Contrasts and Correlations in Effect-size Estimation

General Article

CONTRASTS AND CORRELATIONS IN EFFECT-SIZE ESTIMATION

By Ralph L. Rosnow,[1] Robert Rosenthal,[2] and Donald B. Rubin[3]

[1]*Department of Psychology, Temple University;* [2]*Department of Psychology, University of California, Riverside; and* [3]*Department of Statistics, Harvard University*

This article describes procedures for presenting standardized measures of effect size when contrasts are used to ask focused questions of data. The simplest contrasts consist of comparisons of two samples (e.g., based on the independent t *statistic). Useful effect-size indices in this situation are members of the* g *family (e.g., Hedges's* g *and Cohen's* d*) and the Pearson* r*. We review expressions for calculating these measures and for transforming them back and forth, and describe how to adjust formulas for obtaining* g *or* d *from* t*, or* r *from* g*, when the sample sizes are unequal. The real-life implications of* d *or* g *calculated from* t *become problematic when there are more than two groups, but the correlational approach is adaptable and interpretable, although more complex than in the case of two groups. We describe a family of four conceptually related correlation indices: the alerting correlation, the contrast correlation, the effect-size correlation, and the BESD (binomial effect-size display) correlation. These last three correlations are identical in the simple setting of only two groups, but differ when there are more than two groups.*

Even the most cursory glance at leading research journals in psychology will reveal that authors seldom report effect sizes. One plausible explanation for this neglect may be that few researchers have a clear idea of when or how to calculate and interpret them. For example, the initial printing of the American Psychological Association's (1994) publication manual wisely recommended Cohen's *d* as one useful measure of the effect size. But the manual also prescribed certain squared indices (i.e., r^2, η^2, ω^2, R^2, ϕ^2), all of which, regrettably, are susceptible to the expository problem that small, but sometimes very meaningful, effects may appear to essentially disappear when squared, and the real-life importance of even substantial effects may be lost (e.g., Abelson, 1985; Ozer, 1985; Rosenthal & Rubin, 1982). The American Psychological Association's manual further advised that "in most cases such measures are readily obtainable whenever the omnibus test statistics (e.g., *t* and *F*) and sample sizes (or degrees of freedom) are reported" (p. 18). However, *t* is not an omnibus test but is instead a focused test, and although generally *F* tests are omnibus tests, those with one numerator degree of freedom are focused tests. Moreover, informative effect-size measures such as Hedges's *g*, Cohen's *d*, and the product-moment correlation (*r*) cannot be obtained from an omnibus *F*.

We begin by reviewing the standard calculations of *g*, *d*, and *r* in two-group designs and describe how to adjust some familiar (and perhaps not so familiar) formulas when the sample sizes are unequal. When contrasts are used to address focused questions in more than two groups, the real-life meaning of *g* or *d* calculated from *t* statistics becomes problematic, but we describe a correlational approach that is both adaptable and interpretable, although more complex than with two groups. In particular, we define and illustrate a family of four conceptually related indices, which we call the alerting correlation (symbolized r_{alerting}), the contrast correlation (r_{contrast}), the effect-size correlation ($r_{\text{effect size}}$), and the binomial effect-size correlation (r_{BESD}). This approach is fully developed in Rosenthal, Rosnow, and Rubin (2000), which includes additional equations that can be used with different raw ingredients.

THE *g* FAMILY AND ITS RELATION TO *r* IN TWO-GROUP DESIGNS

Effect Sizes in Designs With Equal Sample Sizes

Basically, contrasts are statistical procedures for asking focused questions of data. For example, when we compare $k = 2$ groups or conditions with equal sample sizes using the standard *t* statistic (or the *F* statistic with numerator $df = 1$), the statistical procedure is intrinsically "focused" because of the implicit idea that one sample (e.g., the experimental group) will have a different score on the dependent variable than the other sample (e.g., the control group).

