Effect Sizes for Experimenting Psychologists

RALPH L. ROSNOW, EMERITUS, Temple University
ROBERT ROSENTHAL, University of California, Riverside

Abstract This article describes three families of effect size estimators and their use in situations of general and specific interest to experimenting psychologists. The situations discussed include both between- and within-group (repeated measures) designs. Also described is the counternull statistic, which is useful in preventing common errors of interpretation in null hypothesis significance testing. The emphasis is on correlation (*r*-type) effect size indicators, but a wide variety of difference-type and ratio-type effect size estimators are also described.

Until quite recently in the history of experimental psychology, when researchers spoke of "the results of a study," they almost invariably were referring to whether they had been able to "reject the null hypothesis," that is, to whether the *p* values of their tests of significance were .05 or less. Spurred on by a spirited debate over the failings and limitations of the "accept/reject" rhetoric of the paradigm of null hypothesis significance testing, the American Psychological Association (APA) created a task force to review journal practices and to propose guidelines for the reporting of statistical results. Among the ensuing recommendations were that effect sizes and interval estimates (e.g., confidence intervals) be reported for principal outcomes (Wilkinson & Task Force on Statistical Inference, 1999). The fifth edition of the *Publication Manual of the American Psychological Association* (APA, 2001) emphasizes the importance of reporting effect size indicators for "one-degree-of-freedom effects – particularly when these are the results that inform the discussion" (p. 26). Examples of such effects are those naturally associated with focused statistical procedures, including all *t* tests, *F* tests with numerator $df = 1$, *z* contrasts on proportions, and 1-*df* χ^2 tests. The expectation is that guidelines promulgated by the APA task force will be absorbed into the mainstream of psychological experimentation and reflected in statistical training practices. Thus, it will be the degree of the relationship between the independent and dependent variables, the effect size (i.e., the magnitude of the research findings), that will become the primary coin of the realm when psychological experimenters speak of "the results of a study." Ideally, there will also be an indication of the accuracy or reliability of the estimated effect size, which would be indexed by an interval estimate placed around the effect size estimate.

The thrust of this article is the description of various ways of estimating effect sizes in situations of general and specific interest to experimenting psychologists. The cases discussed include between- and within-group (repeated-measures) designs, continuous and categorical data, and an effect size for comparing effect sizes. Table 1 lists by family and subtype the effect size estimators described in this article. By and large, we prefer the correlation (*r*-type) family of effect size indices (Rosenthal, Rosnow, & Rubin, 2000). However, it seems natural to employ certain difference-type indices (such as Hedges's *g* and Cohen's *d*) when the original studies have compared two groups and the difference between means and within *S* (or σ) are available. It is sometimes necessary to make a decision to convert all the effect size indices to one particular index (e.g., in meta-analytic work), usually to *r* or z_r for the correlation family, or to Cohen's *d* or Hedges's *g* for the difference family. In that situation, there are reasons to view *r*-type indices as the more generally useful effect size measures (Rosenthal, 2000).

Suppose the data came to us as *r*s. We would not want to convert *r*s to difference-type indices, as the concept of a mean difference index makes little sense in describing a linear relationship over a great many values of the independent variable. On the other hand, given effect sizes that are reported as Hedges's *g* or Cohen's *d*, for example, the *r* index makes perfectly good sense in its point-biserial form (i.e., two levels of the independent variable). If the data were structured

TABLE 1
Three Families of Effect Size Estimators Discussed in This Article

Family	Subtype: Raw	Subtype: Standardized	Subtype: Transformed
Difference	M_1 - M_2 (raw difference) Cohen's *g* (14) Π (16) *d'*, Risk Difference (RD) (21)	Hedges's *g* (2) Cohen's *d* (4) Glass's Δ (6) BESD-based RD (24)	probit *d'* (10) logit *d'* (11) Cohen's *b* (15) Cohen's *q* (18)
Correlation	r_ϕ (9) $r_{equivalent}$ (12) $r_{contrast}$ (26-30) $r_{alerting}$ (31) $r_{effect\ size}$ (32) r_{BESD} (34) $r_{counternull}$ (39)	Fisher z_r	
Ratio	Relative Risk (RR) (19) Odds Ratio (OR) (20)	BESD-based RR (22) BESD-based OR (23)	

Note. Numbers in parentheses refer to equation numbers defining these estimators. Fisher's z_r is the log transformation of *r*, that is, $log_e\left[(1 + r)/(1 - r)\right]$.

in a 2 x 2 table of counts, then the phi form of the *r* index would be suitable. But suppose a hypothesis called for five levels of arousal, and the experimenter predicted better performance on the dependent measure in the middle levels of arousal than in the more extreme levels, and the very best performance in the midmost level of arousal. The magnitude of an effect associated with a curvilinear trend is quite naturally indexed by *r*, but not so naturally by difference-type or ratio-type indices. To represent the predicted quadratic trend, the experimenter could choose contrast weights (λs) of -2, +1, +2, +1, -2. The experimenter's effect size *r* would index the degree to which these λs accurately predicted the actually obtained performance. Still another convenience of the *r*-type index is that it requires no computational adjustment in going from the two-sample or multisample case to the one-sample case, whereas with Cohen's *d* or Hedges's *g*, the definition of the size of the study will change by a factor of 2 in going from a *t* test for two samples to a *t* test for one sample. Finally, the magnitude of the research findings can be quite simply interpreted on the basis of *r*-type indices (as illustrated later).

The *r*-type family includes the Pearson product-moment correlation in any of its customary incarnations, such as those discussed in this article. These include (a) the point-biserial correlation (r_{pb}) when one variable is continuous and one variable is dichotomous; (b) the phi coefficient (ϕ or r_ϕ) when both variables are dichotomous; (c) other variants such as $r_{contrast}$, $r_{alerting}$, $r_{effect\ size}$, r_{BESD}, and $r_{counternull}$; (d) the Fisher *z* transformation of *r* (z_r); and (e) squared indices of *r* and *r*-like quantities such as r^2, ω^2, ε^2, and η^2. One reason to avoid the use of squared indices of effect size, however, is that they lose their directionality (i.e., is the treatment helping or hurting, or is the correlation positive or negative?), and thus are of little use in scientific work for which information on directionality is essential. Another problem is that the implications of squared indices of effect size are likely to be misconstrued as being much less important than is often true. A little further on, we will illustrate how r^2 is susceptible to the expository problem that very small, but sometimes quite meaningful, effects may seem to essentially disappear.

The APA publication manual also recommends that authors of research reports provide the exact probabilities (*p* values) of their tests of significance (i.e., except in tables where it would be cumbersome to do so). Thus, it is useful to reiterate the general relationship between the *p* value and the effect size (Cohen, 1965; Rosenthal & Rosnow, 1991), which is given by

significance test = effect size x study size,

a relationship that is described in detail in the pages that follow. To anticipate, the larger the study in terms of the total number (*N*) of units or observations, or the larger the effect size, the larger will be the value of the significance test and, therefore, the smaller (and more

coveted) the *p* value. As sample sizes (*n*s) become increasingly unequal, the significance test will become less efficient. Simple procedures and adjustments are available for estimating the relative loss of power in unequal-*n* designs, for estimating Hedge's *g* or Cohen's *d* from *t*, and for converting Hedges's *g* or Cohen's *d* to *r* when working with unequal *n*s (Rosenthal et al., 2000; Rosnow, Rosenthal, & Rubin, 2000). Of course, if the size of effect is truly zero, increasing *N* will not produce a result that is any more significant than a smaller *N* will produce (although effect sizes of exactly zero are rarely encountered).

The Two-Sample Case

Perhaps the most commonly employed experimental design is the two-sample case (e.g., an experimental and a control condition), and suppose the researcher were interested in *t* as a test of significance. The experimenter has a choice of equations that can be written in the style of the general relationship between effect size and study size (Rosenthal, 1991, 1994). For example,

$$t = g \times \sqrt{\frac{n_1 n_2}{n_1 + n_2}}, \tag{1}$$

where the effect size component is indicated as Hedges's *g*, which in turn is defined as

$$\text{Hedges's } g = \frac{M_1 - M_2}{S_{\text{pooled}}}, \tag{2}$$

a ratio of the difference between two means divided by the combined estimate of the standard deviation (e.g., Hedges & Olkin, 1985). Equation 2 implies that one way to maximize *t* would be to select an experimental treatment that would drive M_1 and M_2 further apart. For example, were we to investigate the effects of after-school tutoring, we might use 5 hours or 2 hours of tutoring per week, but certainly not 5 minutes or 2 minutes per week. In other words, "maximizing *t*" (or whatever statistical procedure is used) is simply part of sensible experimental design. As Equation 2 also implies, another strategy would be to draw the units from a relatively homogeneous population (i.e., in characteristics that are presumably correlated with the dependent variables) so as to minimize *S*. Other relevant experimental design features, including threats to valid inferences about the existence and magnitude of presumed causal generalizations, are described by Shadish, Cook, and Campbell (2002).

