the hypothesis that there is *no* difference is unrealistic: The real problem is to obtain estimates of the sizes of the differences. (Cochran & Cox, 1957, p. 5)

A few years later, Rozeboom (1960) recommended to psychologists that "Whenever possible, the basic statistical report should be in the form of a confidence interval" (p. 426). Perhaps one reason why psychologists have been slow to heed Rozeboom's advice is because it is often difficult to interpret the size of an effect in a psychological study. With literally thousands of psychological scales in use and relatively little known about the metric of the scale scores, psychologists could argue that unstandardized measures of effect size do not always provide meaningful information. It is now common practice to report a point estimate of a standardized measure of effect size in applications where the metric of the response variable is not well understood. However, reporting only a point estimate of an effect size is not enough: A confidence interval for the population effect size value is needed as well. Until now, a general and simple method for constructing confidence intervals for standardized linear contrasts of means has not been available. Journal editors are implored to require a confidence interval for δ along with its point estimate in studies where δ is the most appropriate measure of effect size.

References

Algina, J., & Keselman, H. J. (2003). Approximate confidence intervals for effect sizes. *Educational and Psychological Measurement, 63,* 537–553.

Algina, J., Keselman, H. J., & Penfield, R. D. (2005). An alternative to Cohen's standardized mean difference effect size: A robust parameter and confidence interval in the two independent groups case. *Psychological Methods, 10,* 317–328.

Bird, K. D. (2002). Confidence intervals for effect sizes in analysis of variance. *Educational and Psychological Measurement, 62,* 197–278.

Bird, K. D. (2004). *Analysis of variance via confidence intervals.* Thousand Oaks, CA: Sage.

Bollen, K. A. (1989). *Structural equations with latent variables.* New York: Wiley.

Bonett, D. G., & Price, R. M. (2002). Statistical inference for linear functions of medians: Confidence intervals, hypothesis testing, and sample size requirements. *Psychological Methods, 7,* 370–383.

Bonett, D. G., & Wright, T. A. (2007). Comments and recommendations regarding the hypothesis testing controversy. *Journal of Organizational Behavior, 28,* 647–659.

Cochran, W. G., & Cox, G. M. (1957). *Experimental design* (2nd ed.). New York: Wiley.

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

Cumming, G., Fidler, F., Leonard, M., Kalinowski, P., Christiansen, A., Kleinig, A., et al. (2007). Statistical reform in psychology: Is anything changing? *Psychological Science, 18,* 230–232.

Glass, G. V., McGaw, B., & Smith, M. (1981). *Meta-analysis in social research.* Thousand Oaks, CA: Sage.

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

Keppel, G., & Wickens, T. D. (2004). *Design and analysis: A researcher's handbook* (4th ed.). Upper Saddle River, NJ: Pearson.

Kline, R. B. (2004). *Beyond significance testing.* Washington, DC: American Psychological Association.

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison approach* (2nd ed.). Mahwah, NJ: Erlbaum.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from person's responses and performances as scientific inquiry into score meaning. *American Psychologist, 50,* 741–749.

Murphy, L. L., Plake, B. S., & Spies, R. A. (2006). *Tests in print VII.* Lincoln: University of Nebraska.

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

Price, R. M., & Bonett, D. G. (2004). Improved confidence interval for a linear function of binomial proportions. *Computational Statistics and Data Analysis, 45,* 449–456.

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

Rozeboom, W. W. (1960). The fallacy of the null hypothesis significance test. *Psychological Bulletin, 57,* 416–428.

Shaw, M. E., & Wright, J. M. (1967). *Scales for the measurement of attitudes.* New York: McGraw-Hill.

Snedecor, G. W., & Cochran, W. C. (1980). *Statistical methods* (7th ed.). Ames: Iowa State University.

Sokal, R. R., & Rohlf, F. J. (1995). *Biometry* (3rd ed.). New York: Freeman.

Steiger, J. H. (2004). Beyond the *F* test: Effect size confidence intervals and tests of close fit in the analysis of variance and contrast analysis. *Psychological Methods, 9,* 164–182.

Stuart, A., & Ord, J. K. (1994). *Kendall's advanced theory of statistics* (6th ed., Vol. 1,). London: Arnold.

Viechtbauer, W. (2007). Approximate confidence intervals for standardized effect sizes in the two-independent and two-dependent samples designs. *Journal of Educational and Behavioral Statistics, 32,* 39–60.

