Psychological Methods
2004, Vol. 9, No. 2, 183–197

Copyright 2004 by the American Psychological Association
1082-989X/04/$12.00 DOI: 10.1037/1082-989X.9.2.183

Biases of Success Rate Differences Shown in Binomial Effect Size Displays

Louis M. Hsu
Fairleigh Dickinson University

The intent of a binomial effect size display (BESD) is to show "the [real-world] importance of [an] effect indexed by a correlation [*r*]" (R. Rosenthal, 1994, p. 242) by reexpressing this correlation as a success rate difference (SRD) (e.g., treatment group success rate − control group success rate). However, SRDs displayed in BESDs generally overestimate real-world SRDs implied by correlations of (a) dichotomous *X* and *Y* variables (ϕ coefficients), (b) dichotomous *X* and continuous *Y* variables (point-biserial coefficients [r_{pb}s]), and (c) continuous *X* and *Y* variables (r_{xy}s). Furthermore, overestimation biases are larger for r_{xy}s than for r_{pb}s. Differences in the sizes of biases linked to different correlations suggest that BESD SRDs reported for different correlations are not comparable. The stochastic difference index (N. Cliff, 1993; A. Vargha & H. D. Delaney, 2000) is recommended as an alternative to the BESD.

Rosenthal and Rubin (1982) proposed "an intuitively appealing general-purpose effect size display whose interpretation is perfectly transparent: the binomial effect size display (BESD)" (p. 166), which is intended to show "the *real-world importance* [italics added] of treatment effects" (Rosenthal, 1987, p. 117). More specifically, Rosenthal and Rubin stated that the BESD "displays the change in success rates (e.g., survival rate, cure rate, improvement rate, selection rate, etc.) attributable to a certain treatment procedure" (p. 166), and Rosenthal (1990) noted that this success rate difference (SRD) shows "the *practical importance* [italics added] of any effect indexed by a correlation coefficient" (p. 775). Emphasis on the BESD's relevance to real populations has been restated in some of Rosenthal's most recent writings: The BESD shows "the *real-world importance* [italics added] of any treatment effect" (Rosenthal, 2002, p. 844).[1]

The operational definition of a BESD SRD for "any obtained effect size *r*" (Rosenthal, 1994, p. 242) is seductively simple: "We obtain the BESD . . . by computing the treatment condition success rate as (.5 plus *r*/2) and the control condition success rate as (.5 minus *r*/2)" (Rosenthal, 1994, p. 242). The BESD SRD is then operationally defined as follows:

$$\text{BESD SRD} = [(.5 + r/2) - (.5 - r/2)] = r, \qquad (1)$$

where *r* is (see Rosenthal, 1994) any obtained Pearson product–moment correlation (e.g., r_{pb}, r_{xy}, or ϕ). Note that if *r* is calculated on population data, the operational definition of the BESD SRD would yield what might be called a BESD SRD effect size parameter.

Rosenthal and Rubin (1982) illustrated how the BESD SRD would be calculated in a 2 × 2 table showing the relation of an independent variable *X* (treatment, control) to a dependent variable *Y* (alive, dead). In this table, which had uniform marginal distributions (100 patients per category of *X* and 100 patients per category of *Y*) the proportions of treated and control patients who survived were, respectively, .66 and .34, and the *r* or ϕ was .32. Applying Equation 1, the BESD SRD = (.5 + .32/2) − (.5 − .32/2) = (.66 − .34) = .32: This BESD SRD was exactly equal to the actual SRD displayed in the table (viz., actual SRD = 66/100 − 34/100) = .32. What may not be apparent from Rosenthal and Rubin's illustration is that the equality of the BESD SRD (calculated from Equation 1) and actual SRD of a 2 × 2 table does not generalize to 2 × 2 tables that do not have uniform marginal distributions.

It can generally be shown (see, e.g., Strahan, 1991) that whenever a study of a real-world population yields a 2 × 2 table with uniform marginal distributions, the ϕ (or *r*) of this real-world population, and therefore its BESD SRD parameter (see Equation 1), will exactly equal its actual SRD parameter (calculated from frequencies reported in the table). None of the major critics of the BESD (viz., Crow,

Correspondence concerning this article should be addressed to Louis M. Hsu, Ph.D. Program in Clinical Psychology, Teaneck–Hackensack Campus, Fairleigh Dickinson University, 1000 River Road, T-WH1-01, Teaneck, NJ 07666. E-mail: lhsu@fdu.edu

[1] For an excellent discussion of a broad range of issues related to the meaning of practically and clinically important treatment effects, see Kazdin (2003, chapter 14).

1991; McGraw, 1991; Preece, 1983; Strahan, 1991; Thompson & Schumacker, 1997) have questioned, in the case of a study that yields a 2 × 2 table with uniform marginals, the validity of Rosenthal and Rubin's (1982) statement that the BESD SRD (operationally defined in Equation 1) is "perfectly transparent" and "easily understood" (p. 166) or their belief that the BESD SRD can provide practically important information about an effect in a real-world population: Then, the BESD SRD defined as BESD SRD $= r$ equals a real-world SRD.

However, Rosnow, Rosenthal, and Rubin (2000) did not suggest restricting the use of the BESD to empirically observed 2 × 2 tables with uniform marginals or limiting the use of the BESD to specific types of research designs (e.g., naturalistic, prospective, retrospective, random; see Carroll, 1961; Fleiss, Levin, & Paik, 2003). Quite the contrary, they noted that (a) BESDs provide useful information about real-world SRDs for any Pearson correlation and for any "effect size estimate that can be converted to [a Pearson *r*]" and (b) "*any* product–moment correlation [can be recast] into such a display, whether the original data is *continuous* or *categorical* [italics added]" (Rosnow et al., 2000, p. 451). Furthermore, Rosenthal (1995b) stated that although "the input to the BESD is a specific effect size estimate, the Pearson *r*, . . . any other effect size estimate can be converted to *r*" (p. 190). In addition, Rosenthal (1995b) stated that "the BESD can be used to display the mean or median effect size estimate of any meta-analysis" (p. 190); examples of this type of application include Rosenthal's (2002) meta-analyses of the literature on interpersonal expectancy effects and Hiller, Rosenthal, Bornstein, Berry, and Brunell-Neuleib's (1999) meta-analyses on relative validities of Minnesota Multiphasic Personality Inventory (MMPI) and Rorschach scales.

The BESD SRD is one of the most recent additions to the existing collection of more than 20 effect size indices examined by Huberty (2002). Despite its recent origin, the BESD has been used extensively: A recent search (August 26, 2003) of the American Psychological Association's full-text journal articles data base yielded 360 documents in response to *binomial effect size display.* One reason for this popularity is undoubtedly the observation (see Rosenthal, 1990, 1995a, 2002; Rosnow & Rosenthal, 1988) that BESDs have often suggested large and practically important effects when other effect size indices have not. In fact Rosenthal (1990) concluded that "the BESD has shown that we are doing considerably better . . . than we may have thought we were doing" (p. 777).

A second reason for the BESD's popularity is undoubtedly its perceived ability to provide a common metric for comparisons of magnitudes of effects across heterogeneous studies—for instance, studies differing in terms of objectives, outcome measures, effect size statistics, and so forth (e.g., Hiller et al., 1999).

Because the BESD SRD parameter, defined in Equation 1 as BESD SRD $= r$, has only been shown to be equal to the actual SRD parameter of a population when the reported *r* is for a 2 × 2 table with uniform marginal distributions (see Strahan, 1991), and because *r*s reported in studies are almost never *r*s (or ϕs) for 2 × 2 tables with uniform marginals, attempts to draw inferences about actual SRDs in real-world populations from BESD SRDs (defined as equal to reported *r*s) generally require "transform[ing]" (Rosnow & Rosenthal, 1996, p. 338) or "recast[ing]" (Rosnow & Rosenthal, 1988, p. 207) the reported data into 2 × 2 tables with uniform marginals. Serious consequences of this recasting generally include lack of realism of resulting tables and large overestimation biases of BESD SRDs relative to the actual real-world SRDs they are intended to estimate. However, nonnegligible overestimation biases of BESD SRDs may also be present when the recasting is realistic. Therefore the real-world meaning of BESD SRDs cannot (as I see it) be considered "perfectly transparent [or] easily understood" (Rosenthal & Rubin, 1982, p. 166).

The present article focuses on the measurement of biases of the SRDs displayed in BESDs: It shows that, as operationally defined, BESD SRD parameters are often larger (and sometimes much larger) than the real-world SRD parameters of interest to BESD users (and are therefore "biased" in that sense), and that factors (including types of correlations—e.g., r_{xy} vs. r_{pb}) that affect the relative sizes of the BESD SRD biases can invalidate comparisons of BESD SRDs. This article shows that biases of BESD SRDs can occur because of a violation of either of two important assumptions: (a) the uniform marginals assumption and (b) the constancy of correlations assumption. The exact real-world meanings of these assumptions differ for different types of correlations (ϕ, r_{pb}, r_{xy}) and are fully described under separate headings (see Cases 1–4). However, for any type of *r*, both of these assumptions must be tenable if BESD SRDs are to address, as intended, practically important real-world questions.

Authors who have endorsed the BESD (viz., Rosenthal, 1990, 1991, 1995a, 2002; Rosenthal & Rubin, 1982; Rosnow et al., 2000) have not ignored the BESD SRD bias problem but have suggested (see, e.g., Rosenthal, 1991; Rosenthal & Rosnow, 1991; Rosenthal & Rubin, 1982; Rosnow & Rosenthal, 1988) that biases are generally negligible, given uniform marginal distributions in the recast tables (see Case 2). However, these authors have paid little attention to biases that result when real-world marginal distributions are not uniform in the case of r_{xy}s, r_{pb}s and ϕs or when constancy of correlations is not present in the case of ϕs—that is, when the ϕ coefficient of the observed 2 × 2 table (regardless of marginal distributions) is not equal to the ϕ for real-world data that would yield a 2 × 2 table with uniform marginals (see Case 1 below).