In fact, any test of significance can be obtained by one or more definitions of effect size multiplied by one or more definitions of study size. For instance, another way to think about *t* in the style of the relationship between effect size and study size is

$$t = d \times \frac{\sqrt{df}}{2}, \tag{3}$$

where the effect size component is now indicated as Cohen's *d*. This popular effect size index can be estimated by

$$\text{Cohen's } d = \frac{M_1 - M_2}{\sigma_{\text{pooled}}}, \tag{4}$$

where the difference between independent means is divided by the pooled population standard deviation (Rosenthal & Rosnow, 1991), and

$$\sigma_{\text{pooled}} = S_{\text{pooled}} \sqrt{\frac{df}{N}}. \tag{5}$$

In sum, both Hedges's *g* and Cohen's *d* represent the effect size in standard-score units (i.e., *z* scores). However, Cohen's *d* uses *N* for the denominator of the estimated variance to obtain the standard deviation, whereas Hedges's *g* uses *N* - 1, that is, the pooled within-sample unbiased estimate of the population variance to obtain the standard deviation. Also listed in Table 1 is another standardized difference index, Glass's Δ (Glass, McGaw, & Smith, 1981), which is defined as

$$\text{Glass's } \Delta = \frac{M_1 - M_2}{S_{\text{control}}}, \tag{6}$$

where S_{control} is like the *S* in the denominator of Hedges's *g*, but is computed only for the control group. Like Hedge's *g*, Glass's Δ is an inferential measure, whereas Cohen's *d* is descriptive of a population of scores.

Yet another way of thinking about the relationship between effect size and study size is reflected in the following identity:

$$t = \frac{r}{\sqrt{1 - r^2}} \times \sqrt{df}, \tag{7}$$

where, in the two-group case with one dichotomous and one continuous variable, *r* would be the Pearson point-biserial correlation (r_{pb}) and $df = N - 2$.

When dichotomously scored independent and dependent variables are employed, still another way of representing the relationship between the effect size and study size is

$$\chi^2_{(1)} = \phi^2 \times N, \tag{8}$$

where χ^2 is based on a 2 x 2 table of independent frequencies (counts), ϕ^2 is the squared Pearson product-moment correlation between membership in the row category (scored 1 or 0) and membership in the column category (scored 1 or 0), and N is the total number of counts in all four cells. Rearrangement of Equation 8 defines the effect size estimator as

$$r_\phi = \sqrt{\frac{\chi^2_{(1)}}{N}}, \tag{9}$$

that is, the Pearson product-moment r when independent and dependent variables are both dichotomous.

Another difference-type index is the raw difference between population proportions, called the Risk Difference (RD) in biomedical research, similar to what Fleiss (1994) termed *d'*. Still another option is Cohen's *h*, the difference between arcsin transformed population proportions. We will have more to say about RD and *h* shortly, but other relevant difference-type indices are the probit and logit transformations of proportions, which are used to index the difference between independent categorical variables, assuming there are no refined metric scales for the outcome variable but a makeshift dichotomy is feasible (Glass, McGaw, & Smith, 1981). As examples, Glass et al. (1981) mentioned "dropping out versus persisting in school" and "remaining sober versus resuming drinking" (p. 136). The probit *d'* is defined as

$$\text{probit } d' = z_{p_1} - z_{p_2}, \tag{10}$$

that is, the difference between standard normal deviate transformed proportions (p_1 and p_2) of two populations. Equation 10 is predicated on the assumption that experimental and control group scores are distributed normally, and that it is possible to constitute an underlying metric in which a dichotomous cutoff point is presumed. The logit *d'* is defined as

$$\text{logit } d' = \log_e\left(\frac{p_1}{1-p_1}\right) - \log_e\left(\frac{p_2}{1-p_2}\right), \tag{11}$$

or the difference between logit transformed population proportions; this index is predicated on the assumption of skewness in an expected direction. For detailed discussion of the probit and logit transformations, see Glass et al. (1981).

Another way to measure differences requires no special formulas or adjustments. Suppose the dependent variable were the daily number of cigarettes smoked by experimental and control subjects. The raw difference between M_1 and M_2 (where M is the mean number of cigarettes in each condition) is meaningful in and of itself. Another intrinsically meaningful variable would be absences from work. Say an experimenter is interested in comparing a method of vocational rehabilitation against a control, and the experimenter notes the days that employees were reported as absent from work. Finding, for instance, that workers in the control condition averaged five more absences per month than those in the rehabilitation treatment becomes fraught with practical significance.

Before leaving the two-sample case, we should mention that another common situation occurs when experimenters who use very small samples report nonparametric statistics (such as the Mann-Whitney *U*) and accurate *p* values but not effect sizes. Rosenthal and Rubin (in press) described how to estimate the effect size by simply identifying (in a standard table, for example) the value of *t* that corresponds to the accurate *p* with $df = N - 2$, and substituting in the following equation:

$$r_{\text{equivalent}} = \sqrt{\frac{t^2}{t^2 + df}}, \tag{12}$$

where *r* is equivalent to the sample point-biserial correlation between the treatment indicator and an exactly normally distributed outcome in a two-treatment experiment with $N/2$ units in each group and the obtained *p* value.

The One-Sample Case

There are situations, however, in which there is only a single sample in the experiment, perhaps with each subject exposed to two different experimental conditions. For instance, an experimenter is interested in teachers' favourableness of nonverbal behaviour toward children for whom the teachers hold more versus less favourable expectations, or in the health outcomes to the same patients of a new drug versus a placebo taken at different points in time. A simple test of significance of the effect of teachers' expectations on their nonverbal behaviour, or of patients' reactions to the two different treatments, would be the *t* for correlated observations. Equation 12 can be used to obtain the *r* associated with the one-sample *t*-test, because the *r* index in the two-sample case is identical to that in the one-sample case (Rosenthal, 1994). In a similar way, we can obtain Cohen's *d* from the following:

$$d = \frac{t}{\sqrt{df}}, \tag{13}$$

where *t* is based on the one-sample case.

Dichotomous Data

When the one-sample data are dichotomous rather than continuous, three possible effect size indices are Cohen's *g*, Cohen's *h*, and a newer index, Π. Cohen's *g* is simply

$$\text{Cohen's } g = p_{\text{observed}} - .50, \tag{14}$$

the raw difference between an observed proportion and .50. For example, the magnitude of an electoral victory could be given directly by Cohen's *g*. If 60% of the electorate voted for the winner, then Cohen's *g* = .60 - .50 = .10. Such an effect size might be regarded as enormous in the case of an election result, but might be viewed as far less noteworthy as the result of an experimental intervention to boost scores of a class of high school students on a true-false test on Canadian or American history.

Cohen's *h*, noted previously in connection with the two-sample case, is defined as

$$\text{Cohen's } h = \left(\arcsin\ p_1\right) - \left(\arcsin\ p_2\right), \tag{15}$$

or the difference between arcsin transformed population proportions. Suppose that participants were asked to identify four expressions of emotions (e.g., joy, disappointment, anger, and fear) in a series of posed photographs. On each trial, they are presented with a photograph, and the instruction is to choose one of four responses (resembling a multiple-choice test in which one of four answers is always correct and its position assigned at random). Guessing should yield an accuracy rate of .25, and say the actual observed performance is .75. To transform each proportion (*p*), we calculate 2 arcsin $\sqrt{p}$, which yields 2 arcsin $\sqrt{.75}$ for the actual proportion and 2 arcsin $\sqrt{.25}$ for the expected proportion, and Cohen's *h* is 2.09 - 1.05 = 1.04. The reason for the arcsin transformation is to make the *h*s comparable, since differences between raw proportions are not all comparable (e.g., with respect to power). Thus, a difference between proportions of .95 and .90 would yield a Cohen's *h* of .19, whereas the difference between proportions of .55 and .50 yields a Cohen's *h* of only .10 (Cohen, 1988).