Cohen (1965) showed that one useful option for measuring the effect size in this situation is the familiar product-moment correlation (*r*), expressed as a point-biserial correlation between dummy-coded groups or conditions (e.g., coded 1 for

Address correspondence to Ralph L. Rosnow, 177 Biddulph Rd., Radnor, PA 19087-4506; e-mail: rrosnow@nimbus.temple.edu.

experimental and 0 for control) and scores on a continuous variable. Cohen also noted that a shortcut formula for obtaining the effect-size correlation directly from *t* is

$$r = \sqrt{\frac{t^2}{t^2 + df_{\text{within}}}}, \tag{1}$$

where df_{within} is the degrees of freedom for the *t* statistic, equal to $N - 2$ in the case of two groups, and *N* is the total number of subjects in both groups. When working with *F* (with numerator $df = 1$) instead of *t*, to use Cohen's formula, we would simply substitute *F* for t^2 in Equation 1. A highly informative convention, coming out of the meta-analytic tradition, is to report the effect-size correlation as positive when the effect is in the predicted direction and as negative when the effect is in the unpredicted direction. However, when the two samples are not independent, the effect-size correlation calculated using Equation 1 is no longer the simple point-biserial correlation, but is instead the correlation between group membership and scores on the dependent variable with indicator variables for the paired individuals partialed out.

Another family of effect-size indices concentrates on the standardized difference between the sample means (M_1 and M_2). Common examples are Hedges's *g* (e.g., Hedges & Olkin, 1985) and Cohen's *d* (e.g., Cohen, 1965, 1988), both of which represent the effect size as standard-score units (*z* scores). Hedges's *g* is defined as

$$g = \frac{M_1 - M_2}{s_{\text{within}}}, \tag{2}$$

where s_{within} is the usual pooled within-sample estimate of the population standard deviation, given by

$$s_{\text{within}} = \sqrt{\frac{\Sigma(X_1 - M_1)^2 + \Sigma(X_2 - M_2)^2}{df_{\text{within}}}}, \tag{3}$$

and X_1 and X_2 are individual scores in Samples 1 and 2, respectively. Cohen's *d* uses *N* for the denominator of the estimated variance, so that *d* is estimated by

$$d = \frac{M_1 - M_2}{\sigma_{\text{within}}}, \tag{4}$$

and

$$\sigma_{\text{within}} = s_{\text{within}} \sqrt{\frac{df_{\text{within}}}{N}}. \tag{5}$$

In the same way that Cohen (1965, 1988) showed that we can obtain *r* from *t*, he showed that standardized difference measures can also be obtained from *t*. Thus, we can calculate *g* by

$$g = \frac{2t}{\sqrt{N}}, \tag{6}$$

assuming equal sample sizes in the two groups.

Cohen's and Hedges's measures of effect size can be readily transformed back and forth, whether the sample sizes are equal or unequal. Thus, we can convert *g* into *d* by

$$d = g \sqrt{\frac{N}{df_{\text{within}}}} \tag{7}$$

or *d* into *g* by

$$g = d \sqrt{\frac{df_{\text{within}}}{N}}. \tag{8}$$

Similarly, in an equal-*n* study, we can transform *g* into *r* by

$$r = \frac{g}{\sqrt{g^2 + 4\left(\frac{df_{\text{within}}}{N}\right)}}. \tag{9}$$

For example, suppose the hypothesis is that treating subjects by a particular clinical intervention (the experimental condition) will, relative to nonintervention (the control condition), result in improvement on some psychological criterion. There are 50 subjects in each of the two randomly assigned conditions, with resulting mean scores of $M_1 = 6.0$ and $M_2 = 4.8$ in the experimental and control groups, respectively, and $s_{\text{within}} = 2.0$. We calculate $t(98)$ to be 3.00, and from Equation 1 find

$$r = \sqrt{\frac{(3.00)^2}{(3.00)^2 + 98}} = .29.$$

We obtain *g* from Equation 2 by

$$g = \frac{6.0 - 4.8}{2.0} = 0.60$$

or directly from *t* (Equation 6) by

$$g = \frac{2(3.00)}{\sqrt{100}} = 0.60.$$

Transforming *g* into *d* by Equation 7 gives

$$d = 0.60 \sqrt{\frac{100}{98}} = 0.61,$$

and transforming *d* into *g* by Equation 8 gives

Table 1. *Effects of unequal sample sizes on loss of relative efficiency (Equation 10) and the effective loss of* N *(Equation 12)*