In addition, neither authors who have recommended the BESD (e.g., Rosenthal, 1990, 1991, 1995a, 1995b, 2002; Rosenthal & Rubin, 1982; Rosnow et al., 2000) nor those who have criticized it (Crow, 1991; McGraw, 1991; Preece, 1983; Strahan, 1991; Thompson & Schumacker, 1997) have provided any quantitative information about biases of BESD SRDs that correspond to Pearson *r*s calculated on continuous (interval- or ratio-scaled) *X* and *Y* variables (r_{xy}s), even though (a) users have been encouraged to determine BESD SRDs for all types of Pearson *r*s (Rosenthal, 1990, 1994, 2002; Rosnow et al., 2000), (b) Pearson r_{xy}s calculated on continuous *X* and *Y* variables are very popular effect size statistics (see Cohen, 1988; Hunter, Schmidt, & Jackson, 1982; Meyer et al., 2001; Rosenthal, 1994), and (c) BESD SRDs have been calculated extensively for Pearson r_{xy}s applied to continuous *X* and *Y* variables in several important meta-analyses (e.g., Hiller et al., 1999; Meyer et al., 2001).

BESD SRD biases are discussed in this article in the context of four types (or cases) of *r*s that have been reported in published studies: Case 1, defined as ϕs for naturally dichotomous *X* and *Y* variables; Case 2, defined as r_{pb}s for naturally dichotomous *X* and continuous *Y* variables; Case 3, defined as r_{xy}s for continuous *X* and *Y* variables; and Case 4, defined as r_{pb}s for artificially dichotomized *X* and continuous *Y* variables.

Case 1: Naturally Dichotomous Independent (*X*) and Dependent (*Y*) Variables: Synthesis and Implications of Published Critiques of BESDs

Almost all of the published criticisms of the BESD (Crow, 1991; McGraw, 1991; Preece, 1983; Strahan, 1991; Thompson & Schumacker, 1997) have focused on the lack of real-world relevance of "data" generated by the transformation or recasting of the raw *X* and *Y* data into 2 × 2 tables with uniform marginals and on the consequent lack of relevance of the BESD SRDs to the real-world questions that are of interest to researchers (see, in particular, Crow, 1991; McGraw, 1991; Strahan, 1991). However, interpretations of these criticisms, as well as of biases of BESD SRDs discussed in the present article, depend on whether the concepts of (a) uniform marginals and (b) constancy of correlations are viewed as postulates (whose realism is not considered relevant) or as assumptions (whose realism is viewed as relevant).[2] Important differences in these interpretations are illustrated below with results of the Physicians' Aspirin Study (Rosenthal, 1995a, p. 135), a real-world study frequently cited in Rosenthal's writings about the BESD.

The aspirin study was a double-blind randomized design in which about half of 22,071 physicians received aspirin and the other half received a placebo. The proportion of all physicians who experienced heart attacks (denoted here as Π) was .0133; the ϕ for the study's 2 × 2 table (aspirin vs. placebo, heart attack vs. no heart attack) was .034; the proportions of heart attacks among those who received, and did not receive, aspirin were .0094 and .0171, respectively; and the observed SRD (difference in the heart attack rates of the two groups) for this real-world population was thus .0077 (about .008).

Curiously, the BESD SRD, which is intended (a) to be a measure of the magnitude of the effect of aspirin in a real-world population and, more specifically, (b) to "display the increase in success rate due to treatment" (Rosnow & Rosenthal, 1989, p. 1279) is not .008 but .034 (more than quadruple the SRD found in the real-world study). Unambiguous explanations of the difference between the reported SRD and the BESD SRD may be found in equations derived in Preece (1983; especially his results shown in Equation 2, below) and Thompson and Schumacker (1997) and in the work of Crow (1991), McGraw (1991), and Strahan (1991):

$$\text{Actual SRD} = 2\phi[\pi(1 - \pi)]^{.5}. \qquad (2)$$

Equation 2 shows the relation of the actual SRD to the success rate for the combined groups (π) and to the correlation (φ) in any 2 × 2 table that has a uniform marginal distribution for *X* (i.e., equal group size in the two groups). Equation 2 is implicit in the calculation of Case 1 BESD SRDs. Application of Equation 2 to the aspirin study data helps in understanding both the logic and the limitations of the BESD SRD for Case 1. Given that the ϕ coefficient does not change across total success rates (π) (i.e., given constancy of the correlation in Case 1), the observed ϕ of .034 in the aspirin study in which $\pi = .013$ and SRD = .008 implies (a) the relation SRD = $(2)(.034)[\pi(1 - \pi)]^{.5}$ (graphed in Figure 1) and, more specifically, (b) a BESD SRD = .034 (because the BESD SRD is the SRD for a uniform marginal distribution of the dichotomous *Y* (success, failure) "scores"—that is, $\pi = .5$, so that SRD = $2(.034)[.5(.5)]^{.5} = .034$). But how should this BESD SRD of .034 be interpreted? The two perspectives on the meaning of the BESD SRD, which differ in terms of whether uniform marginals and constancy of ϕ are viewed as postulates or as assumptions, can now be presented.

Interpreting Uniform Marginals and Constancy of ϕ as Postulates

One interpretation of BESD SRDs considers (a) constancy of the correlation ϕ as a theoretical premise or postulate. The BESD SRD parameter is then simply defined as the value of the SRD when (b) the total success rate is .50 (this may be viewed as a homogeneity of marginal *Y* dis-

[2] I thank an anonymous reviewer for drawing my attention to the "postulates" interpretation of the BESD SRD.

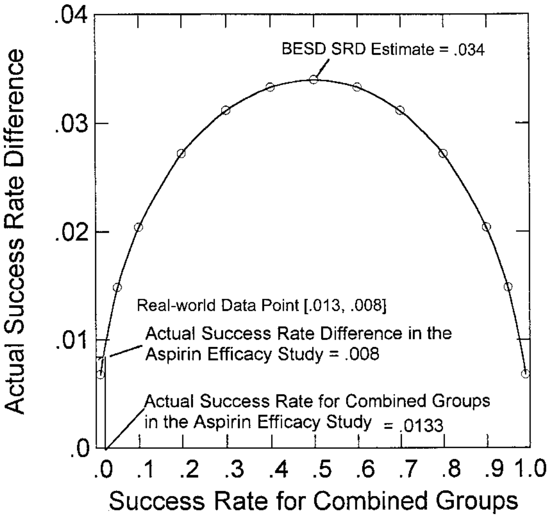

Figure 1. Implication of the binomial effect size display's (BESD's) constancy of ϕ assumption concerning the relation of the success rate difference (SRD) to the total success rate when $\phi = .034$.

tribution postulate). Given postulates (a) and (b), the BESD SRD parameter will always be equal (by definition) to the value of the observed ϕ coefficient, regardless of the observed value of the total success rate (π). Thus from this perspective, the fact that the aspirin study yielded $\phi = .034$ means that (regardless of the observed total success rate π) the BESD SRD parameter must by definition be .034. Questions about realism of (a) and (b) are considered moot; (a) and (b) are, in effect, simply components of the definition of the BESD SRD. The BESD SRD is not a parameter (or estimate of a parameter) in a real-world population but an effect size parameter defined in the context of a two-postulate theory.

This definition of the BESD SRD parameter may be considered analogous to definitions of the unrestricted range validity coefficient parameter (ρ_{xy}) when range restriction is present on variable *X*. One definition of this ρ_{xy} parameter (see Ghiselli, Campbell, & Zedeck, 1981, p. 299) is as follows:

$$\rho_{xy} = \frac{\rho'_{xy}(\sigma_x/\sigma'_x)}{[1 - \rho'^{\,2}_{xy} + \rho'^{\,2}_{xy}(\sigma^2_y/\sigma'^{\,2}_y)]^{.5}}, \tag{3}$$

where ρ_{xy} = the unrestricted range validity coefficient, ρ'_{xy} = the restricted range validity coefficient, σ_x = the full range standard deviation of *X* scores, σ'_x = the restricted range standard deviation of *X* scores, σ^2_y = the full range variance of *Y* scores, and $\sigma'^{\,2}_y$ = the restricted range variance of *Y* scores. Equation 3 is derived from a model that postulates (A) linearity of the regression of *Y* on *X* over the full range of *X* and (B) equality of standard errors of estimate for the restricted and unrestricted ranges (see Ghiselli et al., 1981). From a theoretical perspective (A) and (B) could simply be viewed as components of the definition of the unrestricted range parameter ρ_{xy}, just as from that perspective (a) and (b) could simply be viewed as components of the definition of the BESD SRD. From this perspective the parameters ρ_{xy} and BESD SRD are theoretically meaningful concepts irrespective of the realism of their underlying postulates.

Interpretations of the unrestricted range correlation parameter (ρ_{xy}) and of the BESD SRD effect size parameter, which consider (A) and (B) (in the case of ρ_{xy}) and (a) and (b) (in Case 1 applications of the BESD SRD) exclusively as postulates whose realism can be ignored, are certainly defensible perspectives on the meanings of these parameters. However, these perspectives are of theoretical rather than practical importance. This follows from the fact that the parameters (ρ_{xy} in Equation 3 and BESD SRD in Equation 2) will lack real-world relevance unless their postulates hold (which they may or may not) in real-world populations.