The one-sample effect size index Π is expressed as the proportion of correct guesses if there had been only two choices from which to select. When there are more than two choices, the Π index converts the proportion of hits to the proportion of hits made if there had been only two equally likely choices, where

$$\Pi = \frac{p_{\text{hits}}\left(k-1\right)}{p_{\text{hits}}\left(k-2\right)+1}, \tag{16}$$

with p_{hits} = raw proportion of hits, and *k* = number of alternative choices available. The standard error of Π is

$$SE_{\Pi} = \frac{1}{\sqrt{N}}\left[\frac{\Pi\left(1-\Pi\right)}{\sqrt{p_{\text{hits}}\left(1-p_{\text{hits}}\right)}}\right]. \tag{17}$$

This index would be especially valuable in evaluating performance on a multiple-choice dependent measure in which the number of alternatives varied from item to item. The Π index allows us to summarize the overall performance so we can compare performance on tests made up of varying numbers of alternatives per item. Further details can be found in Rosenthal and Rubin (1989, 1991) and Schaffer (1991).

Effect Sizes for Comparing Effect Sizes

Suppose an experimenter hypothesized that two cognitive performance measures will be more highly correlated in preschoolers than in fifth-graders. Cohen's *q* allows us to assess the degree to which the hypothesis is supported by simply calculating the difference between Fisher z_r-transformed *r*s obtained from the preschoolers and fifth-graders, that is,

$$\text{Cohen's } q = z_{r_1} - z_{r_2}. \tag{18}$$

Tables for Fisher z_r transformations are generally available (e.g., Rosenthal & Rosnow, 1991; Rosnow & Rosenthal, 2002a). This transformation makes equal differences between the z_rs equally detectable, whereas equal differences between the *r*s would not be equally detectable. Significance tests among *r*s are also more accurate when the Fisher z_r transformation is used (Alexander, Scozzaro, & Borodkin, 1989).

It should be noted, however, that evaluating whether two effect size *r*s are significantly different will typically require substantially more sampling units than testing whether an obtained effect size *r* is statistically significant (Cohen, 1988; Rosenthal & Rosnow, 1991). Say we wanted to achieve a power level of .80 in trying to detect at *p* = .05 (two-tailed) a difference of .10 between r_1 and r_2; we would need about 1,600 units (*n*) in *each* sample. On the other hand, if we wanted to test whether an *r* of .10 was different from zero at *p* = .05 (two-tailed) with power of .80, we would need a total *N* of about 800 units.

Cohen's *q* can also be used in the one-sample case, such as when an obtained effect size *r* is to be compared to a theoretical value of *r*. All that is required is to take the difference between z_r associated with the observed sample and z_r associated with the theoretical value of *r* (Cohen, 1988).

TABLE 2
Vaccination Status and Diagnostic Class of 401,974 Children in 1954 Salk Vaccine Trials

Condition	Paralytic polio present	Paralytic polio absent
1. Raw counts in four conditions		
Vaccination	33	200,712
Placebo	115	201,114
2. Percentages in four conditions		
Vaccination	0.016	99.984
Placebo	0.057	99.943
3. Binomial effect size display of r = .011		
Vaccination	49.5	50.5
Placebo	50.5	49.5
Total	100.0	100.0

The Interpretation of Effect Sizes

Despite growing awareness of the importance of estimating effect sizes, there remains a problem in how to interpret popular effect size estimators (Cooper, 1981). Rosenthal and Rubin (1982) found that neither experienced behavioural researchers nor experienced statisticians had a good intuitive feel for the meaning of common effect size estimators, and this was particularly true for such squared indices as r^2, η^2, ω^2, ε^2. To provide a real-world perspective on this problem, we will consider some important biomedical findings, starting with the 1954 Salk vaccine trial – called "the biggest public health experiment ever" (Meier, 1988, p. 3).

The Salk Trial in the Context of Other Effect Sizes

The purpose of this famous biomedical experiment was to evaluate the effects of inoculating young children with the Salk poliomyelitis vaccine versus a placebo consisting of a simple salt solution (Francis, Korns, Voight, Boisen, Hemphill, Napier, & Tolchinsky, 1955). There were serious problems with the study, however, which were discussed by Brownlee (1955) in a prominent journal. Originally, it had been proposed that only second-grade children receive the vaccine and that first-grade and third-grade children serve as an "observed control" group. Once the folly of this proposal was realized, a second plan proposed that all the children be combined, with half of them blindly receiving a placebo solution. But, as Brownlee caustically commented, "only 41 per cent of the trial was rescued and the remaining 59 per cent blundered along its stupid and futile path" (p. 1007). Nonetheless, he concluded that there was "convincing evidence for the effectiveness of the vaccine" (Brownlee, 1955, p. 1010).

And what was the magnitude of the effect that Brownlee found so convincing? Psychologists have grown accustomed to referring to *r*s of .10, .30, and .50 as small, moderate, and large, respectively. For example, Smith, Glass, and Miller (1980) found out meta-analytically that the average effect size of psychotherapy outcome studies was r = .39 (a moderate-to-large effect). Might we expect that the magnitude of the effect in the Salk vaccine trial was perhaps much larger than .39? To find the answer, we compute a 1-*df* chi-square on the raw data in Table 2 (i.e., the data that Brownlee found so persuasive) and then use Equation 9 (with N = 401,974) to obtain the effect size. We find that χ^2 = 45.25, p = 1.7^{-11} and the effect size r = .011, and thus the corresponding r^2 = .000 or, to four decimal places, .0001.

Most experimenters would be surprised to learn that an effective biomedical intervention could be associated with an r as small as .011 and an r^2 of .0001. But *r*s smaller than .10 are not at all unusual in biomedical research. In 1987, at a specially called meeting, it was decided to end, prematurely, a randomized double-blind experimental study of the effects of aspirin on reducing heart attacks (Steering Committee of the Physicians Health Study Research Group, 1988). The reason for the unusual termination was that it had become clear that aspirin prevented heart attacks (and death from heart attacks), and thus it would have been unethical to continue to give half of the approximately 22,000 physician research subjects a placebo. The effect size index r_ϕ was .034 (an r^2 of .0011). Table 3 lists effect sizes obtained in a convenience sample of 23 studies, including those previously summarized by Rosenthal (2000). Nine of these studies employed dependent variables of paralytic polio, convulsions, AIDS events, alcohol problems, heart attacks, and death, with associated *r*-type effect sizes of less than .10. One result of our consideration of these biomedical effect size estimates is to make us more sanguine about the magnitude and importance of research findings in the behavioural and social sciences (Rosenthal, 1995, 2000). Although the effect size is mathematically determined by characteristics of the study design and results, the interpretation of its real-life implications would, of course, depend upon the context (e.g., Rosnow & Georgoudi, 1986) and the nature of the dependent variable.

The Binomial Effect Size Display

Returning to Table 2, the bottom subtable recasts the r that we estimated from the chi-square on the Salk vaccine data as a Binomial Effect Size Display (BESD). This standardized display gives us an idea of the practical value of any effect indexed by a correlation coefficient. The BESD shows the *r*-type effect size to be a

TABLE 3
Effect Sizes of Various Independent Variables

Independent variable	Dependent variable	r	r^2
Salk vaccine[a]	Paralytic polio	.01	.00
Aspirin[b]	Heart attacks	.03	.00
Beta carotene[c]	Death	.03	.00
Streptokinase[d]	Death	.03	.00
Propranolol[e]	Death	.04	.00
Magnesium[f]	Convulsions	.07	.00
Vietnam veteran status[g]	Alcohol problems	.07	.00
Garlic[h]	Death	.09	.01
Indinavir[i]	Serious AIDs events	.09	.01
Testosterone[j]	Adult delinquency	.12	.01
Compulsory hospitalization versus treatment choice[k]	Alcohol problems	.13	.02
Cyclosporine[l]	Death	.15	.02
Low dose warfarin[m]	Blood clots	.15	.02
Ganzfeld perception[n]	Accuracy	.16	.03
Cisplatin and Vinblastine[o]	Death	.18	.03
AZT for neonates[p]	HIV infection	.21	.04
Cholesterol-lowering regimen[q]	Coronary status	.22	.05
AZT[r]	Death	.23	.05
Treatment choice versus AA[s]	Alcohol problems	.27	.07
Psychotherapy[t]	Improvement	.39	.15
Compulsory hospitalization versus AA[u]	Alcohol problems	.40	.16
Anxiety[v]	Rumormonger	.48	.23
Progesterone[w]	SIV infection	.65	.42

[a]Francis, Jr. et al. (1955); [b]Steering Committee of the Physicians Health Study Research Group (1988); [c]Alpha-Tocopherol, Beta Carotene Cancer Prevention Study Group (1994); [d]GISSI (1986); [e]Kolata (1981); [f]Foreman (1995); [g]Centers for Disease Control Vietnam Experience Study (1988); [h]Goldfinger (1991); [i]Knox (1997); [j]Dabbs and Morris (1990); [k]Cromie (1991); [l]Canadian Multicentre Transplant Study Group (1983); [m]Grady (2003); [n]Chandler (1993); [o]Cromie (1990); [p]Altman (1994); [q]Roberts (1987); [r]Barnes (1986); [s]Cromie (1991); [t]Smith, Glass, and Miller (1980); [u]Cromie (1991); [v]Rosnow (1991); [w]Contraceptive trials set for a link to AIDS risk (1996).