Size of subgroup		Arithmetic mean *n*	Harmonic mean *n*	Loss of relative	Effective
n_1	n_2	($\bar{n}$)	(n_h)	efficiency	loss of *N*
50	50	50	50.00	.00	0
55	45	50	49.50	.01	1
60	40	50	48.00	.04	4
65	35	50	45.50	.09	9
70	30	50	42.00	.16	16
75	25	50	37.50	.25	25
80	20	50	32.00	.36	36
85	15	50	25.50	.49	49
90	10	50	18.00	.64	64
95	5	50	9.50	.81	81
99	1	50	1.98	.96	96

$$g = 0.61\sqrt{\frac{98}{100}} = 0.60.$$

Transforming *g* into *r* by Equation 9 gives

$$r = \frac{0.60}{\sqrt{(0.60)^2 + 4\left(\frac{98}{100}\right)}} = .29.$$

Adjustments for Unequal Sample Sizes: A General Approach

As mentioned, the equation for obtaining *g* directly from *t* (Equation 6) is predicated on the assumption of a two-group design with equal sample sizes.[1] However, when the sample sizes in the two groups are unequal, the "*g* from *t*" formula will tend to underestimate the actual *g*. Furthermore, even in the presence of a large total *N*, there may be insufficient power to obtain a *p* value at some predetermined level of significance if the sample sizes are unequal (Cohen, 1988). Hsu (1993) discussed this problem in the context of two-sample tests of means, proportions, and correlations, and described how to use Cohen's (1988) power tables to estimate the maximum power attainable when the sample size in one group (n_1) is fixed and the sample size in another group (n_2) is larger than n_1. Although sample sizes smaller than 30 have often been considered acceptable in psychology, it would be difficult (power ≈ .12) for effects that are commonly characterized as "small" ($g = 0.20$) or "medium" ($g = 0.50$; power = .46) to be found significant at the .05 level when the smaller of the two samples is less than 30.

The ratio of the harmonic mean sample size (n_h) to the arithmetic mean sample size ($\bar{n}$) is a useful index of the retention of power in the unequal-*n* design relative to the equal-*n* design. Subtracting this ratio from unity will reveal the proportional loss of relative efficiency, with its implications for loss of power (Rosenthal et al., 2000), that is,

$$\text{loss} = 1 - \left(\frac{n_h}{\bar{n}}\right), \tag{10}$$

where the harmonic mean sample size in $k = 2$ samples of n_1 and n_2 size is

$$n_h = \frac{2n_1n_2}{n_1 + n_2}. \tag{11}$$

Because the harmonic mean sample size equals the arithmetic mean sample size when $n_1 = n_2$, the ratio of n_h to $\bar{n}$ is always 1.0 in equal-*n* designs, and Equation 10 therefore yields a value of zero loss in such designs. In samples of unequal sizes, the harmonic mean is less than the arithmetic mean, and so the value given by Equation 10 will increase with corresponding increases in the inequality of the sample sizes.

Table 1 illustrates this relationship with independent samples of size n_1 and n_2, when $N = n_1 + n_2$ is fixed (e.g., $N = 100$). The last column indicates the effective loss of total

1. One of the assumptions underlying the usual *t* statistic comparing two means is that the variances of the variable in the populations from which the two samples were drawn are equal. The *t* statistic still works quite well even if the variances are fairly different, especially if sample sizes are equal or nearly so. However, if both the population variances are very different and the two sample sizes are quite different, the *t* statistic used may not follow the *t* distribution very well. One approach to this problem for significance testing is to transform the data to make the variances in the two samples more nearly equal (Box, Hunter, & Hunter, 1978; Tukey, 1977); another approach (useful when suitable transformations are unavailable or ineffective, or when inference on the original scale is more meaningful) is to use Satterthwaite's approximate method (Snedecor & Cochran, 1989).

sample size, obtained by multiplying the right-hand side of Equation 10 by the total N:

$$\text{effective loss of } N = N\left[1 - \left(\frac{n_h}{\bar{n}}\right)\right], \tag{12}$$

which is relevant to considerations of cost when the cost per sampling unit is constant. For example, a 60:40 split of 100 cases is equivalent to losing 4 of 100 total cases, whereas an 85:15 split is equivalent to losing virtually half of the total sample, and a split of 99:1 is equivalent to the loss of all but 4 of 100 cases.