Interpreting Uniform Marginals and Constancy of ϕ as Assumptions About the Real World

The BESD SRD = .034 of the aspirin study indicates, according to Rosnow and Rosenthal (1989), "that approximately 3.4% fewer persons who would probably experience a myocardial infarction . . . will not experience it if they follow the regimen as prescribed in the aspirin treatment condition" (p. 1279). This view is consistent with the intent of showing in the BESD "the real-world importance of a treatment effect" (Rosenthal, 2002, p. 844) and of "display-[ing] the increase in success rate due to treatment" (Rosnow & Rosenthal, 1989, p. 1279).

However, it is suggested that six caveats—implied by Equation 2 and Figure 1 when (a) and (b) are viewed as assumptions rather than postulates—should be considered in relation to any attempt to interpret the BESD SRD of .034 as a measure of the magnitude of the effect of aspirin on heart attacks in a real-world population. It should be noted that the BESD SRD of .034, first, is not an SRD that has been observed in any real-world population but, second, is instead only an estimate or prediction of what the SRD would be, given uniform marginals—that is, in a relatively high-risk ($\pi_{\text{BESD}}/\pi_{\text{OBSERVED}} = 0.5/.0133 > 37$) real-world population, which, third, is based on a functional relation (shown in Figure 1) between the actual SRD and the total success rate (π), in which constancy of ϕ across real-world populations that differ in π values is assumed. Because this function, fourth, is empirically supported by exactly 1 datum point (.0133, .008; see Figure 1), every other point including the BESD SRD = .034 is therefore a risky extrapolation.

In addition, it should be noted that, fifth, if the function (see Figure 1) is realistic, the BESD SRD will overestimate the SRD for all populations in which the total heart attack rate (Π) differs from .50—that is, whenever the uniform marginals assumption is not realistic. Finally, it should be mentioned that, sixth, if the constancy of ϕ assumption is not realistic, BESD SRDs provide no information whatsoever about the magnitude of the effect of aspirin on heart attacks for any real-world population, irrespective of the tenability of the uniform marginals assumption. Clearly, these six caveats raise questions about the extent to which any Case 1 observed BESD SRD value provides, as claimed, practically important information about the effect of a treatment in a real-world population.

What is apparent from Equation 2 (illustrated in Figure 1) is that the choice of a 2 × 2 table with uniform marginals (made in the definition of the BESD) is, in fact, an arbitrary choice to recast data into a 2 × 2 table that maximizes the SRD for any fixed ϕ coefficient. This is because, as noted by Thompson and Schumacker (1997) and Preece (1983), the maximum value of $[\pi(1 - \pi)]^{1/2}$ in Equation 2 occurs for $\pi = .5$ (i.e., uniform Y marginals). From this perspective, then, the BESD SRD is the most optimistic of an infinite number of SRDs that would be consistent with the assumption of constancy of the ϕ coefficient. Therefore, aside from the main problem of the realism of the constancy of ϕ assumption, the next most important problem is that if all known real-world populations yield uneven splits on Y (i.e., not 50% success and 50% failure), BESD SRDs, as operationally defined (see Equation 1), will not only be inaccurate but also will be consistently too large; that is, the BESD SRDs will overestimate the actual SRDs of all known real-world populations.

Case 2: Naturally Dichotomous Independent Variable and Continuous Dependent Variable (the Binormal Model)

The typical comparative efficacy study involves a naturally dichotomous independent variable X (e.g., treatment and control groups, preferably of about equal size; see Hsu, 1993) and a continuous dependent variable Y (e.g., MMPI scale scores; see, e.g., Rosenthal, 2002; Svartberg & Stiles, 1991). Tests of significance of group mean differences that are generally carried out in studies of this type are pooled variance t tests or other parametric tests that assume normality of distributions of Y scores, independence, and homogeneity of variance. This model is referred to as the *binormal* model[3] in the present article (for descriptions of situations in which the homogeneity of variance assumption is unrealistic, see Grissom, 2000).

Effect size indices often used to measure the magnitudes of the effects of X on Y in Case 2 are Cohen's d (see Cohen, 1988) and the r_{pb} (the point-biserial correlation of a dichotomous and a continuous variable that is often inferred from d, t, F ratios, or Cohen's overlap statistics [Us; see Cohen, 1988]). The SRD of the BESD is set equal (see Equation 1) to this r_{pb} (e.g., Rosenthal, 2002; Svartberg & Stiles, 1991). Formulas for r_{pb} and d can be found in the Appendix.

For Case 2 the recasting of the raw data into a 2 × 2 table with uniform marginals (on both X and Y) to determine the BESD SRD requires considering that the same number of persons have been assigned to each of the two levels of X (e.g., the same number of persons have been assigned to the treatment as to the control condition) and that the cut score separating success from failure is the median of the Y score distribution of the combined groups (see Lipsey, 1990). Thus, the uniform marginals assumption for the dependent variable in Case 2 has different real-world implications than the same assumption in Case 1: In Case 1, in which Y is a naturally dichotomous outcome (e.g., heart attack vs. no heart attack, alive vs. dead, etc.), the uniform marginal distribution of Y scores is achieved by hypothesizing that there exists a population in which the success rate is .5; the BESD SRD is then an estimate of an SRD for a population that was not actually observed. Given the tenability of the assumption that such an unobserved population exists, the BESD may (if the constancy of correlations assumption is tenable) provide useful information about the size of an effect in the real world.

In Case 2, on the other hand, the uniform marginals assumption involves changing a cut score for labeling an outcome as a success but does not imply that the BESD SRD is an estimate of an SRD in a different population than the one that was actually observed; rather, the BESD SRD is an estimate of the SRD in the population that was observed, but for a cut score that has been adjusted to yield the uniform marginal Y distribution required by the BESD model. Questions about the realism of this assumption in Case 2 concern whether the median split yields success and failure categories that make real-world sense.

The second assumption of the BESD method for Case 2 is that the observed point-biserial correlation (r_{pb}) of the naturally dichotomous and continuous variables is equal to the ϕ coefficient for the recast 2 × 2 table. This differs from the second assumption for Case 1, which was that the ϕ coefficient of the observed 2 × 2 table equals the ϕ coefficient that would be observed in a study yielding a uniform marginals 2 × 2 table. The tenability of the second assump-

[3] The expression *binormal distribution* is sometimes used interchangeably with *bivariate normal distribution.* However, in this article *binormal* is used to refer to two normal distributions whose variances are equal, one associated with each level of the independent variable: for example, one associated with a treatment condition and the other associated with a control condition. Homogeneity of variance is also a characteristic of the bivariate normal model (see Sheskin, 2000).

tion in Case 1 has to be addressed by determining empirically the ϕ of the new population implied by the first assumption—for example, the actual ϕ in a real-world high-risk ($\Pi = .50$) population of physicians. In contrast, the tenability of the second assumption in Case 2 can be addressed mathematically (given known distributions of Y scores—e.g., given the binormal model). This section focuses on biases of BESD SRDs for the binormal model when the assumption of uniform Y marginals is realistic (Case 2a) as well as when it is unrealistic (Case 2b).

Case 2a: Bias of BESD SRDs When Median Splits on X and Y Are Realistic (Binormal Model)

Rosenthal and Rubin (1982) showed how to calculate exact values of SRDs for the recast binormal data (see the Appendix). After comparing these values to the BESD SRD (viz., r_{pb}), they concluded that the biases of the BESD SRDs (in relation to the exact SRDs) were in most cases small and negligible (Rosenthal & Rubin, 1982). Their method was used to generate, for the binormal model, the exact values of these biases corresponding to r_{pb}s from 0 to 1.00 (see the Case 2a curve in Figure 2). It should be emphasized that these "biases" are differences between actual SRD parameters and BESD SRD parameters when the median split used to define success and failure yields classifications consistent with real-world definitions of success and failure (i.e., when the uniform marginals assumption is realistic).

Consider that the r_{pb} reported in a study is .30 (a realistic value for the point-biserial correlation—e.g., Rosenthal, 2002, noted that the mean r_{pb} obtained in a recent meta-analysis of 479 studies on the experimenter expectancy effect was .30). The BESD SRD would then be .30 (see Equation 1). The vertical distance between the Case 2a curve in Figure 2 and the no-bias line at the abscissa (i.e., the BESD SRD) of .30 is .05, showing that the BESD SRD overestimates the SRD by .05. What becomes immediately apparent from an inspection of the Case 2a curve in Figure 2 is that the bias of BESD SRD is (a) generally small (consistent with comments of Rosenthal & Rubin, 1982) and, more important, (b) generally positive (overestimation) for values of r_{pb} that are most frequently reported in the literature ($r_{\mathrm{pb}} < .76$). Point (b) is particularly important because of the intent that BESDs provide practically useful information about magnitudes of effects in real-world populations: Note that an r_{pb} correlation of .76 corresponds to a d value of 2.34 in the binormal model (see Cohen, 1988) and that ds > 2.34 are rarely reported in meta-analyses (see Lipsey & Wilson, 1993); values of $r_{\mathrm{pb}} > .76$ are equally rare (see Meyer et al., 2001).[4] Meta-analyses that rely on BESD SRDs can therefore be expected to consistently overestimate effect sizes of Case 2a studies.

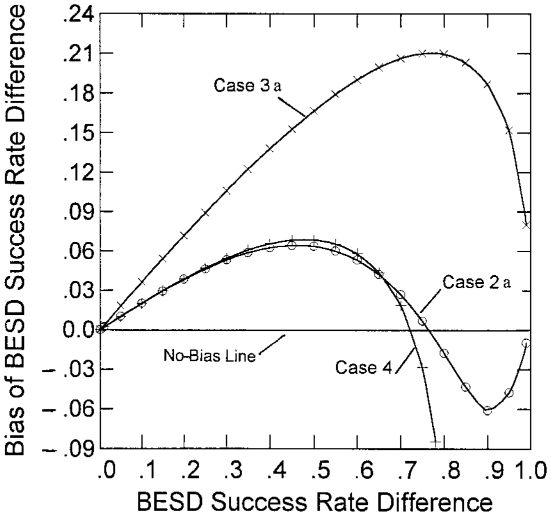

Figure 2. Extent to which the binomial effect size display success rate differences (BESD SRDs) over- and underestimate the actual success rate differences in Cases 2, 3, and 4 when the assumption of uniform marginals for X and Y is tenable.