TABLE 4
Other Examples of Binomial Effect Size Displays

Aspirin's effect on prevention of heart attack (r = .034)		
Condition	Heart attack	No heart attack
Aspirin	48.3	51.7
Placebo	51.7	48.3
Vietnam service and alcohol problems (r = .07)		
Vietnam veteran	Problem	No problem
Yes	53.5	46.5
No	46.5	53.5
Low doses of warfarin on prevention of blood clots (r = .152)		
Condition	Blood clots	No blood clots
Warfarin	42.4	57.6
Placebo	57.6	42.4
Benefits of psychotherapy (r = .39)		
Condition	Less benefit	Greater benefit
Psychotherapy	30.5	69.5
Control	69.5	30.5

TABLE 5
Three Hypothetical Examples of Four Effect Size Estimates

	Die	Live	Relative	Odds	Risk	
Control	A	B	risk	ratio	difference	r_ϕ
Treatment	C	D	(Eq. 19)	(Eq. 20)	(Eq. 21)	(Eq. 9)
Study 1						
	Die	Live				
Control	10	990	10.00	10.09	.01	.06
Treatment	1	999				
Study 2						
	Die	Live				
Control	10	10	10.00	19.00	.45	.50
Treatment	1	19				
Study 3						
	Die	Live				
Control	10	0	10.00	∞	.90	.90
Treatment	1	9				

simple difference in outcome rates between the experimental and control groups in a 2 x 2 table with rows and columns always totaling 100 (Rosenthal & Rubin, 1982). Given that the success rate is higher in the treatment group than in the control, the BESD is obtained from any r-type effect size by computing the treatment condition success rate as 100(.50 + r/2) and the control condition success rate as 100(.50 - r/2). Thus, an r of .011 in the Salk trial yields a vaccination success rate (i.e., paralytic polio absent) of 100(.50 + .005) = 50.5, and a placebo success rate of 100(.50 - .005) = 49.5. The difference between these rates divided by 100 is .01, the effect size indexed by r rounded to two decimal places

Table 4 illustrates the BESD for four different effect sizes, starting with the aspirin study mentioned above, in which the effect size index r_ϕ was .034 for heart attack. The aspirin success rate was therefore 100(.50 + .017) = 51.7, and the placebo success rate, 100(.50 - .017) = 48.3. Experimental psychologists are not used to thinking of rs as small as .034 (aspirin study) or .011 (Salk vaccine study) as implying effect sizes of any practical importance. But when we think of an r of .034 as reflecting a 3.4% decrease in heart attacks, or an r of .011 as reflecting a 1.1% decrease in paralytic polio, these rs do not appear to be quite so "small." The second BESD shown in Table 4 refers to a nonexperimental study of 4,462 Army veterans of the Vietnam War era (1965-1971). The correlation between having served in Vietnam (as opposed to serving elsewhere) and subsequent alcohol abuse or dependence was r = .07 (Centers for Disease Control, 1988). In other words, it is equivalent to the difference between the problem rates of 53.5 and 46.5 per 100. The third display of Table 4 shows the BESD for the results of a recent clinical trial of low doses of warfarin on the prevention of blood clots in 508 high-risk patients at 52 hospitals in the United States, Canada, and Switzerland (Grady, 2003), where r = .152 (an r^2 of .023). This is another example of a biomedical study with results that were considered so dramatic as to lead to its premature termination on the ethical grounds that it would be improper to continue to administer a placebo to the control group patients. The bottom display of Table 4 reflects Smith et al.'s (1980) meta-analytic finding of r = .39 for the effect size of psychotherapy (i.e., substantially greater than the effects of a good many breakthrough biomedical interventions).

Relative Risk

We turn now to three popular effect size estimators in biomedical research, beginning with relative risk (RR). With reference to the cells labeled A, B, C, D in Table 5 (Rosenthal, 2000), relative risk is defined as

$$RR = \left(\frac{A}{A+B}\right) / \left(\frac{C}{C+D}\right), \qquad (19)$$

that is, the ratio of the proportion of the control patients at risk to the proportion of treated patients at risk. Applied to the Salk vaccine trial (Table 2), but with cells arranged as shown in Table 5, the relative risk is (115/201,229)/(33/200,745) = 3.48.

A limitation of this effect size estimate can be seen in Table 5. We ask readers to examine the three study outcomes closely and to ask themselves the following question: "If I had to be in the control condition, would it matter to me whether I was in Study 1, Study 2, or Study 3?" We think most people would rather have been in Study 1 than Study 2. We also think that virtually no one would prefer to be a member of the control group in Study 3. Yet, despite the very important phenomenological differences among these three studies, Table 5 shows that all three relative risks are identical: 10.00. That feature may be a serious limitation to the value and informativeness of the relative risk index.

The Odds Ratio

With A, B, C, D cells defined in Table 5, the odds ratio (OR) is

$$OR = \left(A / B\right) / \left(C / D\right), \qquad (20)$$

that is, the ratio of the not-surviving control patients to the surviving control patients divided by the ratio of the not-surviving treated patients to the surviving treated patients. Applied to the Salk vaccine trial (Table 2), with the arrangement of the cells conforming to the template in Table 5, the odds ratio is (115/201,114)/(33/200,712) = 3.48.

Notice in Table 5 that the odds ratio behaves more as expected than does the relative risk. That is, the OR increases with our phenomenological discomfort as we go from the results of Study 1 to Study 2 to Study 3. But the high odds ratio for Study 1 seems alarmist. Suppose the data were as shown in Results A of Table 6 (Rosenthal, 2000), which indicates an even smaller proportion of patients at risk; the odds ratio is still 10, which is an even more alarmist result. The odds ratio for Study 3 in Table 5 is also unattractive; but because all the controls die, perhaps we could exonerate the infinite odds ratio. However, very different phenomenological results yield an identical odds ratio. If the data resembled Results B of Table 6 (Rosenthal, 2000), we would again have an infinite odds ratio, definitely an alarmist result. In this case, even the problematic relative risk index would yield a phenomenologically more realistic result of 1.00.

The Risk Difference

With cells again labeled as shown in Table 5, the risk difference (RD) is defined as

$$RD = \left(\frac{A}{A+B} - \frac{C}{C+D} \right), \qquad (21)$$

the difference between the proportion of the control patients at risk and the proportion of the treated patients at risk. Applied to the Salk vaccine results (Table 2), the risk difference is (115/201,229) - (33/200,745) = .0004, or somewhat smaller than the effect size index r_ϕ previously calculated to be .011.

TABLE 6
Further Illustrations of Extreme Outcomes

Results A			
	Die	Live	Totals
Control	10	999,990	10^6
Treated	1	999,999	10^6
Totals	11	1,999,989	$2(10^6)$
Results B			
	Die	Live	Totals
Control	1,000,000	0	10^6
Treated	999,999	1	10^6
Totals	1,999,999	1	$2(10^6)$

Note. In Results A, the risk ratio (RR) = 10.00, the odds ratio (OR) = 10.00, the risk difference (RD) = .000009, $\chi^2_{(1)} = 7.36$, and r_ϕ = .0019. In Results B, the RR = 1.00, OR = infinity, RD = .000001, $\chi^2_{(1)} = 1.00$, and $r_\phi = .00071$.

The last column of Table 5 shows the Pearson product-moment *r* between independent variable of treatment (scored 0, 1) and dependent variable of outcome (scored 0, 1). Comparing risk differences with *r* in Table 5 (and elsewhere) shows that RD is never unreasonably far from the value of *r*. For that reason, the RD index may be the one least likely to be quite misleading under special circumstances. Thus, if we had to choose among RR, OR, and RD, we would select RD as our all-purpose index among these three. But even here we feel we can do better.

Standardizing the Three Risk Indices

In other work, we have proposed a simple adjustment that standardizes the RR, OR, and RD indices (Rosenthal et al., 2000). We compute the *r* between treatment and outcome, and then display *r* in a BESD as described above. Table 7 shows these BESD-based results for the three studies of Table 5. The *N*s in the tables of counts of Table 5 varied considerably (2,000, to 40, to 20), but the corresponding BESD-based indices of Table 7 all show the standard margins of 100, which is a design feature of the BESD. The calculation of our new effect size indices is straightforward. We compute the relative risk, odds ratio, and risk difference on our BESD tables to obtain standardized (BESD-based) relative risk, odds ratio, and risk difference.