In unequal-*n* designs, the valid estimate of g is still given by Equation 2, and the valid estimate of d is given by Equation 4. The transformations between d and g are also still given by Equations 7 and 8. However, the expression to obtain g from t (Equation 6) needs adjustment for the loss in relative efficiency (Rosenthal et al., 2000). In an unequal-*n* design, we can obtain g from t by

$$g = \left(\frac{2t}{\sqrt{N}}\right)\sqrt{\frac{\bar{n}}{n_h}}, \tag{13}$$

and we can obtain d from t by

$$d = \left(\frac{2t}{\sqrt{df_{within}}}\right)\sqrt{\frac{\bar{n}}{n_h}}. \tag{14}$$

Of course, there will be no difference in an equal-*n* design whether we use the equal-*n* or unequal-*n* equation. For example, in the equal-*n* situation given earlier, where $t(98) = 3.00$, $n_1 = n_2 = 50$, and g calculated from Equation 6 was 0.60, the accurate g is unchanged by the use of Equation 13 because $\bar{n} / n_h = 1.0$ when sample sizes are equal.

However, in the unequal-*n* situation, the result may be dramatically different. Suppose $n_1 = 85$ and $n_2 = 15$, still with means of 6.0 and 4.8, and $s_{within} = 2.0$, resulting in $t(98) = 2.14$. Calculating g directly from the means using Equation 2 gives 0.60 (the correct value), but calculating g from t using Equation 6 (the equal-*n* equation) gives 0.43, notably less than the correct value. Using Equation 13, with its relative-efficiency-loss adjustment, gives the correct value:

$$g = \frac{2(2.14)}{\sqrt{100}}\sqrt{\frac{50}{25.5}} = 0.60.$$

When transforming g into the point-biserial effect size in an unequal-*n* design, we also need to make an adjustment in Equation 9, because the effect size now changes with the ratio $n_h/\bar{n}$, even with the same M_1, M_2, s_{within}, and N. For example, consider two very large studies with the same N and the same population values of M_1, M_2, and s_{within}, but suppose that in one study we had $n_h/\bar{n} = 1$, and in the other study we had $n_h/\bar{n} = 1/100$. Both studies will have approximately the same effect-size value for g (and d) because of the large sample size and identical population values. However, the t produced by the equal-*n* study will be approximately 10 times the t produced by the unequal-*n* study, and the value of the effect-size correlation for the equal-*n* study, calculated using Equation 9, will also be larger than that from the unequal-*n* study.

The effect-size correlation for the equal-*n* study can be obtained from g by Equation 9, but for the unequal-*n* study, we need the following modification:

$$r = \frac{g}{\sqrt{g^2 + 4\left(\frac{\bar{n}}{n_h}\right)\left(\frac{df_{within}}{N}\right)}}. \tag{15}$$

Thus, in the case of $n_h/\bar{n} = .01$, the effect-size correlation is approximately

$$r = \frac{g}{\sqrt{g^2 + 400\left(\frac{df_{within}}{N}\right)}}$$

rather than

$$r = \frac{g}{\sqrt{g^2 + 4\left(\frac{df_{within}}{N}\right)}}$$

as with the equal-*n* study. In the example with means of 6.0 and 4.8, sample sizes of 85 and 15, and valid $g = 0.60$, Equation 15 gives

$$r = \frac{0.60}{\sqrt{(0.60)^2 + 4\left(\frac{50}{25.5}\right)\left(\frac{98}{100}\right)}} = .21.$$

FOUR CORRELATION INDICES IN DESIGNS WITH MORE THAN TWO GROUPS

The Alerting Correlation

In designs with more than two groups, many researchers have the habit of using omnibus F tests that are only indirectly related to any focused question of interest. For example, suppose the researcher's hypothesis is that grade level is an effective predictor of psychological resilience. The researcher tests 10 children at each of five grade levels (6th, 7th, 8th, 9th, and

10th grades) and obtains mean performance scores of 25, 30, 40, 50, and 55, respectively, with a pooled standard deviation of 39.69. However, the omnibus test yields $F(4, 45) = 1.03$, $p = .40$, and the researcher laments the failure of the hypothesized increase in resilience.