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

Appendix

Variance of $\hat{\delta}$

An approximation of the variance of $\hat{\delta}$ that does not assume homoscedasticity may be obtained by approximating the variance of a ratio of two random variables, $\hat{\theta} = \sum_{j=1}^{k} c_j \hat{\mu}_j$ and $\hat{\sigma} = (k^{-1}\sum_{j=1}^{k} \hat{\sigma}_j^2)^{1/2}$. Assuming normality, it follows that $\hat{\theta}$ and $\hat{\sigma}$ are independent. Application of the delta method (Stuart & Ord, 1994, p. 350) gives the following approximate variance for $\hat{\theta}/\hat{\sigma}$:

$$\theta^2 \text{var}(\hat{\sigma})/\sigma^4 + \text{var}(\hat{\theta})/\sigma^2, \tag{A1}$$

where $\text{var}(\hat{\theta}) = \sum_{j=1}^{k} c_j^2 \sigma_j^2 / n_j$ in between-subjects designs with k independent samples of sizes n_j and $\text{var}(\hat{\theta}) = (\sum_{j=1}^{k} c_j^2 \sigma_j^2 + 2\sum_{r=1}^{k} \sum_{t=r+1}^{k} c_r c_t \sigma_{rt})/n$ in within-subjects designs with a single sample of size n. Note that $\sigma_{rt} = \rho_{rt}\sigma_r\sigma_t$ is the covariance between measurements r and t. Note also that the first term of Equation A1 may be expressed as $\delta^2 \text{var}(\hat{\sigma})/\sigma^2$.

Assuming normality, $\text{var}(\hat{\sigma}_j^2) = 2\sigma_j^4/n_j$, where $\hat{\sigma}_j^2$ is estimated from a sample of size n_j. Assuming bivariate normality, $\text{cov}(\hat{\sigma}_r^2, \hat{\sigma}_t^2) = 2\sigma_{rt}^2/n$ (Bollen, 1989, p. 427) where $\hat{\sigma}_r^2$ and $\hat{\sigma}_t^2$ are estimated from the same sample of size n. Application of the delta method gives $\text{var}(\hat{\sigma}) = (\sum_{j=1}^{k} \sigma_j^4/n_j)/2k^2\sigma^2$ for between-subjects designs and $\text{var}(\hat{\sigma}) = (\sum_{j=1}^{k} \sigma_j^4 + 2\sum_{r=1}^{k} \sum_{t=r+1}^{k} \sigma_{rt}^2)/(2k^2\sigma^2 n)$ for within-subjects designs. Substituting the approximations for $\text{var}(\hat{\sigma})$ and $\text{var}(\hat{\theta})$ into Equation A1 gives

$$(\delta^2/2k^2\sigma^4)\sum_{j=1}^{k} \sigma_j^4/n_j + \left(\sum_{j=1}^{k} c_j^2\sigma_j^2/n_j\right)/\sigma^2 \tag{A2}$$

for between-subjects designs and

$$(\delta^2/k^2\sigma^4 n)\left(\sum_{j=1}^{k} \sigma_j^4 + 2\sum_{r=1}^{k} \sum_{t=r+1}^{k} \sigma_{rt}^2\right) + \left(\sum_{j=1}^{k} c_j^2\sigma_j^2 + 2\sum_{r=1}^{k} \sum_{t=r+1}^{k} c_r c_t \sigma_{rt}\right)/\sigma^2 n \tag{A3}$$

for within-subjects designs using a single sample of size n. To obtain an estimate of $\text{var}(\hat{\delta})$, the unknown parameter values in Equation A2 and A3 are replaced with sample estimates.

Following Snedecor and Cochran (1980, p. 81), n_j is replaced with $n_j - 1$ in $\text{var}(\hat{\sigma})$. Hedges and Olkin (1985, p. 104) gave the exact variance of $(\hat{\mu}_1 - \hat{\mu}_2)/\hat{\sigma}_p$ in which the component corresponding to $\text{var}(\hat{\theta})/\sigma^2$ was expressed as $[(n_1 + n_2 - 2)(n_1 + n_2)]/(n_1 + n_2 - 4)$. With equal sample sizes, this term simplifies to $2(n - 1)/[n(n - 2)]$ and is well approximated by $2(n - 1)$. This suggests that replacing n_j with $n_j - 1$ in $\text{var}(\hat{\theta})$ could improve the small-sample performance of Equation 11. An improvement was in fact observed in a preliminary simulation study for between-subjects designs and also for within-subject designs.

Received June 26, 2007
Revision received January 14, 2008
Accepted January 24, 2008 ■