With cells of the 2 x 2 table labeled A, C, C, A from upper left to lower right (as shown in Table 7), the cal-

TABLE 7
Standardized Outcomes of Table 5

	Die	Live	BESD-based RR	BESD-based OR	BESD-based RD (r)
Control	A	C	(Eq. 22)	(Eq. 23)	(Eqs. 24 & 9)
Treatment	C	A			
Study 1					
	Die	Live			
Control	53	47	1.13	1.27	.06
Treatment	47	53			
Study 2					
	Die	Live			
Control	75	25	3.00	9.00	.50
Treatment	25	75			
Study 3					
	Die	Live			
Control	95	5	19.00	361.00	.90
Treatment	5	95			

culation is further simplified. The BESD standardized risk ratio is calculated as

$$\text{BESD-based RR} = \left(A / C\right). \tag{22}$$

To apply Equation 22 to the Salk vaccine BESD, where the values of cells A and C are 50.5 and 49.5, respectively, the BESD-based RR = 50.5/49.5 = 1.02. The odds ratio standardized is calculated as

$$\text{BESD-based OR} = \left(A / C\right)^2, \tag{23}$$

which, applied to the Salk vaccine BESD, yields BESD-based OR = $(50.5/49.5)^2$ = 1.04 Finally, the standardized risk difference, which is now actually equivalent to r_ϕ, is calculated as

$$\text{BESD-based RD} = \frac{A - C}{100}, \tag{24}$$

and applied to the Salk vaccine BESD, yields BESD-based RD = (50.5 - 49.5)/100 = .01.

Table 7 compares these standardized indices using the outcomes of Table 5. We see the BESD-based RR in Table 7 increasing, as it should, in going from Study 1 to Study 3. The BESD-based OR in Table 7 also increases from Study 1 to Study 3, but without the alarmist value for Study 1 and the infinite value for Study 3. (A standardized odds ratio could go to infinity only if r_ϕ were exactly 1.00, an unlikely event in behavioural or biomedical research.) The BESD-based RD is shown in Table 7 to be identical to the effect size index r_ϕ, which is an attractive feature emphasizing the interpretability of r-type indices as exhibited in a BESD.

TABLE 8
Number of Tracking Errors in Four Independent Groups, With Subjects in Each Group Exposed to a Particular Level of Complexity of Background Noise

Level of complexity of background noise				
	Low	Moderate	High	Extreme
	0	1	5	7
	1	2	4	8
	2	3	6	6
	1	3	6	7
	2	2	5	8
M =	1.20	2.20	5.20	7.20
S^2 =	.70	.70	.70	.70

The Multiple-Sample Case

We turn now to designs with more than two independent groups. Suppose an experimenter who is interested in object-tracking hypothesizes that the error rate will increase linearly as a function of the complexity of background noise. Based on the four-group experimental design and raw scores in Table 8, the experimenter reports only that the overall $F(3,16)$ = 54.17, with p considerably smaller than .05. The problem is that the experimenter specifically predicted a linear increment in error rate progressing from low to extreme noise condition, but the omnibus F is oblivious to that prediction. The same F will result if the ordering of the four groups were reversed, or indeed for any other arrangement of these four groups. To address the experimenter's prediction, we need to compute a linear contrast.

The contrast weights (λs) we select can take on any convenient numerical value as long as $\Sigma\lambda = 0$. We will choose single-digit λs of -3, -1, +1, +3, and substitute in the following equation:

$$t = \frac{\Sigma(M\lambda)}{\sqrt{\left(\sum \frac{\lambda^2}{n}\right)S^2}}, \tag{25}$$

where M = group mean, S^2 = the usual MS_{error} from a between-subjects ANOVA, n = number of units (e.g., subjects) in group, and λ = contrast weight. We find

$$t_{(16)} = \frac{(1.2)(-3)+(2.2)(-1)+(5.2)(+1)+(7.2)(+3)}{\sqrt{\left[\frac{(-3)^2}{5}+\frac{(-1)^2}{5}+\frac{(+1)^2}{5}+\frac{(+3)^2}{5}\right]0.70}} = \frac{21.00}{\sqrt{2.80}} = 12.55,$$

and one-tailed $p = 5.4^{-10}$. Squaring the contrast t gives us $F(1,16)$ = 157.50, $p = 1.1^{-9}$. We now have four com-

plementary ways of thinking about *r*-type effect sizes: the contrast *r*, alerting *r*, effect size *r*, and BESD *r* (Rosenthal et al., 2000).

The Contrast r

The contrast *r* is a partial correlation between individual sampling unit scores on the dependent variable (*Y*) and the predicted mean score (represented by λ, the contrast weight) of the group to which they belong, with other between-group variation (i.e., noncontrast variation, *NC)* removed. The contrast *r* can thus be written as $r_{contrast}$ or $r_{Y\lambda \bullet NC}$. It is the simplest effect size to calculate, and when all we have are minimal ingredients (as in meta-analytic work), it is sometimes the only one that can be estimated from the published results. For example, the contrast *r* can be conveniently estimated from focused tests of significance (i.e., *F* with numerator *df* = 1, all *t* tests, 1-*df* χ^2, and all *z* tests) by any of the following equations:

$$r_{contrast} = \sqrt{\frac{F_{contrast}}{F_{contrast} + df_{within}}}, \quad (26)$$

$$r_{contrast} = \sqrt{\frac{t^2_{contrast}}{t^2_{contrast} + df_{within}}}, \quad (27)$$

$$r_{contrast} = \sqrt{\frac{\chi^2_{(1)}}{N}}, \quad (28)$$

$$r_{contrast} = \frac{z}{\sqrt{N}}, \quad (29)$$

and in the two-sample case, the equal-*n* *r* can be obtained from the following:

$$r_{contrast} = \sqrt{\frac{d^2}{d^2 + 4}}. \quad (30)$$

Notice that Equation 27 is similar to Equation 12, and Equation 28 to Equation 9. The reason is that Equations 12 and 9 are actually both contrast correlations. But because there can be no noncontrast variation in the two-sample case, $r_{contrast} = r_{effect\ size}$ in two-group designs. For the example that we have been discussing, applying Equation 26 or 27 yields $r_{contrast}$ = .95.

The Alerting r

The alerting *r* is the Pearson product-moment correlation between group (or condition) means and contrast weights, and thus can be written as $r_{alerting}$ or $r_{M\lambda}$. We call it the alerting *r* because it can signal overall trends of interest and when squared, alerts us to the proportion of between-condition sum of squares accounted for by the particular contrast. If the leftover (noncontrast) between-condition variability is minimal, it tells us the contrast *r* is a close approximation of the effect size *r*. Using a handheld calculator that computes *r*s, it is easy to obtain the alerting *r*. In this case, with linear lambda weights of -3, -1, +1, +3 and group means of 1.2, 2.2, 5.2, 7.2, the correlation between λs and means is .9845 (and the squared alerting *r* is .969).

Table 9
Summary ANOVA for Results in Table 8, Including the Linear Contrast

Source	*SS*	*df*	*MS*	*F*	*p*
Between	113.75	(3)	37.92	54.17	1.3^{-8}
Contrast	110.25	1	110.25	157.50	1.1^{-9}
Noncontrast	3.50	2	1.75	2.50	.11
Within error	11.20	16	.70		

Note. For the linear contrast *F*, the $r_{effect\ size}$ = .94, $r_{contrast}$ = .95, $r_{alerting}$ = .98, and r_{BESD} = .88. As the overall *F* and noncontrast *F* are omnibus tests (i.e., numerator *df* > 1), effect sizes would not be reported for them.

Table 9 shows the ANOVA on the data of Table 8, including the linear contrast *F* of 157.50. As we expected, the contrast *SS* (110.25) consumes 97% of the between-condition *SS* (113.75). Suppose all we knew were the group means and the omnibus *F* – which is not unusual in meta-analytic work. We can compute a contrast *F* in three easy steps. First, we correlate the contrast weights (λs) and the group means, and then square this value. Second, we multiply the omnibus *F* by its numerator *df*, which gives the maximum possible contrast *F*. Finally, we multiply the results of steps 1 and 2: (.969)(54.17)(3) = contrast *F*(1,16) = 157.5.