Despite the researcher's disappointment, we can see that there was a progression in the group means. When we correlate a set of lambda (λ) coefficients of −2, −1, 0, +1, and +2 (representing the predicted linear trend, where $\Sigma\lambda = 0$) with the five group means, we find $r = .9923$. We characterize such an aggregate correlation (i.e., based on means rather than individual scores) as an "alerting correlation" ($r_{alerting}$) because it can alert us to overall trends of interest, but it would be a poor estimate of the relation between individual children's grades and resilience. Indeed, alerting correlations (i.e., based on sample means) can be substantially larger or smaller than (or even in the opposite direction from) effect-size correlations, which are based on individual scores. This aggregate correlation "alerts us" that, despite the fact the omnibus F was not significant at the desired p level, the researcher may have been too hasty in dismissing the hypothesis that grade level is an effective predictor of psychological resilience. Also, in secondary analyses of data (e.g., as in meta-analytic work), sometimes the only information available is the set of condition means, and then $r_{alerting}$ may be the only effect-size estimate available.

The omnibus F addressed the diffuse question of whether there were any differences among the five grade levels, but was insensitive to their ordinal arrangement. The number of possible permutations of five samples is 120, and any of these permutations would have yielded the same F with numerator $df = 4$. However, had the researcher computed a contrast to address the predicted linear pattern corresponding to grade levels, the statistical result would have been more informative (and more gratifying to the researcher). We can easily obtain such a contrast using a simple procedure described elsewhere (Rosnow & Rosenthal, 1995, 1996). First, we multiply the omnibus F by its numerator degrees of freedom, which gives $(1.03)(4) = 4.12$ (i.e., the largest value of F that any contrast carved out of the sum of squares for the numerator of F can possibly achieve). Then, we multiply this value by the squared alerting correlation to obtain a contrast F, which gives $(4.12)(.9846) = F(1, 45) = 4.06$ (i.e., the F for the hypothesized linear trend).

The alerting correlation is a convenient way of evaluating the "success" of any contrast, because the squared alerting correlation tells us the proportion of the between-condition sum of squares ($SS_{between}$) that is accounted for by the contrast. In this example, given $k = 5$ groups (and, therefore, 4 df between groups), the contrast far exceeds the 25% of $SS_{between}$ (i.e., 25% = the reciprocal of the df) that we might have expected by chance. Indeed, the contrast accounts for more than 98% of $SS_{between}$. Table 2 shows the analysis of variance table corresponding to this illustration. The lesson? Common as omnibus significance tests are, they may not tell us anything we really want to know. It should also be noted that the alerting correlation can be employed for any contrast, not only for contrasts examining linear trends.

Table 2. *Summary analysis of variance table reconstructed from available information in the study of children's resilience*

Source	*SS*	*df*	*MS*	*F*
Grade level	6,488	4	1,622	1.03
Linear contrast	6,395	1	6,395	4.06
Noncontrast	93	3	31	0.02
Within grade level	70,875	45	1,575	

The Contrast Correlation

As noted, when the contrast is a simple comparison between two independent groups, the effect-size correlation (hereafter denoted as $r_{effect\ size}$) is the point-biserial correlation between each subject's group membership (coded as 0 or 1) and the score on a continuous variable. The standard expression for computing $r_{effect\ size}$ from t with two groups was given by Equation 1, with $df_{within} = N - 2$. However, when the contrast is computed on more than two independent groups, $r_{effect\ size}$ is no longer a point-biserial correlation. Thus, when $k > 2$, we regard Equation 1 as the contrast correlation ($r_{contrast}$) rather than the $r_{effect\ size}$, because Equation 1 then gives the partial correlation between scores on the outcome variable and the lambdas associated with the groups, after eliminating all between-group noncontrast variation.

Therefore, with the understanding that all sources of variation other than the contrast have been removed, we obtain $r_{contrast}$ from t by

$$r_{contrast} = \sqrt{\frac{t^2}{t^2 + df_{within}}} = \frac{t}{\sqrt{t^2 + df_{within}}}. \qquad (16)$$

Because any contrast F equals t^2, in our continuing example (Table 2), Equation 16 yields

$$r_{contrast} = \sqrt{\frac{4.06}{4.06 + 45}} = \frac{2.015}{\sqrt{4.06 + 45}} = .29.$$

With $k = 2$ groups, there are no sources of noncontrast variation to be eliminated, thereby implying that $r_{contrast} = r_{effect\ size}$ in two-group designs. Similarly, if $r^2_{alerting}$ revealed that a contrast had accounted for virtually all the between-group variation in a design with three or more groups, then $r_{contrast}$ would be virtually equivalent to $r_{effect\ size}$. However, $r_{contrast}$ can be quite large, yet not be a reflection of a similarly large $r_{effect\ size}$. The reason is that $r_{effect\ size}$ is the unpartialed

correlation, whereas $r_{contrast}$ has all the noncontrast variation removed.[2] If we know that $r_{contrast}$ will be a good approximation to $r_{effect\ size}$ in designs with more than two groups, then Equation 16 is a convenient way of estimating it.