Case 2b: BESD SRD Biases When the Median Split Definition of Success and Failure Is Not Realistic (in the Binormal Model)

The principal published criticism of the BESD applied to binormal data concerns the lack of realism of the median split of the dependent variable Y required for the uniform marginal distributions in BESDs. As noted by Preece (1983),

> the use of a median split to dichotomize the outcome variable may be inappropriate. Success may be a cure following psychotherapy or gaining entrance to university . . . and in such cases the overall success rate [π] cannot be controlled by the experimenter and the value of 0.50 may be quite unrealistic. (p. 764)

Similar points appear in Thompson and Schumacker (1997), together with the warning that "the BESD [can] greatly [exaggerate] the change in success rate . . . when the binomial success rate departs markedly from 50%" (p. 114). Realism of the uniform marginals assumption for the independent variable has not been called into question by most critics (and is also not considered an issue in the present article) because the independent variable is usually under the control of researchers, and because researchers generally choose to include an (approximately) equal number of persons in treatment and control groups (see, e.g., Cohen, 1988;

[4] The presence of publication and accessibility biases in meta-analyses, especially when targeted studies involve treatment versus control group contrasts (Hsu, 2000, 2002b), implies that ds > 2.34 and rs $> .76$ actually occur even more rarely in real-world populations than has generally been reported in meta-analyses.

Hsu, 1993; Kazdin, 2003; Thompson & Schumacker, 1997). Furthermore, when X is a continuous predictor variable, it is almost always both feasible and meaningful to dichotomize this variable at the median of the combined groups (see Case 4 below).

Although critics have drawn attention to the inaccuracies of BESD SRDs when the Case 2 median Y split is unrealistic, none have actually calculated the biases caused by this lack of realism. Furthermore, neither the formulas presented in Rosenthal and Rubin (1982) nor those proposed and used in Preece (1983) and Thompson and Schumacker (1997) are relevant to the measurement of Case 2b biases. However, the directions and sizes of Case 2b biases may easily be determined: Define the cut score (z_c) that separates successes from failures on the outcome measure on Cohen's d scale (whose unit of measurement is σ, the common within-group standard deviation), equating the zero point on the scale with the mean of the control group. Then, given that in the binormal model distributions of the outcome measure are normal with equal variances, the success rate for each group (X value) may be determined for any cut score z_c:

$$P(\text{Success}|\text{Control Group}) = P(z > z_c)$$

$$P(\text{Success}|\text{Treatment Group}) = P[z > (z_c - d)], \qquad (4)$$

where z = the standard normal variable. The actual SRD for the binormal model will therefore be as follows:

$$\text{Actual SRD} = P[z > (z_c - d)] - P(z > z_c). \qquad (5)$$

Because the BESD SRD is defined, in Case 2, as follows (see Equation 1),

$$\text{BESD SRD} = r_{pb}, \qquad (6)$$

the bias of the BESD SRD that is caused by lack of realism of the median split on the outcome measure (Y) is therefore as follows:

$$\text{Case 2b BESD SRD bias} = r_{pb} - \{P[z > (z_c - d)] - P(z > z_c)\}. \qquad (7)$$

The following example illustrates how Equations 4–7 can be used to measure the BESD SRD biases for Case 2b. Consider that the d reported for a study is 0.63 (i.e., the means of the treatment and control groups are 0.63 standard deviation units apart) and that a realistic definition of success places the cut score at 2.00. (Note that a cut score 2 standard deviations from the mean of the functional population is one of Jacobson & Truax's, 1991, three definitions of a cut score that can be used to identify patients who benefit in a clinically significant way in therapy; see also Hsu, 1996.) That is, $z_c = 2.00$. Then Equations 3 and 4 imply that the success rates for the control and treatment groups would be $P(z > 2.00) = .023$, and $P[z > (2.00 - 0.63)] = .085$, respectively. The actual SRD_{pb} would therefore be (by Equation 5) $.085 - 0.023 = .062$. The BESD SRD is the r_{pb} corresponding to $d = 0.63$, namely, $r_{pb} = .30$ (as determined from Cohen's, 1988, p. 23, formula relating d to r_{pb} in the binormal model):

$$r_{pb} = d/(d^2 + 4)^{.5}.$$

Therefore, using Equation 7, the bias of the BESD SRD relative to the actual SRD yielded by Equation 5 would be $.300 - .062 = .238$. In other words, the BESD SRD overestimates the SRD by .238, or, interpreted as a ratio, the BESD SRD is almost 5 times (i.e., $.300/.062 = 4.839$) too large. Equations 4–7 were used to generate actual SRDs for two r_{pb}s (.30 and .70) for various cut scores extending from $z_c = -2.00$ to $z_c = 2.60$ for the binormal model. These SRDs are shown in Figure 3, together with the BESD SRDs (horizontal lines) corresponding to the two r_{pb}s. For each cut score (defined on Cohen's d scale) the vertical distance between the actual SRD and the corresponding BESD SRD indicates the size of the bias. Thus, for the example, the actual SRD for a cut score of 2.00 is indicated by an ordinate of .062 on the lower curve, and the BESD SRD for this cut score (or any other cut score) is .30 for an $r_{pb} = .30$ and is indicated by an ordinate of .30, so that the vertical distance between these two points indicates a bias of .238 (the value calculated for the above example). This assumes that the realistic definition of success involves the Jacobson and Truax (1991) clinical significance cut score—that is, an outcome measure greater than $z_c = 2.00$. Vertical distances

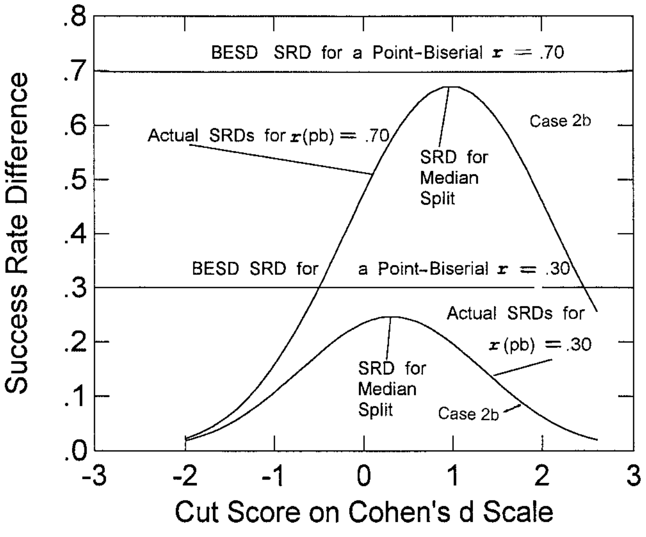

Figure 3. Comparisons of binomial effect size display success rate differences (BESD SRDs) to actual SRDs for various cut scores (and total success rates) and for two values of point-biserial correlations (.30, .70) given binormal distributions.

corresponding to cut scores $z_c = 0.50$, that is, abscissas of .32 (for $r_{pb} = .30$) and .98 (for $r_{pb} = .70$), indicate biases corresponding to median splits in the Y distributions of the combined groups.

Several important facts are illustrated in Figures 2 and 3. First, for r_{pb}s that are less than about .76 (i.e., $d < 2.34$), the biases, when the Y median split is realistic, are consistently overestimation biases (assuming that results of the study were in the predicted direction), suggesting overly optimistic estimates of efficacy of the treatment (see the Case 2a curve in Figure 2). Second, paralleling Case 1 findings (see Equation 2) of Thompson and Schumacker (1997) and Preece (1983), the overestimation biases for Case 2 (based on Equations 3–7) are very large for Y splits that differ considerably from a median split (see Figure 3). Third, the minimum overestimation biases occur for the median splits (see Figure 3). Fourth, the size of the bias is calculable from information about the realistic cut score (on Cohen's d scale) that differentiates successes from failures and from the value of the point-biserial correlation (see Figures 2 and 3). Fifth, because the point-biserial correlation is generally reported (or may be calculated from reported statistics), this implies that the direction and size of the BESD SRD bias can generally be determined, using readily available information, for any cut score that a researcher considers to provide a reasonable definition of success and failure.

Case 3: Continuous Independent (X) and Dependent (Y) Variables (the Bivariate Normal Model)

The typical Pearson correlation for two continuous variables that has been interpreted using the BESD is probably a concurrent or predictive validity correlation involving two scales or one scale and one continuous criterion variable (see, e.g., Hiller et al., 1999). Tests of significance of correlations of continuous variables (which are usually carried out in studies that have yielded the r_{xy}s) have often been derived under assumptions of bivariate normality (see, e.g., Walker & Lev, 1953). This section focuses on biases of the BESD SRD when the X and Y variables associated with an empirically determined r_{xy} have a bivariate normal distribution; these biases are measured when the assumption of uniform marginals is realistic (Case 3a) as well as when it is not realistic (Case 3b).