We can also use the following equation to obtain the alerting *r*:

$$r_{alerting} = \sqrt{\frac{F_{contrast}}{F_{contrast} + F_{noncontrast}\left(df_{noncontrast}\right)}}, \quad (31)$$

with terms defined in Table 9. Applying Equation 31, we find

$$r_{alerting} = \sqrt{\frac{157.50}{157.50 + 2.50(2)}} = .984.$$

The Effect Size r

The effect size *r* (written as $r_{effect\ size}$ or $r_{Y\lambda}$) is the simple (unpartialed) correlation between the contrast weights associated with membership in a group or con-

dition and scores on the dependent variable. As it involves no partialing of other between-condition effects out of the error term, the $r_{\text{effect size}}$ is never larger than the r_{contrast}, and is usually smaller than r_{contrast} (sometimes dramatically so). The $r_{\text{effect size}}$ can be computed from

$$r_{\text{effect size}} = \sqrt{\frac{F_{\text{contrast}}}{F_{\text{contrast}} + F_{\text{noncontrast}}\left(df_{\text{noncontrast}}\right) + df_{\text{within}}}}. \quad (32)$$

Applied to the results in Table 9, we find

$$r_{\text{effect size}} = \sqrt{\frac{157.50}{157.50 + 2.50(2) + 16}} = \sqrt{\frac{157.50}{178.50}} = .939.$$

Just as we expected from the squared alerting *r*, the r_{contrast} was a good approximation of the $r_{\text{effect size}}$ in this example.

The BESD r

The binomial effect size *r* tells us the $r_{\text{effect size}}$ that we would expect to see in a two-group replication with the same total *N*, and the lower-scoring group set at $-1\sigma_\lambda$ and the upper-scoring group at $+1\sigma_\lambda$, where

$$\sigma_\lambda = \sqrt{\frac{\Sigma\lambda^2}{k}}, \quad (33)$$

and *k* = number of conditions in the contrast. Space limitations prevent us from describing this estimator in detail, but further discussions are available elsewhere (Rosenthal et al., 2000; Rosnow et al., 2000). The r_{BESD} can be computed from

$$r_{\text{BESD}} = \sqrt{\frac{F_{\text{contrast}}}{F_{\text{contrast}} + F_{\text{noncontrast}}\left(df_{\text{noncontrast}} + df_{\text{within}}\right)}}, \quad (34)$$

with the restriction that if $F_{\text{noncontrast}}$ is less than 1.00, it is entered in Equation 34 as 1.00. The purpose of this restriction is to formalize the assumption that the noncontrast variation is noise, and forces r_{BESD} to be less than (or at most equal to) $r_{\text{effect size}}$. We can also compute $F_{\text{noncontrast}}$ from

$$F_{\text{noncontrast}} = \frac{F_{\text{between}}\left(df_{\text{between}}\right) - F_{\text{contrast}}}{df_{\text{between}} - 1}. \quad (35)$$

Applying Equation 35, we find

$$F_{\text{noncontrast}} = \frac{54.17(3) - 157.50}{3 - 1} = 2.50,$$

and substituting in Equation 34:

$$r_{\text{BESD}} = \sqrt{\frac{157.50}{157.50 + 2.50\left(2 + 16\right)}} = .882.$$

In the two-group case, r_{contrast}, $r_{\text{effect size}}$, and r_{BESD} will be identical. With $k > 2$, it is possible for r_{contrast}, r_{alerting}, $r_{\text{effect size}}$, and r_{BESD} to be identical, but typically $r_{\text{effect size}}$ will be larger than r_{BESD}, and r_{contrast} will be larger than $r_{\text{effect size}}$. For other equivalences among effect size estimates, see Rosenthal (1991, 1994) and Rosenthal and Rosnow (1991).

TABLE 10
Four Repeated Measures with Associated Contrast (*L*) Scores for Three Age Groups

	Sessions				Contrast (*L*)
	1	2	3	4	scores[a]
Age 8					
Child 1	3	2	3	3	1
Child 2	1	2	1	2	2
Child 3	4	5	5	5	3
M =	2.67	3	3	3.33	2
S^2 =					1.00
Age 10					
Child 4	4	5	4	6	5
Child 5	5	6	5	6	2
Child 6	5	7	6	7	5
M =	4.67	6	5	6.33	4
S^2 =					3.00
Age 12					
Child 7	6	6	7	8	7
Child 8	5	6	6	8	9
Child 9	7	8	8	9	6
M =	6	6.67	7	8.33	7.33
S^2 =					2.33
Grand *M* =	4.44	5.22	5	6	4.44
Mean S^2 =					2.11

[a]These are the *L* scores, which for each child is a linear trend *L* score, as the experimenter had predicted that children's performance would show a linear improvement with practice in going from the first to the fourth session (contrast weights of -3, -1, +1, +3). To illustrate, the linear trend *L* score for Child 4 is computed as $L = \Sigma(Y_i\lambda_I) = (4)(-3) + (5)(-1) + (4)(+1) + (6)(+3) = 5$.

The Case of Multiple Repeated Measures

We turn next to another common design in experimental psychology, a factorial arrangement in which one or more factors involve repeated measures and the primary prediction of interest involves those repeated measures.[1] Table 10 illustrates this case by showing the raw scores of *N* = 9 children at three age levels who

TABLE 11
Preliminary Analysis of Variance of the Data of Table 10

Source	*SS*	*df*	*MS*	*F*	*p*
Between	130.50	(8)			
Age	98.00	2	49.00	9.04	.015
Children nested in age	32.50	6	5.42		
Within	18.50	(27)			
Sessions	11.22	3	3.74	19.24	7.6^{-6}
Sessions x age	3.78	6	0.63	3.24	.024
Sessions x children	3.50	18	0.19		
Total	149.00				

Note. Because all of these *F*s are omnibus tests (i.e., numerator *df* > 1), rather than focused tests (or contrasts), effect size estimates are not reported in this table.

were each measured on a 1-9 scale of performance accuracy on four equally spaced occasions. The experimenter's hypothesis was that the children would improve in a linear fashion from the first to the fourth session and that this linear trend would appear more clearly as the children's ages increased from 8 to 10 to 12. The experimenter is aware that another researcher implied somewhat different predictions, not a general linear increase. Instead, that other researcher's prediction was that children would require two sessions before a noticeable improvement would occur and that this pattern would be increasingly evident in older rather than younger children.

Table 11 shows the first experimenter's overall ANOVA and omnibus *F*s on the results of Table 10. This analysis tells us "where the action is" in terms of the relative magnitude of the mean squares; it gives omnibus *F*s and *p* values for the age effect, the sessions effect, and the Sessions x Age interaction. However, it tells us nothing of any specific interest, as it neither addresses this nor the other researcher's predictions. All the main effect for age tells us is that "ages matter," but not *how* they matter. The sessions effect tells us that "sessions matter," but not *how* they matter. The Sessions x Age interaction effect tells us that "session effects vary with age," but not *how* they vary.

Once again, we need to structure the focused questions in the form of contrasts that will yield interpretable answers with respect to both effect sizes and significance levels. When employing contrasts with repeated measures, it is necessary to compute for each sampling unit (each child in this example) an *L* score (or contrast score) that indexes for each unit the degree to which it reflects the accuracy of the particular prediction (Rosenthal & Rosnow, 1985; Rosenthal et al. 2000). The final column of Table 10 shows the *L* scores for each child, defined as

TABLE 12
Contrast Scores for Linear Trend (-3, -1, +1, +3) Over Sessions for Children of Three Age Groups

	Age		
	8	10	12
	1	5	7
	2	2	9
	3	5	6
M	2.00	4.00	7.33
S^2	1.00	3.00	2.33

$$L = \Sigma\left(Y_i \lambda_i\right) = Y_1 \lambda_1 + Y_2 \lambda_2 + \ldots + Y_k \lambda_k . \qquad (36)$$

The experimenter's first prediction was that, on the whole, children would show a linear improvement over sessions. We can represent this prediction by contrast weights (λs) of -3, -1, +1, +3. All we need do is compute a one-sample *t*-test on the grand mean *L* score:

$$t_{(df)} = \frac{\bar{L}}{\sqrt{\frac{1}{N}\left(S_L^2\right)}}, \qquad (37)$$

where $\bar{L}$ is the mean of the *L* scores, *N* is the number of children in the study, and S_L^2 is the variance of the *L* scores collected within each of the three age groups and aggregated. For the data of Table 10, we find

$$t_{(6)} = \frac{4.44}{\sqrt{\frac{1}{9}(2.11)}} = 9.17,$$

and $p = .000047$. The effect size $r_{contrast}$ estimate is computed from Equation 27 as

$$r_{contrast} = \sqrt{\frac{(9.17)^2}{(9.17)^2 + 6}} = .97.$$

The experimenter's more interesting question was the degree to which increasing age of the children would be associated with increasing linearity of improvement over the course of the four sessions. Table 12 shows the *L* scores obtained by children of

[1] If there were no repeated measures and we were primarily interested in the predicted pattern of condition means, the procedure described in the preceding section could be used. For example, if the design were a 2 x 2 ANOVA, we could address the overall prediction by means of a 1 x 4 contrast and correlational indices along the lines illustrated in the previous example (Rosnow & Rosenthal, 1995).