The Effect-Size Correlation

To reiterate, $r_{effect\ size}$ should be understood as the simple correlation (unpartialed) between membership in a group or condition and scores on the dependent variable. To compute this simple correlation, we treat the noncontrast between-variability as additional error variance, then

$$r_{effect\ size} = \sqrt{\frac{F_{contrast}}{F_{contrast} + F_{noncontrast}(df_{noncontrast}) + df_{within}}}, \quad (17)$$

which from the data in Table 2 gives

$$r_{effect\ size} = \sqrt{\frac{4.06}{4.06 + 0.02(3) + 45}} = .29.$$

Alternatively, we can calculate the omnibus F (i.e., $F_{between}$) from the mean squares, if that information is available, and then calculate

$$r_{effect\ size} = \sqrt{\frac{F_{contrast}}{F_{between}(df_{between}) + df_{within}}}, \quad (18)$$

which yields, in our example,

$$r_{effect\ size} = \sqrt{\frac{4.06}{1.03(4) + 45}} = .29.$$

The contrast and effect-size correlations are identical after rounding in this example because the squared alerting correlation nearly equals 1.0.

The Binomial Effect-Size Correlation

A number of strategies are possible for tying real-life implications to effect sizes in two-group designs with continuous or categorical data. For example, Cohen (1965, 1988) discussed the practical meaning of d in psychological research, and medical researchers often tie real-life meaning to categorical data by estimating odds ratios, relative risks, and risk differences (e.g., Rosenthal et al., 2000). Cohen (1965) noted that it is also possible to use phi (ϕ)—another member of the product-moment family—to reflect an effect size in a 2 × 2 table, and Rosenthal and Rubin (1982) showed how to recast any product-moment correlation into such a display, whether the original data are continuous or categorical. Called the binomial effect-size display (or BESD), its purpose is to represent $r_{effect\ size}$ in a population in which both the independent and the dependent variables are cast as dichotomous and each variable is split at its median, with row and column margins set at 100 observations. The purpose of this section is not to reintroduce the BESD, as it has been described in detail elsewhere (e.g., Rosenthal & Rosnow, 1991; Rosenthal & Rubin, 1982); rather, the purpose of this section is to describe how to generalize the use of the BESD to the situation of three or more groups (Rosenthal et al., 2000). We believe this generalization to be quite useful given the widespread use of the BESD because of its simplicity and transparency.

Table 3. *Binomial effect-size display of effect size* r = *.10*

	Level of improvement		
Condition	Above median outcome	Below median outcome	Total
New drug	55	45	100
Old drug	45	55	100
Total	100	100	200

For example, suppose $r_{effect\ size}$ = .10 in a clinical trial comparing the level of improvement in subjects who were given either a newly developed drug or a standard drug, with the drugs randomly assigned to the subjects using a simple between-groups design. Table 3 shows the corresponding BESD, in which the cell values can be interpreted as percentages. In the upper-left and lower-right cells, 55% was calculated by adding one half the value of $r_{effect\ size}$ to .50 and then multiplying by 100; 45% in the upper-right and lower-left cells was calculated by subtracting one half the value of $r_{effect\ size}$ from .50 and then multiplying by 100. The difference between 45% and 55% (when divided by 100) gives the original value of the effect-size correlation, and tells us that it is equivalent to a rate of improvement of 10% in a population in which half the subjects would receive the new drug and half would not, with the outcome variable cast as split at the median.

We turn now to the use of the BESD when there are three or more groups involved in a contrast. In this situation, it is not immediately obvious how to exhibit $r_{effect\ size}$ as a BESD or what might be gained from such a display. Under the assumption that the noncontrast sum of squares can be considered as "noise," Rosenthal et al. (2000) presented a simple way of recasting the $r_{effect\ size}$ as a BESD with real-life implications.