Case 3a: Bias of BESD SRDs When Median Splits on X and Y Are Realistic

The uniform marginals assumption of the BESD implies median splits on both X and Y in the bivariate normal distribution. Designating above-median scores on X and Y as $X+$ and $Y+$ and below-median scores as $X-$ and $Y-$, respectively, and defining $Y+$ as success, we may define the actual SRD for the bivariate normal distribution as follows:

$$\text{SRD} = P(Y+|X+) - P(Y+|X-), \tag{8}$$

where $P(Y+|X-)$ = the success rate given a performance below the X median and $P(Y+|X+)$ = the success rate given a performance above the X median. In contrast with the actual SRD, the BESD SRD was defined by Rosenthal and Rubin (1982) as follows (see Equation 1):

$$\text{BESD SRD}_{xy} = r_{xy}. \tag{9}$$

Given that the median splits on X and Y are realistic, the bias of BESD SRD in relation to Pearson r_{xy}s for continuous X and Y scores is therefore the difference between Equations 9 and 8:

$$\text{Case 3a BESD bias} = r_{xy} - [P(Y+|X+) - P(Y+|X-)]. \tag{10}$$

The calculation of actual SRDs defined in Equation 8 requires determination of volumes under bivariate normal distributions. Stuart and Ord (1987, p. 482) showed that the calculation of volumes under bivariate normal distributions is very complicated except in the case of median splits on both X and Y. However, given median splits, they noted that the volume corresponding to ($X+ \cap Y+$) could be obtained from the following:

$$P(X+ \cap Y+) = (¼ + \arcsin r_{xy}/2\pi_c), \tag{11}$$

where $\pi_c = 3.1416$ (vs. π = total success rate in Equation 2; Stuart & Ord, 1987, credited Sheppard, 1898, for deriving Equation 11).

Vargha, Rudas, Delaney, and Maxwell (1996, pp. 268–269) recently derived an equation relating r_{xy} to ϕ in bivariate normal distributions given median splits on X and Y:

$$\phi = (2/\pi_c) \arcsin (r_{xy}) = .637 \arcsin (r_{xy}). \tag{12}$$

Equation 12 clearly shows that the ϕ coefficient obtained with median splits of X and Y variables that have a bivariate normal distribution will, in general, not be equal to the correlation of the two continuous variables X and Y (viz., r_{xy}); thus, Equation 12 explicitly shows the degree to which the constancy of correlation assumption of the BESD (in this case the assumption that $\phi = r_{xy}$ given median splits on X and Y) is violated. Equation 11 was used (Equation 12 could also have been used) to determine actual values of the SRDs (for r_{xy}s from 0 to 1.00) that were then used to calculate the biases of BESD SRDs defined in Equation 10 considering that the assumption of uniform marginals (median splits) was tenable. These biases have been plotted against BESD SRDs in Figure 2 (see the Case 3a curve).

What is immediately apparent from Figure 2 (see the Case 3a curve) is that when the uniform marginals assumption is realistic (i.e., when the median splits yield dichotomies that make sense in the real world), (a) the BESD SRDs will consistently overestimate the targeted SRDs, where the targeted SRDs are the actual SRDs that would result from the median splits, and (b) the overestimation biases are generally not negligible.

Case 3b: Biases of BESD SRDs Given a Bivariate Normal Distribution of X and Y and a Median Split on Y That Is Not Realistic

When the median split in the distribution of outcome measures is not realistic, Equations 11 and 12 will not provide the information needed to determine the biases of the BESD SRDs. Fortunately, tables from which this information can be determined are available: Those provided in Taylor and Russell (1939) were used in order to generate actual SRDs for the bivariate normal case for cut scores on the outcome measure corresponding to success rates (for the combined groups) ranging from .05 to .95, and these actual SRDs are depicted in Figure 4 for two values of the Pearson correlation coefficient r_{xy} (.30 and .70). Several important facts are illustrated in Figure 4 concerning biases of BESD SRDs for Case 3b. First, irrespective of whether or not the median split assumption is tenable, the biases are consistently overestimation biases (assuming that results of the study were in the predicted direction), suggesting overly optimistic estimates of efficacy of the treatment. Second, paralleling findings of Thompson and Schumacker (1997) and Preece (1983) for Case 1, and findings presented in relation to Case 2b, the overestimation biases for Case 3b can be very large for *Y* splits that differ considerably from a median split. Third, minimum biases occur for median splits, although in contrast with Case 2a, these biases (for Case 3a) are larger and generally not negligible. Fourth, the size of the bias is predictable from information that is generally reported (the Pearson *r*) or that can be calculated from reported statistics (*t* ratios, *F* ratios, etc.) for any cut score that a researcher considers realistic in defining success and failure. The major difference between biases for Cases 2b and 3b is that, for the same degree of violation of the uniform marginals assumption, the bias in Case 3b is always larger than that in Case 2b.

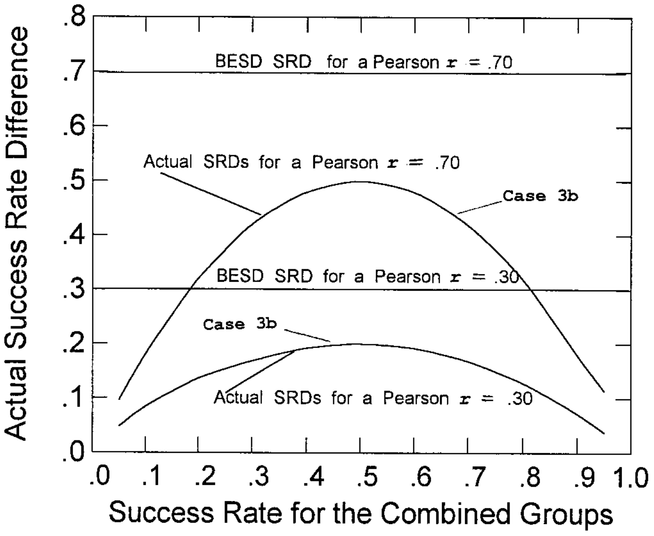

Figure 4. Comparisons of binomial effect size display success rate differences (BESD SRDs) to actual SRDs for various total success rates and for two values of Pearson correlations (.30, .70) given bivariate normal distributions.

Case 4: Artificially Dichotomized *X* and Continuous *Y* Variables: Underlying Bivariate Normal Distribution of *X* and *Y*

Recent derivations by Vargha et al. (1996) allow the measurement of BESD SRD biases in Case 4 in which (a) the reported correlation is a point-biserial correlation (r_{pb}) of an artificially dichotomized variable *X* (median split on *X*) and continuous variable *Y* and (b) the underlying distribution of the continuous variables *X* and *Y* is bivariate normal. Vargha et al. showed (see also Nunnally, 1978) that

$$r_{\text{pb}} = .798(r_{xy}). \tag{13}$$

Because the BESD SRD is set equal to the reported r_{pb} of a study and because the ϕ coefficient for median splits on *X* and *Y* is related to the correlation of the continuous variables *X* and *Y* as indicated in Equation 12, it is clear that (for any fixed r_{xy}) the bias of the BESD SRD in Case 4 must be as follows:

$$\text{Case 4 BESD SRD bias} = [.798(r_{xy}) - .637 \arcsin(r_{xy})], \tag{14}$$

where r_{xy} is the correlation of the continuous variables whose distribution is bivariate normal. Equation 14 was used to generate the curve (see Case 4 curve in Figure 2) showing biases of the BESD SRD for Case 4.

Vargha et al. (1996) showed that for any r_{xy} below .9075 (which corresponds to $r_{\text{pb}} = .724$ by Equation 13), the ϕ coefficient obtained with median-split dichotomies on both *X* and *Y* will be smaller than the r_{pb} that would be obtained with a median-split dichotomy of just one of these variable. What this implies, in relation to the BESD SRD biases for Case 4, is that the BESD SRD (which is defined as equal to the reported r_{pb} by Rosenthal, 1994) will overestimate the actual ϕ coefficient for median splits on both *X* and *Y* whenever the empirically determined r_{pb} is less than .724. In general, Figure 2 shows similar biases for Cases 2a and 4: In both of these cases the BESD SRDs associated with r_{pb}s overestimate targeted SRDs whenever the reported r_{pb} is in the range of correlations most commonly reported in meta-analyses (i.e., <.724 for Case 4 and <.76 for Case 2).

Some Implications of Biases and of Differences in Biases of BESD SRDs Corresponding to r_{pb}s and r_{xy}s

Figure 2 draws attention to several important facts concerning biases of BESD SRDs in situations involving the r_{pb}s and r_{xy}s when the assumption of uniformity of marginals is tenable and when the observed success rates of contrasted groups are in the predicted direction. First, BESD SRD biases associated with r_{xy}s are generally larger and at times much larger than those associated with Cases 2a and 4 r_{pb}s. Second, positive biases (i.e., overestimation of actual SRDs) are consistently present for BESD SRDs associated with all values of r_{xy} and are consistently present for BESD SRDs associated with r_{pb}s that are (a) less than about .76 (i.e., $d < 2.34$) in Case 2 and (b) less than .724 (i.e., $r_{xy} < .9075$) in Case 4. Third, generally small negative biases are present for BESD SRDs associated with values of r_{pb} (a) greater than about .76 (i.e., $d > 2.34$) in Case 2 and (b) greater than .724 (i.e., $r_{xy} > .9075$) in Case 4. All of this information suggests that in real-world applications, the BESD SRDs for Case 2a, Case 3a, and Case 4 statistics will virtually always overestimate treatment effects, even when the uniformity of marginals assumption is tenable.[5]

Figure 2 also implies that even when the uniformity of marginals assumption is true, very different SRDs may correspond to identical values of r_{xy} and r_{pb} (that must yield the same BESD SRD). For example, if the reported correlation is .80, and therefore the BESD SRD is .80, the actual SRD will be about .59 (because the bias $= +.21$) if the correlation is an r_{xy} and will be about .83 (because the bias $= -.03$) if it is a Case 2 r_{pb}. This type of information raises questions about the validity of inferences about equality of effect sizes (which have appeared in recent meta-analyses—e.g., Hiller et al., 1999) based on equal BESD SRDs corresponding to different types of correlations (r_{xy} and r_{pb}). Figure 2 clearly shows that equality of these BESD SRDs does not imply equality of actual SRDs. In fact Figure 2 leads to the paradoxical conclusion that two equal BESD SRDs, one based on a Case 3a r_{xy} and the other on a Case 2a or 4 r_{pb}, should be interpreted as evidence of a larger SRD for the r_{pb} because of the larger overestimation bias for the r_{xy}.