TABLE 13
Contrast Scores for Alternative Prediction (-1, -1, +1, +1) for Improvement Over Sessions

	Age		
	8	10	12
	1	1	3
	0	0	3
	1	1	2
M	0.67	0.67	2.67
S^2	0.33	0.33	0.33

the three age groups. Applying Equation 25, with S^2 estimated by aggregating the S^2 values in Table 12 (i.e., $S^2_{\text{aggregated}} = 2.11$), we find

$$t_{(6)} = \frac{2(-1)+4(0)+7.33(+1)}{\sqrt{\left[\frac{(-1)^2}{3}+\frac{(0)^2}{3}+\frac{(+1)^2}{3}\right]2.11}} = \frac{5.33}{1.19} = 4.50,$$

$p = .0021$, $r_{\text{contrast}} = .88$, $r_{\text{alerting}} = .99$, and $r_{\text{effect size}} = .87$.

The other researcher's general prediction was that children would require two sessions before a noticeable improvement would occur. To address this prediction, we choose contrast weights of -1, -1, +1, +1. Correlating these weights with the -3, -1, +1, +3 linear weights, we find $r = .89$, which tells us the predictions, although different, are highly correlated. Table 13 shows the *L* scores for each of the nine children based on the prediction with weights of -1, -1, +1, +1. Applying Equation 37, we find

$$t_{(6)} = \frac{1.33}{\sqrt{\frac{1}{9}(.33)}} = 6.95,$$

and $p = .00022$. From Equation 29, the effect size r_{contrast} estimate is .94.

A more interesting question was whether this pattern would be shown increasingly more by older than by younger children. Applying Equation 25, we find

$$t_{(6)} = \frac{0.67(-1)+0.67(0)+2.67(+1)}{\sqrt{\left[\frac{(-1)^2}{3}+\frac{(0)^2}{3}+\frac{(+1)^2}{3}\right]0.33}} = \frac{2}{0.47} = 4.26,$$

and $p = .0027$, $r_{\text{contrast}} = .87$, $r_{\text{alerting}} = .87$, and $r_{\text{effect size}} = .77$. In sum, the prediction of Table 13 (with contrast weights of -1, -1, +1, +1) was supported almost as well as was the prediction of Table 12 (with contrast weights of -3, -1, +1, +3). If we wanted a more direct and more precise comparison of these two hypotheses, we could do so by procedures described elsewhere (Rosenthal et al., 2000, pp. 165-169; Rosnow & Rosenthal, 2002b).

Minimizing Errors in Thinking About Effect Sizes

At the beginning of this article, we alluded to discussions about null hypothesis significance testing (NHST) and its discontents. We will conclude by mentioning a recently introduced statistic, the counternull value of the effect size (Rosenthal & Rubin, 1994; Rosnow & Rosenthal, 1996), which may help to eliminate two common errors associated with NHST. The first error occurs when the researcher mistakenly infers that failure to reject the null implies an effect size of zero; the second error occurs when the researcher mistakenly equates rejection of a null hypothesis on the basis of a significance test with having demonstrated a scientifically important effect. The counternull value of the effect size refers to the nonnull magnitude of the effect size that is supported by exactly the same amount of evidence as is the null value of the effect size. That is, if the counternull value were taken as the null hypothesis, the resulting *p* value would be the same as the obtained *p* value for the actual null hypothesis.

For effect size estimates that are based on symmetric distributions (e.g., *d*, *g*, Δ, z_r), no matter what the magnitude of the effect size (*ES*) is under the null, the counternull value is

$$ES_{\text{counternull}} = 2\left(ES_{\text{obtained}}\right) - ES_{\text{null}}. \qquad (38)$$

Because the effect size expected under the null is zero in many of its applications, the value of the counternull is often simply twice the obtained effect size, or $2(ES_{\text{obtained}})$. For asymmetric distributions (e.g., the *r* between two continuous variables), it is best to transform the effect size to a symmetric distribution, then to calculate the counternull on the symmetric scale, and finally to transform back to obtain the counternull on the original scale (Rosenthal et al., 2000). The following equation can be used to estimate the counternull statistic of the obtained *r*:

$$r_{\text{counternull}} = \sqrt{\frac{4r^2}{1+3r^2}}, \qquad (39)$$

where *r* is simply the estimated magnitude of the obtained effect.

Suppose an experimenter calculated an obtained effect size *r* of .10, with null defined as $r = .00$, and (using Equation 7 or looking in a table of significance levels) found $p = .20$. Applying Equation 39, the

experimenter finds $r_{counternull}$ = .20 (rounded), which tells us that the counternull value of r = .20 is "as likely" as the null value of r = .00. Rather than concluding that "nothing happened" because the obtained p was greater than the chosen level of .05, the experimenter instead accepts the conclusion that an effect size of r = .20 is just as tenable as an effect size of zero. In the same way, the counternull value of $2d$ or $2z_r$ would be just as defensible a conclusion as concluding d = 0 or z_r = 0.

The counternull value is conceptually related to confidence intervals (which provide limits for such fixed probabilities as, for example, 95% and 99%) but involves the null hypothesis and the obtained p value. As Cohen, with his customary wisdom, pointed out, the behavioural and medical sciences would be far more advanced had researchers routinely reported not just p values, but effect size estimates with confidence intervals as well (Cohen, 1990, 1994).

Portions of this article draw on some of our earlier writing in Rosenthal (2000); Rosenthal and Rosnow (1991); Rosenthal, Rosnow, and Rubin (2000); Rosnow and Rosenthal (1989, 1996); and Rosnow, Rosenthal, and Rubin (2000). Eric K. Foster (2003) has developed free software for Microsoft Windows and the HP49G programmable calculator for many of the procedures described in this article and in our earlier writing noted above (http://www.netaxs.com/~efoster). Correspondence concerning this article may be addressed to Ralph L. Rosnow, 177 Biddulph Road, Radnor, PA 19087 USA (E-mail: rosnow@temple.edu).

References

Alexander, R. A., Scozzaro, M. J., & Borodkin, L. J. (1989). Statistical and empirical examination of the chi-square test for homogeneity of correlations in meta-analysis. *Psychological Bulletin, 106,* 329-331.

Alpha-Tocopherol, Beta Carotene Cancer Prevention Study Group (1994). The effect of vitamin E and beta carotene on the incidence of lung cancer and other cancers in male smokers. *New England Journal of Medicine, 330,* 1029-1035.

Altman, L. K. (1994, February 21). In major finding, drug limits H.I.V. infection in newborns. *The New York Times,* pp. A1, A13.

American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: Author.

Barnes, D. M. (1986). Promising results halt trial of anti-AIDS drug. *Science, 234,* 15-16.

Brownlee, K. A. (1955). Statistics of the 1954 Polio vaccine trials. [Electronic version]. *Journal of the American Statistical Association, 272,* 1005-1013.

Canadian Multicentre Transplant Study Group. (1983). A randomized clinical trial of cyclosporine in cadaveric renal transplantation. *New England Journal of Medicine, 309,* 809-815.

Centers for Disease Control Vietnam Experience Study. (1988). Health status of Vietnam veterans: 1. Psychosocial characteristics. *Journal of the American Medical Association, 259,* 2701-2707.

Chandler, D. L. (1993, February 15). Study finds evidence of ESP phenomenon. *Boston Globe,* pp. 1, 8.

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. (1990). Things I have learned (so far). *American Psychologist, 45,* 1304-1312.

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

Contraceptive trials set for a link to AIDS research. (1996, May 7). *Boston Globe*, p. B1.

Cooper, H. M. (1981). On the significance of effects and the effects of significance. *Journal of Personality and Social Psychology, 41,* 1013-1018.

Cromie, W. J. (1990, October 5). Report: Drugs affect lung cancer survival. *Harvard Gazette,* 1, 10.

Cromie, W. J. (1991, September 13). Study: Hospitalization recommended for problem drinkers. *Harvard Gazette,* 3-4.