2. For a more concrete description of these partial correlations, including the calculation of adjusted scores for $r_{contrast}$, see Rosenthal et al. (2000). More specifically, each score will have a value predicted from the contrast and a residual from its group mean (i.e., the score minus its group's mean). The group mean will exactly equal the predicted value for that group when the alerting correlation is 1.00. Thus, the adjusted score is the original score with the between-group variation not explained by the contrast "partialed out." The correlation between the adjusted scores and the lambda coefficients is $r_{contrast}$.

That is, we assume that the contrast of interest does, in fact, capture the full predictable relation between the outcome variable (*Y*) and the treatment groups. In that case, we conceptualize the BESD as reflecting the $r_{\text{effect size}}$ that we would expect to see in a two-group replication of the current study with the same total *N*; by implication, the lower group (or treatment condition) is set at $-1\sigma_\lambda$ and the upper group (or treatment condition) is set at $+1\sigma_\lambda$, where

$$\sigma_\lambda = \sqrt{\frac{\Sigma\lambda^2}{k}} \tag{19}$$

with, as before, k = number of conditions in the contrast.

For example, in a two-group design (with lambdas of +1 and −1), we find

$$\sigma_\lambda = \sqrt{\frac{(+1)^2 + (-1)^2}{2}} = 1,$$

which tells us that the BESD will have the same conditions as the current design. In designs with more than two groups, the BESD will capture only conditions defined by $-1\sigma_\lambda$ and $+1\sigma_\lambda$, so in the five-group study of children's resilience discussed previously (with lambdas of −2, −1, 0, +1, and +2), we find

$$\sigma_\lambda = \sqrt{\frac{(-2)^2 + (-1)^2 + (0)^2 + (+1)^2 + (+2)^2}{5}} = 1.41$$

and then use this value to set the lower and upper limits for treatment conditions of the BESD. To set the lower limit, we subtract $\sigma_\lambda = 1.41$ from the mean grade of 8, and for the upper limit we add 1.41 to the mean of 8. (Table 4 shows various values of σ_λ for linear predictions in designs consisting of 2–10 ordinal conditions.)

We now need to obtain the value of the $r_{\text{effect size}}$ to be represented in our BESD. If this were a two-group design with equal sample sizes, we could estimate $r_{\text{effect size}}$ from *t* by Equation 1. If the sample sizes of our two-group design were unequal, we would calculate r_{BESD} by

$$r_{\text{BESD}} = \sqrt{\frac{t^2}{t^2 + df_{\text{within}}\left(\frac{n_h}{\bar{n}}\right)}}. \tag{20}$$

When $k > 2$, r_{BESD} is defined to be the $r_{\text{effect size}}$ that we would expect to see in a $\pm 1\sigma$ two-group replication (with stipulations noted earlier), which can be calculated by

$$r_{\text{BESD}} = \sqrt{\frac{F_{\text{contrast}}}{F_{\text{contrast}} + F_{\text{noncontrast}}(df_{\text{noncontrast}} + df_{\text{within}})}}. \tag{21}$$

If $F_{\text{noncontrast}}$ is less than 1.00, it is entered into Equation 21 as equal to 1.00. This restriction on $F_{\text{noncontrast}}$ requires the noise level underlying MS_{within} to be at least as large as the noise level of $MS_{\text{noncontrast}}$, and arises because we are viewing the noncontrast variability as the appropriate index of the noise level, which must be at least as large as the within variability.

In our continuing example, if we enter $F_{\text{noncontrast}} = 1.00$ (because it was less than 1) into Equation 21, we find

$$r_{\text{BESD}} = \sqrt{\frac{4.06}{4.06 + [1.00(3 + 45)]}} = .28,$$

which we can interpret as the $r_{\text{effect size}}$ we would expect to see in a replication that compared the resilience of 9.4th-grade children with the resilience of 6.6th-grade children, assuming the same total *N* as in the study for which we computed the r_{BESD} and equal sample sizes in the two grade levels. Table 5, interpreted in the usual way, is the BESD corresponding to this result.