The above conclusion is, of course, based on assumptions of binormal distributions of Y scores for Case 2a and underlying bivariate normal distributions of X and Y scores for Cases 3a and 4. These are assumptions under which tests of significance of r_{pb} and r_{xy} are generally derived (as noted above) and are assumptions that are usually not questioned by researchers who use these significance tests. However, it is not necessarily the case that the binormal and/or bivariate normal distributional assumptions are true (see Cliff, 1993, 1996; Grissom, 2000). Unfortunately, information about BESD SRD biases for other distributional assumptions has apparently been provided for only one case—namely, the case in which X is a naturally dichotomous variable and Y is a continuous variable with a "bi-t-distribution." That is, the distribution of Y scores associated with each value of X is a Student's t distribution. Rosenthal and Rubin (1982), who examined this case, indicated that BESD SRDs overestimated targeted SRDs more (or underestimated these SRDs less) for the binormal distribution model than for the bi-t-distribution model. The major implication of this information, in relation to the biases associated with r_{xy} that were determined in the present article, is that the statement that BESD SRDs overestimate targeted SRDs more for r_{xy}s than for r_{pb}s of equal size is even more true for the bi-t-distribution model than it is for the binormal model.

Figures 3 and 4 show that violation of the uniform marginals assumption generally results in larger biases than those shown for corresponding values of r in Figure 2. Also, for the same departure from the uniform Y marginals distribution, the bias is generally larger for r_{xy} than for r_{pb}.

Alternatives to the BESD: Stochastic Superiority and Difference Indices

All of the indices reviewed in Huberty's (2002) "history of effect size indices" (p. 227) could be viewed as alternatives to the BESD SRD. Similarly, all of the effect size indices proposed by earlier critics of the BESD—including "relative [risk]" (Crow, 1991, p. 1083) and success rate ratios (Preece, 1983; Strahan, 1991)—are defensible alternatives to the BESD. However, none of these indices are as similar to the BESD in generality of application, objective, definition, and simplicity as the "stochastic difference" (Vargha & Delaney, 2000, p. 104) index (δ) recently discussed in Cliff (1993, 1996) and Vargha and Delaney.

Given Populations 1 and 2 (a dichotomous or dichotomized X variable) the "stochastic difference [δ] of Populations 1 (say, Treatment) and Population 2 (say, Control) with respect to the dependent variable [Y]" (Vargha & Delaney, 2000, p. 102) is defined in general (see also Cliff, 1993, 1996) as follows:

$$\delta = P(Y_1 > Y_2) - P(Y_1 < Y_2), \tag{15}$$

[5] The argument that uniformity of marginals and constancy of correlations could be viewed as postulates (see Case 1) faces an incompatibility problem in Cases 2a, 4, and especially 3a, in addition to problems of realism: Splitting the X and/or Y distributions to satisfy the uniform marginals postulate generally implies inequality of correlations—that is, the ϕ for the 2 × 2 table will not equal the r_{pb} (Cases 2a and 4) or the r_{xy} (Case 3a)—and is therefore incompatible with the constancy of correlations postulate, for example, median splits on X and Y, given a Case 3a $r_{xy} = .70$, imply $\phi = .49$ and not $\phi = .70$. Uniformity of marginals (realistic or not) can therefore clearly imply nonnegligible inequality of correlations.

Table 1
AZT Data Set (see Thompson & Schumacker, 1997) and Notation for Comparison of the Stochastic Difference Index of Effect Size to the Actual Success Rate Difference, Binomial Effect Size Display Success Rate Differences, ϕ, and Equation 2

	Notation		Data		
Outcome	Control	Treatment	Placebo	AZT	Total
Success	$(Y_2 = 1)$ b	$(Y_1 = 1)$ a	121	144	265
Failure	$(Y_2 = 0)$ c	$(Y_1 = 0)$ d	16	1	17
Group size	*m*	*n*	137	145	282

Note. a, b, c, d, *m,* and *n* are counts or frequencies. *Y* is an indicator variable: $Y_2 = 1$ indicates success for a control participant, and b = the number of control participants who were successful; $Y_1 = 1$ indicates success for a treatment participant, and a = the number of treatment participants who were successful, and so on. $N = \text{a} + \text{b} + \text{c} + \text{d} = m + n$. For the AZT (also known as Retrovir) data set, success equals survival and failure equals death.

where Y_1 and Y_2 are independently and randomly drawn scores from Populations 1 and 2, respectively, and δ is defined in terms of the "stochastic superiority" index *A* (see Grissom, 1994a, 1996; Vargha & Delaney, 2000, p. 102). When *Y* is a continuous variable, the stochastic superiorities of Population 1 over Population 2, and of Population 2 over Population 1, are defined, respectively, as follows:

$$\text{A}_{12} = P(Y_1 > Y_2)$$

and

$$\text{A}_{21} = P(Y_1 < Y_2).$$

Thus, δ can also be defined as $A_{12} - A_{21}$. For a special case that would be relevant to Case 2—the binormal model—see McGraw and Wong (1992) and Grissom (1994a, 1994b) and for a discussion of the relation of this case to various diagnostic validity indices, see Hsu (2002a).[6] Therefore, δ is the difference between (a) the probability that the outcome measure *Y* of a randomly drawn person from Population 1 exceeds that of an independently and randomly drawn person from Population 2 and (b) the probability that the outcome measure *Y* of a randomly drawn person from Population 1 is less than that of a randomly drawn person from Population 2; δ is a quantification of the statement that persons exposed to one treatment tend to score higher on an outcome measure (*Y*) than persons exposed to another (see Cliff, 1993).

Both the BESD SRD and δ are magnitudes of effect indices that are intended to provide practically useful information; both indices focus on contrasts of proportions (rates, probabilities) to provide this information; both indices require dichotomous (or dichotomized) *X* variables; both indices can be determined from empirical data that yield values of ϕ, r_{pb}, and r_{xy} (with dichotomization of *X* in the case of r_{xy}).

The principal difference between the BESD SRD and the stochastic difference index is that the parameter δ generally provides meaningful information about the magnitudes of effects in real-world populations, irrespective of the tenability of the BESD's assumptions of uniformity of *X* or *Y* marginals and of constancy of correlations. In fact δ does not even involve any marginal assumptions concerning distributions of *X* or *Y* scores and does not even require assumptions of interval or ratio scaling of *Y* scores (see Cliff, 1993, p. 495): It is even interpretable when the level of measurement of *Y* is only ordinal.

The sample estimate *d** (note that the asterisk is used to avoid conflict in notation with Cohen's *d;* note also that *d** was called a "dominance statistic" in Cliff, 1993, p. 494) of δ, given *m* persons in one group and *n* in the other, is (see Cliff, 1993) as follows:

$$d^* = \frac{\#(Y_1 > Y_2) - \#(Y_1 < Y_2)}{(m)(n)} = \frac{\#(Y_1 > Y_2)}{(m)(n)} - \frac{\#(Y_1 < Y_2)}{(m)(n)}, \quad (16)$$

where

> # denotes *the number of* or *number of times*) . . . [and] each of the *n* [*Y*s] in one group is compared with each of the *m* [*Y*s] in the other, and counts are made of how many times [the score of a] member of the first group is higher and how many times it is lower. (Cliff, 1993, p. 495)

Thus, in Case 1 (the case in which both *X* and *Y* are naturally dichotomous variables—this case is most directly relevant to comparison of the stochastic difference with the 2 × 2 BESD) there will be (*m*)(*n*) pairings of *Y* scores from the two groups, (a)(c) yielding ($Y_1 > Y_2$) and (b)(d) yielding ($Y_1 < Y_2$) (see Table 1). Thus, *d** for Case 1 is as follows:

$$d^* = \frac{(\text{ac} - \text{bd})}{\text{mn}}. \quad (17)$$

Note that *d** is identical (in Case 1) to a form of Kendall's τ and to Somer's *d* statistic (see Cliff, 1993; Vargha & Delaney, 2000). The relation of *d** to (a) BESD SRDs, (b)

[6] An anonymous reviewer noted that the relative sizes of A_{12} and A_{21} can also be defined in terms of an odds ratio (see Agresti, 1989).

actual SRDs, (c) ϕs, and (d) Preece's (1983) Equation 2 can now be illustrated for Case 1 using results of an AZT study (see Table 1) that has been cited in several of Rosenthal's (1990, 1994, 1995a) writings. This study was concerned with the efficacy of AZT in the treatment of AIDS (see the right side of Table 1). The number of possible pairings of AZT and placebo participants is $(m)(n) = (145)(137) = 19{,}865$ (because the m AZT participants can be paired with the n placebo participants in (m)(n) ways. Of these pairings, $(\mathrm{a})(\mathrm{c}) = (144)(16) = 2{,}304$ would yield $(Y_1 > Y_2)$, that is, better results for the AZT than for the placebo participant in the pairing, and $(\mathrm{b})(\mathrm{d}) = (121)(1) = 121$ would yield ($Y_1 <$ Y_2). Therefore, if we apply Equation 17,

$$d^* = (2{,}304)/(19{,}865) - (121)/(19{,}865) = .1099.$$

In order to understand the relation of d^* to the actual SRD for the AZT study, we may rewrite Equation 17—because $\mathrm{c} = (m - \mathrm{b})$ and $d = (n - \mathrm{a})$; see Table 1—as follows:

$$d^* = \frac{\mathrm{a}(m - b)}{mn} - \frac{\mathrm{b}(n - \mathrm{a})}{mn} = \frac{\mathrm{a}}{n} - \frac{\mathrm{b}}{m}. \qquad (18)$$

In other words, for Case 1 studies, d^* = observed SRD. Note that in the AZT study the actual SRD calculated using Equation 18 is $(144/145) - (121/137) = .1099$, which is exactly equal to the stochastic difference calculated using Equation 17.