Dabbs, J. M., Jr., & Morris, R. (1990). Testosterone, social class, and antisocial behavior in a sample of 4,462 men. *Psychological Science, 1,* 209-211.

Fleiss, J. L. (1994). Measures of effect size for categorical data. In H. Cooper & L. V. Hedges (Eds.), *Handbook of research synthesis* (pp. 245-260). New York: Russell Sage Foundation.

Foreman, J. (1995, July 27). Medical notebook: A new confirmation for a pregnancy drug. *Boston Globe,* p. B3.

Foster, E. K. (2003). METASTATS: Behavioral science statistics for Microsoft Windows and the HP49G programmable calculator. *Behavior Research Methods, Instruments, & Computers, 35,* 325-328.

Francis, T., Jr., Korns, R. F., Voight, R. B., Boisen, M., Hemphill, F., Napier, J., & Tolchinsky, E. (1955). An evaluation of the 1954 poliomyelitis vaccine trials – summary report. *American Journal of Public Health, 45*(5), 1-63.

GISSI: Gruppo Italiano per lo Studio della Streptochinasi Nell'Infarto Miocardico. (1986, February 22). *Lancet,* 397-402.

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

Goldfinger, S. E. (1991, August). Garlic: Good for what ails you. *Harvard Health Letter, 16*(10), 1-2.

Grady, D. (2003, February 25). Safe therapy is found for high blood-clot risk. *The New York Times,* pp. A1, A22.

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

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings*. Newbury Park, CA: Sage.

Knox, R. A. (1997, February 25). AIDS trial terminated: 3-drug therapy hailed. *Boston Globe,* pp. A1, A16.

Kolata, G. B. (1981). Drug found to help heart attack survivors. *Science, 214,* 774-775.

Meier, P. (1988). The biggest public health experiment ever: The 1954 field trial of the Salk poliomyelitis vaccine. In J. M. Tanur, F. Mosteller, W. H. Kruskal, E. L. Lehmann, R. F. Link, R. S. Pieters, & G. R. Rising (Eds.), *Statistics: A guide to the unknown* (3rd ed., pp. 3-14). Pacific Grove, CA: Wadsworth.

Roberts, L. (1987). Study bolsters case against cholesterol. *Science, 237,* 28-29.

Rosenthal, R. (1991). *Meta-analytic procedures for social research* (Rev. ed.). Newbury Park, CA: Sage.

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

Rosenthal, R. (1995). Progress in clinical psychology: Is there any? *Clinical Psychology: Science and Practice, 2,* 133-150.

Rosenthal, R. (2000). Effect sizes in behavioral and biomedical research: Estimation and interpretation. In L. Bickman (Ed.), *Validity and social experimentation: Donald Campbell's legacy* (Vol. 1, pp. 121-139). Newbury Park, CA: Sage.

Rosenthal, R., & Rosnow, R. L. (1985). *Contrast analysis: Focused comparisons in the analysis of variance*. Cambridge, UK: Cambridge University Press.

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*. Cambridge, UK: 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. (1989). Effect size estimation for one-sample multiple-choice type data: Design, analysis, and meta-analysis. *Psychological Bulletin, 106,* 332-337.

Rosenthal, R., & Rubin, D. B. (1991). Further issues in effect size estimation for one-sample multiple-choice-type data. *Psychological Bulletin, 109,* 351-352.

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

Rosenthal, R., & Rubin, D. B. (in press). $r_{equivalent}$: A simple effect size indicator. *Psychological Methods*.

Rosnow, R. L. (1991). Inside rumor: A personal journey. *American Psychologist, 46,* 484-496.

Rosnow, R. L., & Georgoudi, M. (Eds.). (1986). *Contextualism and understanding in behavioral science: Implications for research and theory*. New York: Praeger.

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.

Rosnow, R. L., & Rosenthal, R. (2002a). *Beginning behavioral research: A conceptual primer* (4th ed.). Upper Saddle River, NJ: Prentice Hall.

Rosnow, R. L., & Rosenthal, R. (2002b). Contrasts and effect sizes in theory assessment. *Journal of Pediatric Psychology, 27,* 59-66.

Rosnow, R. L., Rosenthal, R., & Rubin, D. B. (2000). Contrasts and correlations in effect-size estimation. *Psychological Science, 11,* 446-453.

Schaffer, J. P. (1991). Comment on "Effect size estimation for one-sample multiple-choice-type design: Design, analysis, and meta-analysis" by Rosenthal and Rubin (1989). *Psychological Bulletin, 109,* 348-350.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.

Smith, M. L., Glass, G. V., & Miller, T. I. (1980). *The benefits of psychotherapy*. Baltimore, MD,: John Hopkins University Press.

Steering Committee of the Physicians Health Study Research Group. (1988). Preliminary report: Findings from the aspirin component of the ongoing physicians' health study. *New England Journal of Medicine, 318,* 261-264.

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

Sommaire

Le présent article décrit trois familles d'estimateurs d'importance de l'effet et leur utilisation dans des situations d'intérêt général et particulier pour les psychologues en recherche. Les cas décrits sont notamment les suivants : les conceptions (mesures répétées) entre groupes et intrinsèques au groupe, les données continues et catégoriques et une importance d'effet qui permet de comparer l'importance des effets. Nous illustrons aussi des cas précis du rapport général entre la valeur *p* et l'importance de l'effet qui est obtenu par le test de signification = importance de l'effet x envergure de l'étude. Même si l'accent est mis sur la corrélation (type *r*) de la famille d'importance de l'effet (p. ex., le point bisérial *r* quand une variable est continue et qu'une variable est dichotomique; le coefficient phi quand les deux variables sont dichotomiques; les autres variantes comme le contraste, l'avertissement, l'importance de l'effet, le BESD et le contre-nul *r*; et la transformation *z* de *r*) de Fisher, nous abordons le type de différence (p. ex., *d, g, h* et *q* de Cohen; le *g* de Hedges; le probit et le logit *d'*; π; et la différence du risque) ainsi que les estimateurs de l'importance de l'effet du type proportionnel (p. ex., le risque relatif et les odds-ratio).

Malgré la sensibilisation grandissante à l'importance de rapporter l'importance des effets et leurs intervalles de confiance, il reste un problème quant à la façon d'interpréter les estimateurs populaires d'importance de l'effet. Ainsi, nous présentons une perspective tirée du monde réel de ce problème en décrivant la « plus grande expérience de tous les temps en santé publique » (soit l'essai du vaccin Salt en 1954) et en plaçant l'importance de l'effet en contexte en résumant l'importance des effets dans vingt-deux autres études. Nous illustrons aussi les limites de trois estimateurs populaires d'importance de l'effet utilisés en recherche biomédicale (soit, le risque relatif, l'odds-ratio et la différence du risque) et nous montrons comment en standardisant ces indices de risque par la méthode d'ajustement de la présentation binominale de l'importance de l'effet *binomial effect-size display* (BESD) qui produira une différence du risque à BESD qui est identique à l'indice d'importance de l'effet phi, ce qui par conséquent met l'accent sur la possibilité d'interpréter les indices de type *r* comme montré dans une BESD. La BESD montre l'effet de type *r* comme une simple différence des taux du résultat entre les groupes expérimentaux et témoins dans un tableau 2 x 2 avec des rangées et des colonnes qui totalisent toujours 100.

Lorsque des contrastes 1 x *k* sont utilisés dans des conceptions avec des conditions $k > 2$ conditions, les estimateurs de types *r* recommandés sont entre autres les suivants : a) le contraste *r* (la corrélation partielle entre les scores de la variable dépendante et le score moyen prévu avec d'autres variantes entre groupes enlevées); b) l'avertissement *r* (la corrélation simple entre les conditions moyennes et les pondérations de contraste; c) l'importance de l'effet *r* (la corrélation simple entre les scores de la variante dépendante et les pondérations de contraste); et d) la BESD *r* (l'importance de l'effet auquel on pourrait s'attendre dans une réplication de deux groupes, compte tenu de certaines spécifications décrites dans le présent article). L'article conclut avec une brève description de la valeur contre-nulle de l'importance de l'effet qui renvoie à l'ampleur du non nul appuyée sur la même quantité de preuves que pour la valeur nulle de l'importance de l'effet. La valeur contre-nulle qui est liée d'un point de vue conceptuel aux intervalles de confiance (mais qui est fondée sur la valeur *p* obtenue plutôt que sur une probabilité fixe) est particulièrement utile lorsque la valeur *p* obtenue est plus grande que 0,05 (c.-à-d. non « signifiante ») mais l'importance de l'effet est une certaine valeur non nulle.