CONCLUSION

We have concentrated here on effect-size estimation, favoring the *r* family when there are more than two groups. In the

Table 4. *Linear contrast weights and* σ_λ *for* k = *2 to 10 ordinal conditions*

	Ordinal conditions										
k	1	2	3	4	5	6	7	8	9	10	σ_λ
2	−1	+1									1.00
3	−1	0	+1								0.82
4	−3	−1	+1	+3							2.24
5	−2	−1	0	+1	+2						1.41
6	−5	−3	−1	+1	+3	+5					3.42
7	−3	−2	−1	0	+1	+2	+3				2.00
8	−7	−5	−3	−1	+1	+3	+5	+7			4.58
9	−4	−3	−2	−1	0	+1	+2	+3	+4		2.58
10	−9	−7	−5	−3	−1	+1	+3	+5	+7	+9	5.75

Table 5. *Binomial effect-size display of* r_{BESD} *= .28 in the study of children's resilience*

Group	Above median resilience	Below median resilience	Total
9.4th grade ($+1\sigma_\lambda$)	64	36	100
6.6th grade ($-1\sigma_\lambda$)	36	64	100
Total	100	100	200

simple setting of two groups, we have no strong preference among *d*, *g*, or *r*, all of which do a good job. In the two-group case, r_{contrast}, $r_{\text{effect size}}$, and r_{BESD} will, of course, be identical. Although when there are more than two groups it is also possible for r_{alerting}, r_{contrast}, $r_{\text{effect size}}$, and r_{BESD} to have identical values, typically $r_{\text{effect size}}$ will be larger than r_{BESD}, and r_{contrast} will be larger than $r_{\text{effect size}}$, with these differences sometimes being quite substantial. The value of r_{alerting} tends to be larger than the value of the other three indices, but need not be so (Rosnow & Rosenthal, 1996). Using this entire family of indices captures the different meanings of contrasts in a way that cannot be precisely communicated by any single measurement.

Our silence thus far on the issue of significance testing is not a tacit endorsement of dichotomous significance-testing decisions. There has been a growing realization of the failings and limitations of the rhetoric of the "accept/reject" paradigm (e.g., Cohen, 1994; Kirk, 1996; Loftus, 1996; Rosenthal & Rubin, 1985; Rosnow & Rosenthal, 1989; Schmidt, 1996; Thompson, 1996). The American Psychological Association's Task Force on Statistical Inference recently recommended that researchers report interval estimates for effect sizes involving principal outcomes (Wilkinson & the Task Force on Statistical Inference, 1999). Examining the counternull statistic, that is, the nonnull magnitude of *d*, *g*, or *r* that is supported by the same amount of evidence as is the null value of the effect size (Rosenthal & Rubin, 1994; Rosnow & Rosenthal, 1996), and its associated interval (Rosenthal et al., 2000), provides insurance against mistakenly embracing the null hypothesis.

REFERENCES

Abelson, R.P. (1985). A variance explanation paradox: When a little is a lot. *Psychological Bulletin, 97,* 129–133.

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

Box, G.E.P., Hunter, W.G., & Hunter, J.S. (1978). *Statistics for experimenters*. New York: Wiley.

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

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

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

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

Hsu, L.M. (1993). Using Cohen's tables to determine the maximum power attainable in two-sample tests when one sample is limited in size. *Journal of Applied Psychology, 78,* 303–305.

Kirk, R.E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 5,* 746–759.

Loftus, G.R. (1996). Psychology will be a much better science when we change the way we analyze data. *Current Directions in Psychological Science, 5,* 161–171.

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

Rosenthal, R., & Rosnow, R.L. (1991). *Essentials of behavioral research: Methods and data analysis* (2nd ed.). New York: McGraw-Hill.

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

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

Rosenthal, R., & Rubin, D.B. (1985). Statistical analysis: Summarizing evidence versus establishing facts. *Psychological Bulletin, 97,* 527–529.

Rosenthal, R., & Rubin, D.B. (1994). The counternull value of an effect size: A new statistic. *Psychological Science, 5,* 329–334.

Rosnow, R.L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist, 44,* 1276–1284.

Rosnow, R.L., & Rosenthal, R. (1995). "Some things you learn aren't so": Cohen's paradox, Asch's paradigm, and the interpretation of interaction. *Psychological Science, 6,* 3–9.

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.

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

Snedecor, G.W., & Cochran, W.G. (1989). *Statistical methods* (8th ed.). Ames: Iowa State University Press.

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher, 25*(2), 26–30.

Tukey, J.W. (1977). *Exploratory data analysis.* Reading, MA: Addison-Wesley.

Wilkinson, L., & the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals. *American Psychologist, 52,* 685–699.

(RECEIVED 1/23/00; REVISION ACCEPTED 4/24/00)