It is important to note that because Equation 18 involves no restrictions on values of a, b, c, d, m, and n (see Table 1), the stochastic difference estimate (in Case 1 studies) will always be identical to the actual SRD irrespective of the marginal distributions of either of the two dichotomous variables, and that because (unlike the BESD) there is no recasting of the data, the (BESD's) assumption of constancy of correlations across tables is irrelevant. That is, the stochastic difference estimate always coincides with a magnitude of effect that is easy to interpret as well as entirely consistent with the empirically observed SRD.

In contrast, the BESD SRD generally suggests an effect that is difficult to interpret because it is apparently inconsistent with the observed data: For example, the BESD table of the AZT study (see Rosenthal, 1990) shows an AZT-caused reduction of 23 percentage points in the death rate, an SRD that is more than double (viz., .23) the observed SRD (viz., .1099; see Thompson & Schumacker, 1997). The "justification" for the inference (and it must be recognized as an inference) that AZT would reduce the death rate by 23 percentage points in a real-world population requires tenability of the assumption of uniform marginals (viz., that there exists a high-risk population with AIDs in which the death rate of combined AZT and placebo patients would be .50) and of the assumption of constancy of correlations (that the ϕ coefficient that would be observed in such a population would be equal to that reported in the actual [low risk: 17/282 = .06] study). Both of these assumptions have to be true if the BESD SRD of 23% is to provide correct information about effects of AZT in a real-world population. Clearly both the stochastic difference index calculated from the raw data and the observed SRD, whose interpretations are in no way affected by the lack of realism of these assumptions, are more defensible indices of the real-world importance of the effect of AZT than is the BESD SRD.

The relation of the stochastic difference estimate d^* to the ϕ coefficient and to Preece's Equation 2 (in Case 1) is apparent given that ϕ can be expressed (see, e.g., Vargha et al., 1996), using the notation of Table 1, as follows:

$$\phi = \frac{\frac{(\mathrm{a})}{N}\frac{(\mathrm{c})}{N} - \frac{(\mathrm{b})}{N}\frac{(\mathrm{d})}{N}}{\sqrt{\frac{(\mathrm{a}+\mathrm{b})}{N}\frac{(\mathrm{c}+\mathrm{d})}{N}\frac{\mathrm{m}}{N}\frac{\mathrm{n}}{N}}}.$$

Given uniform marginals on X, so that $(m/N) = (n/N) = .5$, and letting $\pi = (\mathrm{a} + \mathrm{b})/N$ and $(1 - \pi) = (\mathrm{c} + \mathrm{d})/N$, it is clear that $d^* = (4/N^2)(\mathrm{ac} - \mathrm{bd})$ and that

$$\phi = \frac{1}{N^2}\frac{(\mathrm{ac} - \mathrm{bd})}{1}\frac{2}{[\pi(1 - \pi)]^{0.5}}$$
$$= (1/4)d^*\frac{2}{[\pi(1 - \pi)]^{0.5}}, \qquad (19)$$

which, when solved for d^*, yields Preece's Equation 2 expressed in terms of the stochastic difference estimate:

$$(d^*) = 2\phi[\pi(1 - \pi)]^{.5}. \qquad (20)$$

It is also apparent from Equation 17 that with a uniform distribution on X, $d^* = \kappa$ (Cohen's kappa is arguably the most popular chance-corrected measure of agreement between categorical X and Y variables; see Hsu & Field, 2003), so that Equation 2 can also be reexpressed in terms of Cohen's kappa:

$$\kappa = 2\phi[\pi(1 - \pi)]^{.5}. \qquad (21)$$

What becomes apparent from Equations 2, 20, and 21 is that the constancy of ϕ assumption, which is indispensable to the definition of the BESD SRD, is equivalent not only to an assumption of a specific functional relation between the overall success rate (π) and the actual SRD (illustrated in Figure 1) but also to an assumption of the same functional relation (in which ϕ is a constant) between the stochastic difference index and π and between Cohen's kappa and π. Whether the expression on the left of these equations is actual SRD, stochastic difference, or Cohen's kappa, no empirical evidence of tenability of this assumption, has (as far as I know) ever been provided by any who have endorsed and/or used the BESD.

The stochastic difference estimate d^* calculated from data of a Case 1 study will (unlike a value of d^* obtained from Equation 20 under the BESD SRD assumption of constancy of ϕ) generally be an unbiased estimate of δ (see Cliff, 1993; Vargha & Delaney, 2000). However, the most attractive property of the stochastic difference estimate d^* (see Equation 16) is not its ability to provide unbiased estimates of δ in Case 1 studies in which the outcome variable is dichotomous, but rather its ability to provide unbiased estimates of δ irrespective of whether the outcome measure is (a) continuous or discrete or (b) ordinal, interval, or ratio scaled and also irrespective of (c) the shapes (skewness, kurtosis, heterogeneity of variances) of the distributions of the outcome measures (see Cliff, 1993; Vargha & Delaney, 2000). Thus, not only will d^* provide realistic and useful information about magnitudes of effects under ordinary assumptions of parametric significance tests (see, e.g., McGraw & Wong, 1992), as in Cases 2, 3, and 4 (above), but it will also provide realistic and useful information about magnitudes of effects in the absence of any of these assumptions (see Cliff, 1993, 1996; Grissom & Kim, 2001; Vargha & Delaney, 2000). Another advantage of the stochastic superiority and stochastic difference indices over the BESD SRD is the recent development of easily calculated point and interval estimates and hypothesis tests for δ and A parameters (see Brunner & Munzel, 2000; Cliff, 1996; Delaney & Vargha, 2002; Vargha & Delaney, 2000). However, certain limitations of the stochastic difference and stochastic superiority indices should be noted (see also Grissom & Kim, 2001): (a) the calculation of these indices is slightly (see Grissom, 1994b, for details) more complicated than the calculation of BESD SRDs (for relevant software, see Wilcox, 1997, 2003); (b) with a continuous outcome measure the calculation of these indices requires access to the raw data (unless certain distributional assumptions are tenable; see, e.g., McGraw & Wong, 1992); (c) although calculations of these indices can be made from values of some nonparametric statistics (e.g., Mann–Whitney–Wilcoxon statistics; see Vargha & Delaney, 2000; Wilcox, 1997, 2003), these have rarely been reported in published studies; and (d) algorithms for calculation of d^* have not yet been incorporated in the most popular statistical software packages. These limitations should be recognized as problems of implementation rather than problems of interpretation. Nevertheless, they are not negligible problems, especially for those using these indices in meta-analyses.

Conclusions

The first conclusion implied by Equations 1–14 and Figures 1, 2, 3, and 4 is that BESD SRDs tend to overestimate targeted real-world SRDs in virtually all real-world applications involving Case 1, Case 2a, Case 2b, Case 3a, Case 3b, and Case 4 parameters. The second conclusion (given tenability of the uniform marginals assumption of the independent variable X and given that results are in the predicted direction) is that overestimation biases will be present in almost all of these applications whether or not the uniform marginals assumption for the dependent variable Y is tenable, but overestimation biases will generally be greater (and possibly much greater) when the uniform marginals assumption is not tenable than when it is. The third conclusion is that four major factors determine the sizes of the BESD SRD overestimation biases: (a) the levels of measurement (categorical, continuous) of the independent and dependent variables (e.g., larger overestimation biases are present in Case 3 r_{xy} studies, involving continuous X and Y variables, than in Case 2 r_{pb} studies, involving one naturally dichotomous and one continuous variable), (b) the degree of violation of the uniform marginals assumption (e.g., overestimation biases generally increase with increases in degree of violation of the uniform marginals' Y distribution), (c) the nature of the distributions of the X and Y variables (e.g., different biases are associated with binormal than with bivariate normal distributions), and (d) the sizes of the BESD SRDs (e.g., generally larger overestimation biases are associated with larger BESD SRDs).

Rosenthal (1990, 2002) has expressed the belief that effect size estimates other than the BESD SRD have a "tendency to underestimate the practical importance of the effects of behavioral or biomedical interventions," (Rosenthal, 2002, p. 844) and that "BESDs [present] a better picture of the real-world importance of any treatment effect" (Rosenthal, 2002, p. 844). However, the first two conclusions drawn in the present article suggest a caveat—BESD SRDs tend to overestimate magnitudes of effects, both when the uniform marginals assumption is met and when it is not met. The third conclusion is at odds with the current practice (see, e.g., Hiller et al., 1999; Rosenthal, 2002; Svartberg & Stiles, 1991) of making comparative statements about effect sizes based on BESD SRDs: Any one or any combination of the four factors (see a, b, c, and d, described previously) that determine sizes of BESD SRD biases can clearly invalidate such comparisons.

The stochastic difference and stochastic superiority effect size indices (Cliff, 1993, 1996; Vargha & Delaney, 2000), which are applicable to dichotomous-, ordinal-, interval-, and ratio-scaled response measures, and which are not invalidated by violations of assumptions of uniform marginals or of constancy of correlations, provide more realistic indices of magnitudes of effects in real-world populations than the BESD SRDs.

References

Agresti, A. (1989). Tutorial on modeling ordered categorical response data. *Psychological Bulletin, 105,* 290–301.

Brunner, E., & Munzel, U. (2000). The nonparametric Behrens–Fisher problem: Asymptotic theory and a small sample approximation. *Biometric Journal, 42,* 17–25.

Carroll, J. B. (1961). The nature of data, or how to choose a correlation coefficient. *Psychometrika, 26,* 347–382.

Cliff, N. (1993). Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin, 114,* 494–509.

Cliff, N. (1996). *Ordinal methods for behavioral data analysis.* Mahwah, NJ: Erlbaum.

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

Crow, E. L. (1991). Response to Rosenthal's comment "How are we doing in soft psychology?" *American Psychologist, 46,* 1083.

Delaney, H. D., & Vargha, A. (2002). Comparing several robust tests of stochastic equality with ordinally scaled variables and small to moderate sized samples. *Psychological Methods, 7,* 485–503.

Fleiss, J. L., Levin, B., & Paik, M. C. (2003). *Statistical methods for rates and proportions* (3rd ed.). New York: Wiley.

Ghiselli, E. E., Campbell, J. P., & Zedeck, S. (1981). *Measurement theory for the behavioral sciences.* San Francisco: Freeman.

Grissom, R. J. (1994a). Probability of the superior outcome of one treatment over another. *Journal of Applied Psychology, 79,* 314–316.

Grissom, R. J. (1994b). Statistical analysis of ordinal categorical status after therapies. *Journal of Consulting and Clinical Psychology, 62,* 281–284.

Grissom, R. J. (1996). The magical number .7 +/− .2: Meta-meta-analysis of the probability of superior outcome in comparisons involving therapy, placebo, and control. *Journal of Consulting and Clinical Psychology, 64,* 973–982.

Grissom, R. J. (2000). Heterogeneity of variance in clinical data. *Journal of Consulting and Clinical Psychology, 68,* 155–165.

Grissom, R. J., & Kim, J. J. (2001). Review of assumptions and problems in the appropriate conceptualization of effect size. *Psychological Methods, 6,* 135–146.

Hiller, J. B., Rosenthal, R., Bornstein, R. F., Berry, D. T. R., & Brunell-Neuleib, S. (1999). A comparative meta-analysis of Rorschach and MMPI validity. *Psychological Assessment, 11,* 278–296.

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.

Hsu, L. M. (1996). On the identification of clinically significant client changes: Re-interpretation of Jacobson's cut scores. *Journal of Psychopathology and Behavior Assessment, 18,* 371–386.

Hsu, L. M. (2000). Effects of directionality of significance tests on the bias of accessible effect sizes. *Psychological Methods, 5,* 333–342.

Hsu, L. M. (2002a). Diagnostic validity statistics and the MCMI-III. *Psychological Assessment, 14,* 410–422.

Hsu, L. M. (2002b). Fail-safe *N*s for one- vs. two-tailed tests lead to different conclusions about publication bias. *Understanding Statistics, 1,* 85–100.

Hsu, L. M., & Field, R. (2003). Interrater agreement measures: Comments on kappa$_n$, Cohen's kappa, Scott's Π, and Aickin's α. *Understanding Statistics, 2,* 205–219.

Huberty, C. J. (2002). A history of effect size indices. *Educational and Psychological Measurement, 62,* 227–240.

Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (1982). *Meta-analysis: Cumulating research findings across studies.* Beverly Hills, CA: Sage.

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 59,* 12–19.

Kazdin, A. E. (2003). *Research design in clinical psychology* (4th ed.). Boston: Allyn & Bacon.

Lipsey, M. W. (1990). *Design sensitivity: Statistical power for experimental research.* Newbury Park, CA: Sage.

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment. *American Psychologist, 48,* 1181–1209.

McGraw, K. O. (1991). Problems with the BESD: A comment on Rosenthal's "How are we doing in soft psychology?" *American Psychologist, 46,* 1084–1086.

McGraw, K. O., & Wong, S. P. (1992). A common-language effect size statistic. *Psychological Bulletin, 111,* 361–365.

Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., Dies, R. R., et al. (2001). Psychological testing and psychological assessment. *American Psychologist, 56,* 128–165.

Nunnally, J. C. (1978). *Psychometric theory.* New York: McGraw-Hill.

Preece, P. F. (1983). A measure of experimental effect size based on success rates. *Educational and Psychological Measurement, 43,* 763–766.

Rosenthal, R. (1987). *Judgment studies: Design, analysis, and meta-analysis.* New York: Cambridge University Press.

Rosenthal, R. (1990). How are we doing in soft psychology? *American Psychologist, 45,* 775–777.

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.), *The handbook of research synthesis* (pp. 231–244). New York: Russell Sage Foundation.

Rosenthal, R. (1995a). Progress in clinical psychology: Is there any? *Clinical Psychology: Science and Practice, 2,* 133–149.

Rosenthal, R. (1995b). Writing meta-analytic reviews. *Psychological Bulletin, 118,* 183–192.

Rosenthal, R. (2002). Covert communication in classrooms, clinics, courtrooms, and cubicles. *American Psychologist, 57,* 839–849.

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

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

Rosnow, R. L., & Rosenthal, R. (1988). Focused tests of significance and effect size estimation in counseling psychology. *Journal of Counseling Psychology, 35,* 203–208.

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

Sheppard, W. F. (1898). On the application of the theory of error to cases of normal distributions and normal correlations. *Proceedings of the Royal Society, 62,* 170.

Sheskin, D. J. (2000). *Handbook of parametric and nonparametric statistical procedures.* New York: Chapman & Hall/CRC.

Strahan, R. F. (1991). Remarks on the binomial effect size display. *American Psychologist, 46,* 1083–1084.

Stuart, A., & Ord, J. K. (1987). *Kendall's advanced theory of statistics* (Vol. 1). New York: Oxford University Press.

Svartberg, M., & Stiles, T. C. (1991). Comparative effects of short-term psychodynamic psychotherapy: A meta-analysis. *Journal of Consulting and Clinical Psychology, 59,* 704–714.

Taylor, H. C., & Russell, J. T. (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection: Discussion and tables. *Journal of Applied Psychology, 23,* 565–578.

Thompson, K. N., & Schumacker, R. E. (1997). An evaluation of Rosenthal and Rubin's binomial effect size display. *Journal of Educational and Behavioral Statistics, 22,* 109–117.

Vargha, A., & Delaney, H. D. (2000). A critique and improvement of the CL common language effect size of McGraw and Wong. *Journal of Educational and Behavioral Statistics, 25,* 101–132.

Vargha, A., Rudas, T., Delaney, H. D., & Maxwell, S. E. (1996). Dichotomization, partial correlation, and conditional independence. *Journal of Educational and Behavioral Statistics, 21,* 264–282.

Walker, H. M., & Lev, J. (1953). *Statistical inference.* New York: Holt.

Wilcox, R. R. (1997). *Introduction to robust estimation and hypothesis testing.* New York: Academic Press.

Wilcox, R. R. (2003). *Applying contemporary statistical techniques.* New York: Academic Press.

Appendix

Formulas for Cohen's d, r_{pb}, and for Calculation of Biases of the BESD SRDs in Case 2a

Cohen's population d, for the binormal model, is defined by the following formula (adapted from Cohen, 1988, p. 20):

$$d = \frac{(\mu_{\mathrm{T}} - \mu_{\mathrm{C}})}{\sigma_{\mathrm{wg}}},$$

where μ_{T} = the mean of the treatment population, μ_{C} = the mean of the control population, and σ_{wg} = the standard deviation of either population (because within-group standard deviations are assumed to be equal in the binormal model).

The point-biserial correlation can be calculated using the following formula (adapted from Walker & Lev, 1953):

$$r_{\mathrm{pb}} = \frac{(\mu_{\mathrm{T}} - \mu_{\mathrm{C}})}{\sigma_{\mathrm{total}}} \sqrt{N_{\mathrm{T}} N_{\mathrm{C}}/[N(N-1)]},$$

where σ_{total} = the standard deviation of the combined groups, N_{T} = the number of treatment participants, N_{C} = the number of control participants, and $N = N_{\mathrm{T}} + N_{\mathrm{C}}$.

Calculation of BESD SRD biases for Case 2a is based on formulas presented in Rosenthal and Rubin (1982). The rationale of these formulas was explained as follows:

> In order to relate ϕ and ρ, we establish the following notation. Let $X = -1, +1$ indicate group membership, and let $\mathrm{E}(Y|X) = X\mu$, $\mu > 0$, and $\mathrm{Var}(Y|X) = 1$. Then $\mathrm{E}(X) = 0$, $\mathrm{Var}(X) = 1$, $\mathrm{E}(Y) = 0$, $\mathrm{Var}(Y) = 1 + \mu^2$, $\mathrm{Var}(Y^*) = 1$ [where $Y^* = -1$ if $Y < \mathrm{median}(Y)$ and $Y^* = 1$ if $Y > \mathrm{median}(Y)$], $\mathrm{Corr}(X, Y) = \rho = \mu/[1 + \mu^2]^{.5}$ or $\mu = \rho/[1 + \rho^2]^{.5}$. Also $\mathrm{Corr}(X, Y^*) = \phi = 1 - 2\mathrm{T}$, where T is the area from 0 to ∞ under the $X = -1$ group's Y distribution, or equivalently, the area from $-\infty$ to 0 under the $X = +1$ group's distribution . . . . Thus we can express ϕ as a function of ρ by $\phi = 1 - 2T$, where T is the area from $\rho/[1 + \rho^2]^{.5}$ to ∞ under the [standard normal distribution]. (Rosenthal & Rubin, 1982, p. 169)

That is, $T = 1 - \Phi(\mu)$, where $\Phi(\mu)$ is the cumulative distribution function of the standard normal distribution and $\mu = \mathrm{E}(Y|X = +1)$.

Received September 3, 2002
Revision received September 2, 2003
Accepted October 21, 2003 ■